

**LEARNING VISUALLY-GROUNDED REPRESENTATIONS
USING CROSS-LINGUAL MULTIMODAL PRE-TRAINING**

**ÇOK DİLLİ ÇOK KIPLI ÖN ÖĞRENME İLE GÖRSEL
TABANLI TEMSİLLERİN ÖĞRENİLMESİ**

MENEKŞE KUYU

ASSOC. PROF. DR. MEHMET ERKUT ERDEM

Supervisor

Submitted to
Graduate School of Science and Engineering of Hacettepe University
as a Partial Fulfillment to the Requirements
for the Award of the Degree of Master of Science
in Computer Engineering

August 2020

ABSTRACT

LEARNING VISUALLY-GROUNDED REPRESENTATIONS USING CROSS-LINGUAL MULTIMODAL PRE-TRAINING

Menekşe Kuyu

Master of Science, Computer Engineering Department

Supervisor: Assoc. Prof. Dr. Mehmet Erkut ERDEM

August 2020, 81 pages

In recent years, pre-training approaches in the field of NLP have emerged with the increase in the number of data and developments in computational power. Although these approaches initially included only pre-training a single language, cross-lingual and multimodal approaches were proposed which employ multiple languages and modalities. While cross-lingual pre-training focuses on representing multiple languages, Multimodal pre-training integrates Natural Language Processing and Computer Vision areas and fuse visual and textual information and represent it in the same embedding space. In this work, we combine cross-lingual and multimodal pre-training approaches to learn visually-grounded word embeddings. Our work is based on cross-lingual pre-training model XLM [1] which has shown success on various downstream tasks such as machine translation and cross-lingual classification.

In this thesis, we proposed a new pre-training objective called Visual Translation Language Modeling (vTLM) which combines visual content and natural language to learn visually-grounded word embeddings. For this purpose, we extended the large-scale image captioning

dataset Conceptual Captions [2] to another language—German using state-of-the art translation system to create a cross-lingual multimodal dataset which is required in pretraining. We finetuned our pre-trained model on Machine Translation (MT) and Multimodal Machine Translation (MMT) tasks using Multi30k [3] dataset. We obtained state-of-the-results on Multi30k test2016 set for both MT and MTT tasks. We also demonstrated attention weights of the model to analyze how it operates over the visual content.

Keywords: Cross-lingual Pre-training, Multimodal Pre-training, Transformer, Machine Translation, Multimodal Machine Translation

ÖZET

ÇOK DİLLİ ÇOK KIPLI ÖN ÖĞRENME İLE GÖRSEL TABANLI TEMSİLLERİN ÖĞRENİLMESİ

Menekşe Kuyu

Yüksek Lisans, Bilgisayar Mühendisliği

Danışman: Doç. Dr. Mehmet Erkut ERDEM

Ağustos 2020, 81 sayfa

Son yıllarda veri sayısındaki artış ve hesaplama gücündeki gelişmeler ile birlikte Doğal Dil İşleme alanında ön eğitilmiş model yaklaşımları ortaya çıkmıştır. Bu yaklaşımlar başta sadece tek dili kapsayacak şekilde olsa da, ardından çok dilli ve multimodal yapılar önerilmiştir. Çok kipli ön eğitilmiş modeller, Doğal Dil işleme ve Bilgisayarlı Görü alanlarının ikisini de kapsıyor olup görsel ve metinsel bilgiyi birleştirerek aynı uzayda ifade edilmesini hedef alır. Bu çalışmada, görsel temelli kelime gösterimlerini öğrenmek için diller arası ve çok kipli ön eğitim yaklaşımları birlikte kullanılmıştır. Çalışmamız, makine çevirisi ve diller arası sınıflandırma gibi çeşitli alt görevlerde başarı gösteren, diller arası ön eğitim modeli XLM tabanlıdır. Bu tez kapsamında, görsel temelli kelime vektörlerini öğrenmek için görsel içerik ve doğal dili birleştiren Görsel Çeviri Dili Modellemesi adı verilen yeni bir ön eğitim hedef önerildi. Bu amaçla, ön eğitimde gerekli olan diller arası çok kipli bir veri kümesi oluşturmak için son yıllarda önerilen en başarılı açık kaynak çeviri modelini kullanarak, büyük ölçekli bir görüntü altyazılama veri kümesi olan Conceptual Captions'ı [2], yeni bir dil; Almanca olarak genişlettik. Önerilen ön eğitilmiş model, Multi30k [3] veri kümesini kullanarak Makine Çevirisi (MÇ) ve Çok Kipli Makine Çevirisi (ÇMÇ) görevlerinde ince ayar yapılmıştır. Hem MÇ hem de ÇMÇ görevleri için Multi30k test2016 setinde literatürdeki

en başarılı sonuçlar elde edilmiştir. Ek olarak, önerilen modelin görsel içerik üzerinde nasıl çalıştığını analiz etmek için dikkat ağırlıkları görselleştirilmiştir.

Anahtar Kelimeler: Çok Dilli Ön Eğitim, Çok Kipli Ön Eğitim, Dönüştürücü, Makine Çevirisi, Çok Kipli Makine Çevirisi

ACKNOWLEDGEMENTS

First and foremost, I would like to thank to my supervisors Assoc. Prof. Dr. İbrahim Aykut ERDEM and Assoc. Prof. Dr. Mehmet Erkut ERDEM for giving me the opportunity to work together since my undergraduate years and always leading me to the right path,

Furthermore, I would like to thank my thesis committee members Prof. Dr. Pınar DUYGULU ŞAHİN, Assoc. Prof. Dr. Nazlı İKİZLER CİNBIŞ, Assoc. Prof. Dr. Sinan KALKAN and Asst. Prof. Dr. Emre AKBAŞ for reviewing my thesis and their valuable comments.

I deeply thank, my mother and my brother; Gül and Yiğit for always carry me forward with their support throughout my educational life.

I would like to thank İlker ŞAHİN, for reviewing my thesis and always supporting and motivating me throughout my master's degree.

Finally I would like to thank my friends Begüm Çıtamak for helping me in my thesis, and members of Hacettepe University Computer Vision Laboratory (HUCVL). I also would like to thank you Ozan Çağlayan for enlightening me with his ideas.

This thesis was supported in part by TUBA GEBIP fellowship awarded to E. Erdem, and the MMVC project funded by TUBITAK and the British Council via the Newton Fund Institutional Links grant programme (grant ID 219E054 and 352343575).

GENİŞLETİLMİŞ ÖZET

Kelimeleri düşük boyutlu vektörler ile temsil etmek, Doğal Dil İşleme alanında son yıllarda oldukça popüler bir konu haline gelmiştir. Kelime vektörleri, kelimelerin bağlam içindeki anlamlarının kodlanması ile oluşturulmaktadır. Bu vektörler, modern Doğal Dil İşleme’de büyük ölçekli verilerin denetimsiz şekilde eğitilmesi ile elde edilmektedir. Teknolojinin ve kaynakların gelişmesi ile birlikte, bu işlem için yapay sinir ağları kullanılmaya başlanmış ve çok büyük ölçekli veri kümeleri ile eğitim gerçekleştirilmiştir. Geleneksel yöntemler elde edilen kelime vektörleri, Doğal Dil İşleme’nin soru cevaplama, makine çevirisi birçok probleminde başarı göstermiştir.

Son zamanlarda geleneksel yöntemlerin aksine, kelimelerin anlamsal yakınlıklarını ve bağlamlarını gözetken yeni kelime vektörleri önerilmiştir. Bu bağlamsal yöntemler, diğerlerinin aksine kelimelerin arasındaki bağlamsal bilgiyi de kodlayabilmektedir. Bu yöntemler literatürde ön eğitim (pre-training) yöntemleri olarak adlandırılmaktadır. Kelime vektörlerinin gelişmesiyle birlikte, metinsel ve görsel bilginin bağlamsal yöntemlerle ifade edilmesi fikri ortaya çıkmıştır. İki farklı kaynak türünden gelen bilgilerin aynı uzayda kodlanmasına çok kipli ön eğitim (multimodal pretraining) adı verilmektedir. Bu yöntem sayesinde, metinsel ve görsel kaynaklardan elde edilen vektörler görsel soru cevaplama, görüntü ve video altyazılma gibi Doğal Dil İşleme ve Bilgisayarlı Görü alanlarının kesiştiği problemlerin çözümünde kullanılabilir.

Bu çalışmada, hem çok dilli (multilingual) hem de çok kipli (multimodal) bir ön eğitim yöntemi üzerinde çalışılmıştır. Bunun için ilk olarak, birden fazla dilde metinsel bilgiyi ve bunlara karşılık gelen görsel bilgiyi içeren büyük ölçekli bir veri seti ihtiyacı doğmuştur. Bunun için yaklaşık 3.3 milyon cümle/görüntü çifti içeren Conceptual Captions [1] veri seti kullanılmıştır. Fakat bu veri seti yalnızca İngilizce açıklamalar ve karşılık gelen görüntüyü içermektedir. Bu veri setini çok dilli hale getirmek için İngilizce-Almanca çeviride yüksek performans gösteren açık kaynak kodlu bir çeviri modeli kullanılmıştır. Bu sayede, büyük ölçekli çok dilli ve çok kipli bir veri seti elde edilmiştir.

Birden fazla dil için kelime vektörleri öneren XLM [2] modeli, önerilen çok dilli çok kipli modelin temelini oluşturmaktadır. XLM modeli, son yıllarda sekanstan sekansa olarak adlandırığımız problemlerde yüksek başarı gösteren Transformer modelini baz almıştır. Biz de bu çalışmada aynı mimariyi koruyarak, çok dilli bir yapıya görüntü bilgisini de dahil ederek hem görsel, hem metinsel bilgiyi aynı anda kodlayan bir model geliştirdik. Bu model, Görsel Dil Modelleme (GDM) adını verdiğimiz bir ön öğrenme hedefini kullanmaktadır.

GDM geliştirilirken XLM tarafından önerilen Çeviri Dil Modelleme (ÇDM) öğrenme hedefinden ilham alınmıştır. GDM, bu çok dilli yapıya görüntü bilgisini de ekleyerek çok-kipli çok dilli bir (multimodal cross-lingual) yaklaşımı önermektedir. Basitçe anlatmak gerekirse GDM, ÇDM hedefinin görsel bilgi ile zenginleştirilmiş versiyonudur. Kaynak görüntüden elde edilen her bir nesne, görsel kelime olarak ifade edilmektedir. Bu görsel kelimeler çok dilli sekansın peşine eklenerek çok kipli yapı elde edilmiş olur.

Önerilen model, yukarıda bahsedilen Conceptual Captions [2] veri kümesinin çok dilli varyasyonu (İngilizce/Almanca) ile eğitilmiştir. Görüntü bilgisi ise, Masked Faster-RCNN [3] tabanlı Open Images [4] veri seti üzerinde bir nesne algılama modeli kullanılarak çıkarılmıştır. Her bir görüntü, en yüksek olasılıklı 36 adet nesne bölgesi (object region) ile ifade edilmiştir. Bu nesne bölgelerinde çıkarılan özellikler (feature), görsel kelimeler olarak yorumlanabilir.

Eğitilen çok dilli çok kipli model, makine çevirisi üzerinde ince ayar (finetune) yapılmıştır. Makine çevirisi problemi üzerinde çalışılmasının sebebi, bazı kelimelerin cümle içindeki anlamlarının belirsiz olması ve görüntü bilgisinin bu belirsizliği ayrıştırabileceği hipotezidir.

Önerilen modelin makine çevirisi görevinde ince ayar yapılması için Multi30k [3] veri seti kullanılmıştır. Bu çalışmada Conceptual Captions İngilizce/Almanca otomatik çeviriler kullanılarak elde edilen modeller, Multi30k İngilizce/Almanca ve insanlar tarafından çevirilerin üzerinde ince ayar yapılmıştır.

Deneyler kapsamında, XLM modelinin bir ön öğrenme görevi olan Çeviri Dil Modelleme (ÇDM) kullanılarak da Conceptual Captions veri seti üzerinde çok dilli bir öğrenme gerçekleştirilmiştir.

Bu model de aynı şekilde Multi30k veri seti üzerinde ince ayar yapılmıştır. Ön eğitilen modellerin her ikisi de Multi30k test 2016 veri seti üzerinde test edilmiştir. Sayısal karşılaştırma için BLEU, METEOR ve MLT metrikleri kullanılmıştır. Elde edilen sayısal sonuçlara göre, geliştirilen model BLEU metriğinde hem doğrulama hem de test kümesinde daha yüksek başarımlar göstermiştir.

Deneysel sonuçlara bakıldığında, GDM ile ön eğitim gerçekleştirilen ve Multi30k üzerinde ince ayar yapılan model, ÇDM'ye göre daha yüksek sonuçlar vermektedir. Bu deneyler, görüntü bilgisi kullanılarak (GDM) elde edilen vektör representasyonlarının, sadece metinsel bilgi kullanılarak elde edilenlerden (ÇDM) anlamsal açıdan daha zengin olduğu hipotezi doğrulanmıştır.

CONTENTS

	<u>Page</u>
ABSTRACT	i
ÖZET	iii
ACKNOWLEDGEMENTS	v
GENİŞLETİLMİŞ ÖZET	vi
CONTENTS	ix
TABLES	xi
FIGURES	xiii
ABBREVIATIONS.....	xiv
1. Introduction	1
1.1. Scope of the Thesis	2
1.2. Contribution.....	3
1.3. Organization	4
2. Background.....	5
2.1. RNN-Based Approaches	5
2.1.1. Long Short Term Memory (LSTM)	6
2.2. Attention-based Approaches	7
2.2.1. Transformer	8
2.3. Neural Machine Translation.....	10
2.4. Multimodal Machine Translation	11
3. Pre-training Approaches for Natural Language Processing	12
3.1. Pre-trained Traditional Word Embeddings	13
3.2. Pre-trained Contextual Embeddings	14
3.2.1. Pre-training Tasks	14
Language Modeling	14
Masked Language Modeling (MLM)	15
Next Sentence Prediction (NSP).....	15
3.3. Pre-trained Models for Contextual Embeddings	15

3.3.1. Monolingual Pre-training	15
3.3.2. Cross-lingual Pre-training	21
3.3.3. Multimodal Pre-training	22
4. Cross-lingual Multimodal Pretraining	25
4.1. Textual Representation	25
4.2. Object Representation	26
4.3. Model	27
4.3.1. Visual Translation Language Modeling (vTLM)	29
4.4. Pre-training Settings	29
4.5. Downstream Tasks.....	31
4.5.1. Machine Translation	31
4.5.2. Multimodal Machine Translation	32
5. Results and Analysis	32
5.1. Datasets	32
5.1.1. Conceptual Captions	32
5.1.2. Multi30K	34
5.2. Evaluation Metrics.....	34
5.2.1. BLEU	35
5.2.2. METEOR.....	36
5.2.3. Multimodal Lexical Translation Accuracy.....	36
5.3. Quantitative Results	37
5.4. Qualitative Results.....	39
6. Conclusion.....	43
REFERENCES	50

TABLES

	<u>Page</u>
Table 3.1. The summarization of the models proposed for multimodal pretraining	25
Table 5.1. Experimental results of our models and recent state-of-the-art models on Multi30k test2016 dataset. There is no MT and MMT systems in the literature employs pre-training and fine-tuning on Multi30k dataset. Thus, the systems shared here does not use pre-training.	39
Table 5.2. Experimental results on Multi30k test2017 dataset.....	39
Table 5.3. Example translations for produced from MT systems using TLM and vTLM models. We back-translated each translation to English using Google Translate which is shown under the translation (ET).	41

FIGURES

	<u>Page</u>
Figure 2.1. The Transformer architecture, taken from [4]	9
Figure 2.2. The architecture of typical NMT system. The source sentences is encoded by encoder. The encoded source representation is passed to the decoder for generating the target sentence.	11
Figure 2.3. An example demonstration of MMT system. The source image and source sentence are fused together to generate target translation.	12
Figure 3.1. The input sequence of BERT	17
Figure 3.2. The input sequence of SpanBERT (Figure from [5]) . The masked span is "an American football game".....	19
Figure 4.1. The TLM objective of XLM model. The words are randomly masked for both languages.....	28
Figure 4.2. The proposed vTLM objective.....	30
Figure 5.1. Samples taken from the Conceptual Captions dataset. Alt-text descriptions are the raw descriptions collected from Internet. Conceptual Captions descriptions are clean and fluent.	33
Figure 5.2. Example sentences from Conceptual Captions [2] and their automatic translations.....	34
Figure 5.3. Example captions from Multi30k dataset [3].	35
Figure 5.4. An example from MLT [6] dataset	37
Figure 5.5. We showed the BLEU score on Multi30k validation set while fine-tuning for TLM and vTLM models	42
Figure 5.6. Example translations from MT and MMT systems for Multi30k test2017 dataset.....	43

Figure 5.7. We visualize the attention weights for each head in pre-training for a German sentence and object regions. The rows contains German tokens and columns are named as the object label. We replaced the token "fußball" with [MASK] token to observe model's attention on visual input.....	47
Figure 5.8. We visualize the attention weights for each head in pre-training for a German sentence and object regions. The rows contains German tokens and columns are named as the object label. We replaced the token "ziege" with [MASK] which means "goat" in English.	49

ABBREVIATIONS

LSTM	Long-Short Term Memory
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
BPE	Byte-Pair Encoding
SPM	Sentence Piece Model
BLEU	Bilingual Evaluation Understudy
METEOR	Metric for Evaluation of Translation with Explicit Ordering
CIDEr	Consensus Based Image Description Evaluation
VQA	Visual Question Answering
VCR	Visual Common Reasoning
MLT	Multimodal Lexical Translation
NLG	Natural Language Generation
MT	Machine Translation
MMT	Multimodal Machine Translation
TLM	Translation Language Modeling
vTLM	Visual Translation Language Modeling

1. Introduction

Today, representing word semantics with low-dimension vectors, which are referred to as word embeddings, has gained popularity in the Natural Language Processing (NLP) community. The most common method to obtain word representations in modern NLP is to train large-scale unlabeled textual data in an unsupervised manner. With the development of technology and resources, methods based on artificial neural networks are frequently used for processing large-scale corpora. Even though, global representation of the word obtained with the traditional methods shows success with different NLP tasks such as question answering [7] and natural language inference (NLI) [8], such representation lacks contextual information about the words.

Recently, contextual word representations have been proposed that take into account the semantic affinities and contexts of words, unlike traditional methods. These contextual methods can encode contextual information between words using a new strategy called pre-training. Contextual word representations have become widely used in the literature and shown success for Natural Language Generation problems. With the enhancement of word representations, the idea of expressing textual and visual information with contextual methods has emerged. The encoding of information from two different source types in the same space is called multimodal pre-training. Multimodal representations obtained from textual and visual sources can be used to solve problems where Natural Language Processing and Computer Vision intersect, such as visual question answering, image and video captioning.

In this thesis, we focused on cross-lingual and multimodal pre-training strategies. We extended the multimodal pre-training approach to multilinguality to enhance word representations with the contribution of visual information for Natural Language Generation tasks involving multiple languages such as Machine Translation and Multimodal Machine Translation. For developing cross-lingual multimodal model, we needed a large-scale dataset containing textual information in more than one language and corresponding visual information. For this purpose, we used Conceptual Captions [2] dataset, which contains approximately

3.3 million sentence / image pairs.. However, this dataset contains only the English descriptions and the corresponding image. To extend this dataset into multiple languages, we used an state-of-the-art open source English to German translation system to translate English descriptions into German language.

In this study, we proposed a cross-lingual multimodal pre-training objective based on XLM [1] architecture which proposes word representations for multiple languages and employs Transformer [4] model, which has shown success in sequence-to-sequence tasks in recent years. We extended the XLM architecture to encode both visual and textual information at the same time by preserving the same architecture and including image information in a multilingual structure. This model uses a preliminary learning objective that we call Visual Translation Language Modeling (vTLM). vTLM proposes a multimodal cross-lingual approach by adding image information to this multilingual structure. In other words, vTLM objective is a version of TLM objective proposed by XLM, and enriched by visual information.

We used the word representations obtained from our cross-lingual multimodal model in two different Natural Language Generation tasks; Machine Translation (MT) and Multimodal Machine Translation (MMT). The reason we focused on the translation tasks is to enriched word representations using visual information and solve the ambiguity problem in the translation. We used a multimodal translation dataset; Multi30k [3] in finetuning of MT and MMT systems. In experimental results, we demonstrated that proposed objective vTLM obtains higher results than XLM's TLM objective for MT and MMT tasks. Results also shows that word representations obtained using visual information are semantically richer than those obtained using only textual information.

1.1. Scope of the Thesis

We proposed a new pre-training approach that fuses visual and textual data to create visually-grounded word embeddings for multiple languages. We inspired from XLM [1] which is

a multilingual pre-training model using Transformer [4] architecture. We extended XLM architecture into multimodal setting by adding visual information in the network.

To train our model, we automatically translated the Conceptual Captions [2] into German to create multilingual multimodal dataset. We compared our model with the textual-only pre-trained model called XLM [1] and evaluated both models on two different downstream tasks; Machine Translation and Multimodal Machine Translation. Our goal is to demonstrate visually-grounded word embeddings are richer than the textual embeddings. Visually grounded embeddings can also be beneficial for various tasks such as image captioning, video captioning and visual question answering.

1.2. Contribution

Our contributions can be summarized as follows:

1. We developed a new pre-training approach which integrates multiple languages and visual content which is called cross-lingual multimodal pre-training. We evaluated the proposed pre-trained model on downstream tasks; Machine Translation and Multimodal Machine Translation and achieve state-of-art results on Multi30k [3] dataset.
2. To train our cross-lingual multimodal model, we needed a large-scale dataset contains aligned image/caption pairs into multiple languages. For this purpose, we created a new multilingual multimodal dataset using Conceptual Captions [2]. We automatically translated English captions into German using state-of-the-art translation system, fairseq [9].
3. We demonstrated the attention weights of the proposed Transformer [4] based model to analyze the how model benefits from visual content in the training phase. We showed that our models learn to attend correct visual information while predicting the words.

1.3. Organization

In Chapter 1, we present our motivation and contributions and the scope of the thesis. In Chapter 2, we provide a general background for Natural Language Generation. In Chapter 3, we give an overview of recent pre-training approaches for language and multimodal pre-training. In Chapter 4, we introduce a new pre-training objective for cross-lingual multimodal pre-training, give details about the representation of the textual and visual data and share the experimental setups and downstream tasks. In Chapter 5, we give detailed information about the datasets and the evaluation results we used and discuss about experimental results. In Chapter 6, we summarize our contributions and share ideas for the future work.

2. Background

With the recent developments in the field of Deep Learning, new approaches have been proposed for the modeling and representing the natural language. Neural Language Generation (NLG) has become a widely studied problem with the increasing need to understand and interpret natural language that is essentially the process of generating natural text obtained from humans by an automated system. NLG has many application areas such as Machine Translation, Text Summarization and Text Correction. In this section, we provided the background for Natural Language Generation (NLG) sub-tasks that we applied in this thesis and present the widely used the Deep Learning architectures that are employed for NLG.

2.1. RNN-Based Approaches

Recurrent Neural Networks (RNNs) is a very popular architecture which have shown encouraging performance in various NLG tasks such as language modeling an machine translation. The reason RNNs are successful in these tasks is that they can model long-range dependencies in textual data adequately. Another reason that RNNs are highly preferred is to encode variable length input sequences into a fixed-length vector embedding.

For an input sequence $w = (w_1, w_2...w_i)$, RNN updates its hidden states $w = (h_1, h_2...h_i)$ for each time step and the last state of the RNN represents the entire input sentence. The tokens in the input sequence are first converted to one-hot vectors which are later transformed into continuous word representations x_i using pre-trained word representations or the representations matrix, which is jointly trained with the network. Then, the RNN's hidden state is updated as follows;

$$h_i = f(w_i, h_{i-1}) \quad (1)$$

In Equation 1, f is the function that changes according to the RNN type, and h_i shows the hidden state of the RNN. The initial hidden state h_0 is usually a vector filled by zeros. In

RNNs, the hidden state of the current timestep is dependent on all of the previous hidden states. Although this shows the usability of RNNs for long sequences, in reality RNN cannot model the a few timestamp cannot model the context longer than a few steps. This issue is called vanishing and exploding gradients [10] which cause the incapability of modeling long-range dependencies.

2.1.1. Long Short Term Memory (LSTM)

Long Short Term Memory (LSTM) [11]) networks are proposed to solve the constraints of RNNs caused by vanishing gradients. LSTMs are able to model long-range dependencies using extra hidden layer and the memory cells for each hidden layer which are capable of storing information for long periods of time-steps. Each memory contains 3 gates; an input gate i_t which determines the cells to be updated, an output gate which determines what the next hidden state should be, and a forget gate f_t which controls the information to be forgotten. In Equation 2, \odot stands for element-wise multiplication and σ is the sigmoid activation function and b is the bias term.

$$\begin{aligned}
 i_t &= \sigma \left(W^{(i)} w_t + U^{(i)} h_{t-1} + b^{(i)} \right) \\
 f_t &= \sigma \left(W^{(f)} w_t + U^{(f)} h_{t-1} + b^{(f)} \right) \\
 o_t &= \sigma \left(W^{(o)} w_t + U^{(o)} h_{t-1} + b^{(o)} \right) \\
 u_t &= \sigma \left(W^{(u)} w_t + U^{(t)} h_{t-1} + b^{(t)} \right) \\
 c_t &= i_t \odot u_t + f_t \odot c_{t-1} \\
 h_t &= o_t \odot \tanh(c_t)
 \end{aligned} \tag{2}$$

With the help of the gating mechanism, LSTMs are able to model dependencies over longer time steps. This gating mechanism accommodates adjust hidden states' values which prevents vanishing or exploding gradients which makes LSTMs powerful architecture in NLG tasks.

2.2. Attention-based Approaches

RNNs show promising results in sequence-to-sequence tasks, but if the length of the input sequence is large, RNN often fails to generate a good summarization vector. Experiments using variant length sentences have shown that the performance of the model dramatically decreases with the increase of the sentence length [12]. To overcome this problem, [13] proposed an attention mechanism to be used in encoder-decoder architectures. The idea behind the attention approach is to represent each token with a vector referred to as an annotation vector rather than representing whole sentence with a single vector. Annotation vectors are later combine into a context vector c which is calculated in each time step. This strategy assists the decoder attend to on different parts of the input sequence while generating the target sentence. Research has demonstrated that a bidirectional encoder with the attention mechanism overcomes the performance problem when the sentences are long.

$$c_t = \sum_{i=1}^N a_{it} h_i \quad (3)$$

$$e_{it} = \text{align}(h_i, h_{t-1}) \quad (4)$$

$$\alpha_{it} = \frac{\exp(e_{it})}{\sum_{t'} \exp(e_{t'})} \quad (5)$$

To generate the i -th target token, context vector c_t is computed as shown in Equation 3 where N corresponds to the length of the sentence, a_{it} is attention weight, and h_i is an annotation vector. Alignment scores e_{it} can be obtained by the alignment function, which aims to capture the relation between annotation vector h_i in encoding phase and the last hidden state of the decoder h_{t-1} . Attention weights α_{it} are calculated with the softmax function which converts alignment scores into probabilities (Equation 5).

2.2.1. Transformer

With the recent advances of the attention mechanism, a fully attention-based approach called Transformer [4] has been proposed and has made a big impact in the NMT literature.

Transformer model employs encoder-decoder architecture where both encoder and decoder includes 6 stacked layers. The encoder includes of two parts; multi-head attention mechanism and fully connected feed-forward layer. The decoder also have the same sub-layers as the encoder and additionally, there is a separate multi-head attention layer which accepts the output of the encoder and the decoder's attention layer.

Transformer architecture (Figure 2.1.) is the first network which is completely based on the self-attention. The self-attention mechanism captures dependencies between input tokens inside the sequence instead of computing the dependency between input and output sequence. In this way, the attention-based source representation is obtained.

The self-attention mechanism uses a dot product as an alignment function, which receives query Q , key K , and value V as inputs. This attention mechanism executes in a multi-head manner. First, inputs are projected into keys, queries, values, and then the attention function is applied (Equation 6).

$$\mathcal{A}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d}} \right) V \quad (6)$$

Self-attention applies only one attention function in each timestep, but the authors proposed Multi-Head Attention Mechanism which is parallelized and is able to attend to different portions of the input sequence. At the end, the output of each attention head is summed to calculate final the context vector C ;

$$C = \sum_{i=1}^h \mathcal{A} \left(QW_i^Q, KW_i^K, VW_i^V \right) W_i^O \quad (7)$$

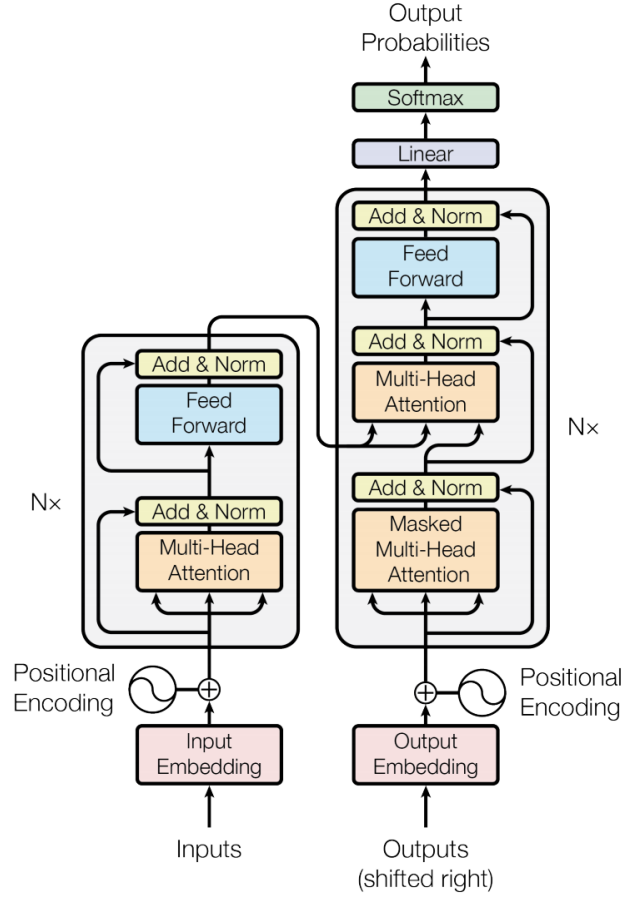


Figure 2.1. The Transformer architecture, taken from [4]

where W matrices are trainable parameters, and h is number of head.

The Transformer architecture does not employ any recurrent and convolutional layers. Therefore, it's essential to encode the locations of tokens in sentences. The authors proposed "positional encoding" which is applied both input and output sequence. For positional encoding function; the authors used sine and cosine functions that is shown in Equation 8 where p is the position, d is the embedding dimension and i is the model's dimension. The authors also experimented with learned and fixed positional encodings which results similar results.

$$\begin{aligned}
 PE_{(p,2i)} &= \sin\left(p/10000^{2i/d_{\text{model}}}\right) \\
 PE_{(p,2i+1)} &= \cos\left(p/10000^{2i/d_{\text{model}}}\right)
 \end{aligned}
 \tag{8}$$

Transformer based encoder-decoder networks outperform the RNNs for many NLG tasks and have become a commonly used architecture for sequence-to-sequence tasks.

2.3. Neural Machine Translation

The automatic translation from source language into the target language without any human supervision is referred to as machine translation (MT) which is a highly popular task among NLP researchers. MT has many applications areas, particularly based on communication applications and this approach reduces the work required substantially.

Over the past decades, MT has maintained its popularity and various effective methods have been introduced to improve MT system performance. In recent years, deep learning approaches have gained attraction for different research topics and the term Neural Machine Translation (NMT) has been proposed. The first attempt for NMT [14] did not perform well as a consequence of hardware limitations. In 2010s, deep learning gained substantial attention with the development of computational powers that easily accessible for anyone. NMT task were also reborned with the deep learning era. The first successful application is proposed by [15] based on deep neural networks.

NMT takes the source sentences and directly translates them into the target language. This is an end-to-end task, and the goal is to determine the correct target sentence for the corresponding input sentence. This can be also interpreted as a classification problem where the labels are the words in the vocabulary. Generally, NMT systems perform with two main processes; encoding and decoding (Figure 2.2.).

There are many different architectures proposed for NMT task, and they can be divided into two category; recurrent and not recurrent. In early years of NMT development, Deep Neural Network (DNN), Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) were the most common selections, but recently attention-based approaches has begun to be used for this task.

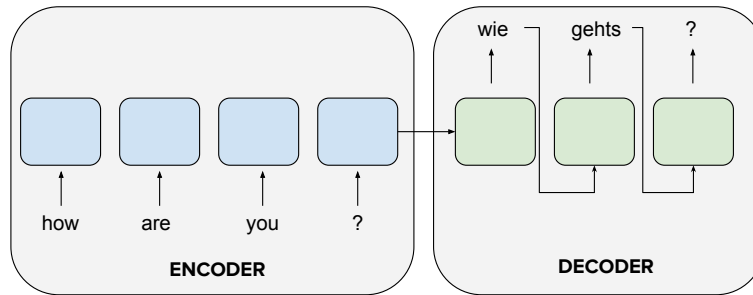


Figure 2.2. The architecture of typical NMT system. The source sentences is encoded by encoder. The encoded source representation is passed to the decoder for generating the target sentence.

2.4. Multimodal Machine Translation

MT involves an automatic translation from source language into the target language without any human supervision. This approach has made a significant contribution to automatic translation systems. Since the standard MT systems uses only the text data, the final translations may contain ambiguity and does not correctly translate the polysemous words. Multimodal Machine Translation task [16] is proposed to enhance the translations with the help of a visual content which integrates the CV and NLP areas. Different from MT, MMT system use an visual content beside the source sentsce which is illustrated in Figure 2.3..

The workshop of Multimodal Machine Translation is organized under Machine Translation conference in 2016 [17], 2017 [18] and 2018 [19]. The tasks under the workshop focused on the generation of image captions into the different languages; French, German and Czech. This can also be interpreted as the composition of the translation and image captioning tasks. The Multi30k dataset [3] which is the multilingual extension of image captioning dataset; Flickr30k [20], is proposed specifically for MMT task and each year, the workshop organizers released a new test set for multiple languages.

The methodology used in MMT task is similar to MT. Before the launch of the Transformer architecture, most of the MMT systems employs RNN-based solutions [21], [22] which is

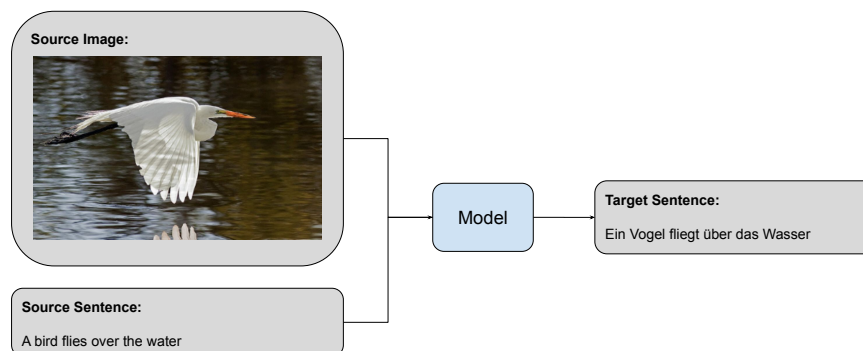


Figure 2.3. An example demonstration of MMT system. The source image and source sentence are fused together to generate target translation.

previously explained in MT section. Last workshop in 2018, almost every MMT system applies Transformer-based methods [23], [24].

3. Pre-training Approaches for Natural Language Processing

With the recent developments of deep learning approaches and the increase of the computational power, model sizes are becoming larger, resulting in a massive amount of model parameters. Hence, a huge amount of data are necessary to train these models and overcome the overfitting problem. Thanks to the rapid growth of the Internet, collecting and accessing huge amounts of data has become easier. Although most NLP tasks need an annotated dataset, most of the time, building large-scale annotated datasets is extremely costly. Because collecting large-scale unlabeled data is easy, we can benefit from these data and learn word representations. Then, we can use these representations to other NLP tasks.

Recent studies have shown that using representations from a pre-trained model yields a significant improvement in the performance of many NLP tasks. The global word representations can be learned with a pre-trained network with an unlabeled large text corpora, and these representations can assist for the downstream tasks. Pre-trained word representations

also enables a better initialization of the model, which can accelerate convergence on the goal task.

Pre-training has also been previously used in different research areas such as Computer Vision. Researchers have trained models with ImageNet [25] dataset, which is a large image corpus, and learn representation for images. Then, they perform finetuning on various vision and multimodal tasks. Many studies showed that using pretrained image features is a powerful initialization method because, in the pre-training phase, the model is able to generate good representation of images. Pre-training a language model with large corpora also showed success in NLP for the many downstream tasks.

3.1. Pre-trained Traditional Word Embeddings

The development of word representations is a problem that has been extensively studied in recent years. One of the first studies using word embeddings in down-stream tasks was performed by [26]. The results showed that using word embeddings from a pre-trained neural network improves the performance of many tasks while avoiding task-specific engineering. [27] proposed two different architectures termed Continuous Bag-of-Words and Continuous Skip-gram to learn word embeddings. Their approach was able to generate high-quality representations with a simple approach and much lower computational cost. In the next study [28], the authors extend the Skip-gram model and generate higher quality embeddings with much less training time. This model is commonly called Word2vec, and it is one of the most popular embeddings in the literature. Another popular word embedding model is GloVe [29], which uses a weighted least squares model. They used word-word co-occurrence statistics from a combination of different large text corpora for training.

Context2Vec [30] is another unsupervised model for learning word embeddings. They learned generic context embeddings with bidirectional LSTM, and it embeds the entire context of the sentence and the target word in the same space. There are many similar works that generate word embeddings from textual data such as paragraphs [31]. All models mentioned above have shown improvement in various NLP tasks, but they all lack of contextual information.

They all embed the word into a fixed-size vector representation in place of the contextual representation.

3.2. Pre-trained Contextual Embeddings

3.2.1. Pre-training Tasks

For learning a representation of a language, pre-training tasks must be defined. In computer vision, pre-trained models use large-scale annotated training data such as ImageNet [25]. But in NLP, annotated datasets are not large enough to pre-train a model except from the MT. In this section, we investigate the most commonly used pre-training tasks literature.

Language Modeling The classical method to learn word embeddings in an unsupervised fashion is with language modelling. A language model can be described as a probability distribution over a sequence of words. For given N tokens (t_1, t_2, \dots, t_N) , a language model models the probability of token t_i :

$$p(t_1, t_2, \dots, t_N) = \prod_{i=1}^N p(t_i | t_1, t_2, \dots, t_{i-1}) \quad (9)$$

Language models train with the maximum likelihood estimation (MLE) method with large text corpora. The term language model is often used as a unidirectional language model which calculates the conditional probability with the benefit of the previous words; $(t_1, t_2, \dots, t_{i-1})$ in a left-to-right manner. Since the unidirectional language models encode words with the context of words on the left, we can not obtain good quality representations. For a richer representation, two unidirectional language models are commonly used to encode sequence in both left-to-right and right-to-left manners. This improved approach is referred to as a bidirectional language model.

Masked Language Modeling (MLM) Another commonly used pre-training task in NLP literature is Masked Language Modeling (MLM) which proposed by [32]. The authors also named this task as Cloze. In MLM, some tokens from the input sequence are masked in the training step and the model learns to predict these masked tokens by looking at the remaining tokens. Most of time time, we can approach MLM as a classification problem. We give the input sequence to a encoder and use a softmax classifier to make a prediction that uses the output representations from the the encoder. Another approach to solve MLM is encoder-decoder networks. We feed the masked input sequence to the encoder and the decoder generates the masked tokens in an auto-regressive manner.

Next Sentence Prediction (NSP) Some of the important NLP tasks, such as question answering, aim to model the connection between two sentences, and language modeling is not able to identify this connection directly. Next Sentence Prediction (NSP) [33] is a task that learns two sequence from the training corpora that are consecutive sentences. To choose pre-training sentences X and Y , 50% of the time Y is followed by X , and 50% the time X and Y are just random sentences in corpora. With this learning manner, the model can understand the relation between two sentences.

3.3. Pre-trained Models for Contextual Embeddings

3.3.1. Monolingual Pre-training

The first work in the NLP literature that propose a pre-trained model to be finetuned on text classification and sentiment analysis [34]. The authors pre-train a model that consist of sequence auto encoder and RNNs by using language modelling. They used weights from the pretrained sequence encoder to initialize a supervised network and they obtained better results than the randomly initialized networks. [35] propose a different method to increase the performance of sequence models. They used both encoder and decoder weights from pretrained network that was trained separately with two language models. They focused on English-to-German translation task and used News Crawl English and German corpora of a

WMT dataset for pretraining. They finetuned a MT system on WMT English-German corpora and obtain better performance than the models that are randomly initialized. Another aspect they highlight is that the machine translation performance significantly reduces (almost the same as without pretraining) if the pretraining network is only trained by the WMT parallel corpus but remains similar if trained with another large corpus. This observation shows that a large corpus is essential for pretraining.

ELMo [36] introduce a model referred to as ELMo which retrieves contextual representations using the internal states from a pretrained bidirectional language model (biLM). The authors used two different LSTM layers —forward and backward— to encode information in both left and right contexts. For finetuning, the linear combination for vectors in internal states are used, which improves the performance compared the using vectors from the top LSTM layer. More specifically, ELMo combines layer representations L into single vector for each token x .

$$\text{ELMO}_x^{\text{task}} = \gamma^{\text{task}} \sum_{j=0}^L s_j^{\text{task}} \mathbf{h}_{x,j} \quad (10)$$

In Equation 10, γ^{task} is the scalar task-specific vector, s^{task} is the weights normalized by softmax, and $(h_{1,j}, h_{1,j}, \dots, h_{N,j})$ is the hidden representation for a sequence length N . For a target downstream task, the authors essentially obtained the all layer representations for each token and the target model learns the combination of the representations. The proposed model evaluated on six different downstream tasks such as sentiment analysis and classification, named entity recognition. The authors also show that ELMo reduces the need for a large amount of training data with experiments on different portions of the SNLI corpus [37].

GPT [38] learn universal word representations, which can be transferred into different NLP tasks only with small adjustments. They proposed a two-stage training process; training a language model using large unlabeled corpora and supervised fine-tuning on downstream tasks. For the language model, they used the Transformer architecture, which has been shown to be a powerful alternative to RNNs and captures long-range dependencies better. The BookCorpus dataset [39] is used to train language models that consist of around 7,000 books. For transferring the pre-trained weights, they employ task-specific adjustments

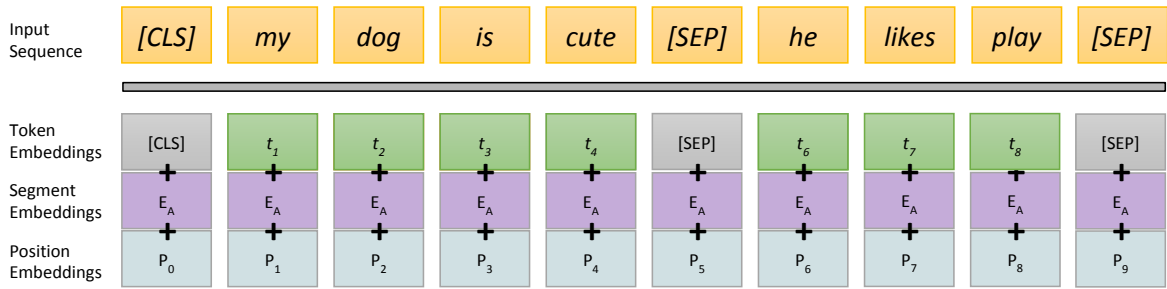


Figure 3.1. The input sequence of BERT

inspired from [40] and converts text input into a single sequence of tokens using the specific tokens. This architecture obtained better results than other task-specific approaches for 9 out of 12 tasks. The following work GPT2 [41] proposes mostly similar architecture but uses a different dataset called WebText, which consists of millions of web pages to show that pre-training with an adequate amount of large and diverse data can obtain to good performance for a wide variety of domains. The authors also demonstrate that the ability of language models can be adapted to various tasks in zero-shot setting. They achieved state-of-the-art performance on 7 out of 8 datasets that were tested for language modeling. Whereas the aforementioned models use character or word-based language models, this study followed a subword-based approach using Byte Pair Encoding (BPE) algorithm [42]. This model has achieved significant success in tasks such as question answering, and it is certain whether this success as a result of the subword approach.

BERT [33] propose a deep bidirectional Transformer model called BERT for contextual language representation. BERT employs MLM objective in pretraining phase which empowers to fusion of left-to-right and right-to-left contexts. They also introduce NSP objective to improve the performance of the downstream tasks such as question answering and natural language inference which depends on the relation between sentence pairs. BERT uses different special tokens in the same manner as GPT. The classification [CLS] token marks the beginning of the sentence which is used after classification tasks as sequence embedding. The other token is the separator [SEP] which is used to distinguish the sentences because input sentences are represented as contiguous sequence pairs.

The input sequence of BERT consists of the summation of token embedding, positional embedding and segment embedding (Figure 3.1.). Segment embedding is used to represent input and output sources and the positional embedding shows the position of the token in the sequence.

BERT uses the same architecture for both pretraining and finetuning. For finetuning step, the final hidden state of the $[CLS]$ token is used for classification tasks. The authors achieved new state-of-the-art performance on different NLP tasks and improved the GLUE [43] score 7.7%.

After the success of BERT, many researchers have worked on the architecture and the objective of BERT model to improve the performance of downstream tasks. RoBERTa [44] analyzed the impact of hyperparameter selection and size of the training set and showed that BERT is undertrained. The proposed solution includes the removal of NSE objective, using larger batches and more data in pre-training and changing the masking strategy. The authors proposed a new large-scale dataset called CC-NEWS to evaluate the dataset size more effectively. In original BERT, random masking is performed only once in the data preprocessing step. RoBERTa changed the strategy and performed dynamic masking which masks the tokens randomly in training phase. The experiments show that dynamic masking improves the accuracy on 2 out of 3 datasets and they also obtained new state-of-the-art results on GLUE.

ERNIE [45] is a model which takes advantage of knowledge-based masking strategies. They use two different masking schemes for pre-training; phrase-level and entity-level masking. Most of the time, entity and phrases consist of more than one word and instead of masking a single token, whole entity or phrase is masked. In this way, semantic and syntactic knowledge can be learned from the masked unit. The following work ERNIE2.0 [46] employs different pre-training tasks through multi task learning and improves the GLUE score over BERT.

SpanBERT [5] is a pre-training approach based on BERT that has been developed to understand and represent the spans of a text. The authors modify BERT by changing the masking strategy and the training objective. Instead of masking the random tokens, they mask random spans and propose a new training objective span-boundary objective (SBO) to predict

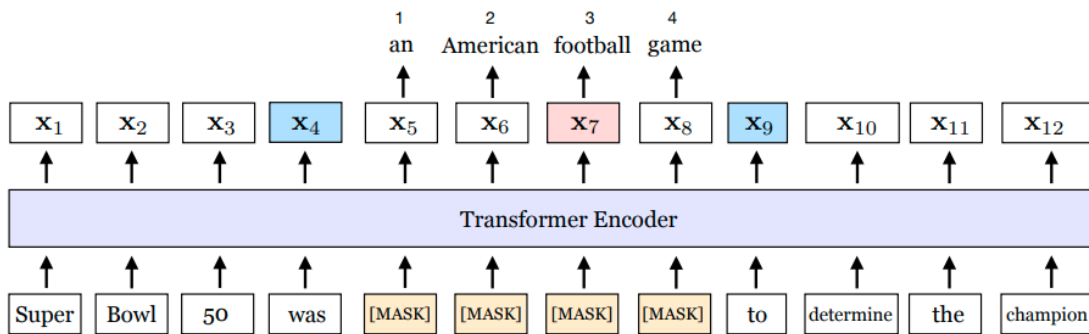


Figure 3.2. The input sequence of SpanBERT (Figure from [5]). The masked span is "an American football game".

the masked span. Figure 3.2. shows the training sequence of SpanBERT. In SBO objective, the model tends to store span information into adjacent tokens x_4 , x_9 , which can be used in finetuning phase easily. The experiments shows that SpanBert outperforms BERT on 14 out of 17 baselines. Similar to ERNIE, they performed different masking strategies which are masking the subword tokens, whole word, named entities, noun phrases and geometric spans. Masking geometric spans achieve better performance than the others but linguistic strategies like named entity and noun phrase masking yield competitive performance.

ALBERT [47] introduced two different techniques for parameter reduction that decreases the memory usage and accelerate the training. The first one is factorized embedding parameterization, which decomposes the vocabulary into two smaller matrices and this reduces the parameter size significantly. The second technique is cross-layer parameter sharing which avoids the parameter size growing with the size of the network. They also propose a new objective; sentence-order-prediction (SOB) in place of NSP objective and discuss that NSP is not a difficult task.

XLNet [48] is a new pretraining approach that leverages autoregressive (AR) and autoencoding (AE) language modeling. The proposed training objective; permutation language modeling is able to model bi-directional context by maximizing the likelihood of the input sequence. The proposed objective can be represented as follows;

$$\max_{\theta} \mathbb{E}_{\mathbf{z} \in Z_N} \left[\sum_{j=1}^N \log p_{\theta} (t_{z_j} | t_{z_1}, t_{z_2}, \dots, t_{z_{j-1}}) \right] \quad (11)$$

For the element t in the sequence, Z_t shows the permutation for that element. They sample a permutation z from Z_t permutations and the probability of the sequence is calculated with factorization according to z which is conditioned on every other tokens in sequence. This objective does not permutes the sequence order. XLNet uses the Transformer-XL [49] architecture with two-stream attention mechanism. The authors achieved significant improvements compared to the original BERT model on many datasets.

UniLM [50] is a pretrained language model which is trained with three language modeling tasks; unidirectional, bidirectional and sequence-to-sequence. UniLM can be employed for both natural language generation and understanding. They introduce a specified self-attention masks to integrate three different objectives. The text representations are jointly trained with multiple language models and yields more generic representations and prevent overfitting. UniLM obtained comparable results to BERT on GLUE scores and achieve state-of-the-art performance on five NLG datasets.

ELEKTRA [51] proposes a novel pretraining objective called replace token detection. In original word, tokens are randomly masked using the $[MASK]$ token. However, ELEKTRA uses samples created by a generator network to replace the masked token. In that way, model can learn from all input sequence instead of the masked tokens. The pre-trained generator network can also be used in the finetuning phase to enhance the input representations. They show that the proposed objective is computationally more efficient than the other BERT-based pre-training models.

Even though BERT performed well on many NLP tasks, it can not be adapted directly into language generation task since it trains only the encoder or decoder. To address this issue, MASS [52] propose a new objective called MAsked Sequence to Sequence learning (MASS) for language generation task. MASS use a encoder-decoder network based on Transformers, the encoder gets an input sequence with masked tokens and the decoder learns the predict

these masked tokens. The tokens which are not masked in encoder step is also masked in decoder step to ensure that decoder concentrate on source sequence. For unsupervised machine translation task, MASS obtained state-of-the-art BLEU scores for different languages including English to German.

3.3.2. Cross-lingual Pre-training

Most of the research in the literature focuses on monolingual pre-training, particularly for the English language. Recently, learning sentence representations for multiple languages has gained attention amount the NLP researches. The multilingual version of BERT [33] is designed by the authors¹. They used XNLI (Cross-lingual Natural Language Inference) [53] corpus for pre-training. XNLI is the extended version of NLI (Natural Language Inference) corpus. It consist 15 languages and 112,500 annotations in total. Authors also used this corpus to evaluate the performance of the sentence embeddings with different tasks including machine translation.

Most of the works in the multilingual NLP area focus on a few languages. [54] proposed universal sentence embeddings for 93 lanugages by using a single bidirectional LSTM model for all languages. The proposed model requires paralell data for training. Thus, the authors combine several parallel datasets from multiple languages. They obtained successful results for different tasks (cross-lingual classification, bitext mining) without additional fine-tuning. They also evaluate the embeddings on the XNLI dataset and achieve better performance than the XNLI baseline and Multilingual BERT for almost every language. For XNLI, they trained a classifier on top of the pre-trained LSTM encoder.

XLM [1] is proposed for cross-lingual pre-training which is based on BERT architecture. The authors introduces new unsupervised pre-training objectives for monolingual and cross-lingual pre-training and unsupervised objectives that benefits from the parallel data and enhance the cross-lingual embedding quality. For monolingual pre-training, they prepared

¹<https://github.com/google-research/bert/blob/master/multilingual.md>

Wikipedia dump and used shared vocabulary among the languages. For supervised pre-training objective, they combined different parallel datasets for multiple languages. They evaluated the pre-trained model on cross-lingual classification, unsupervised and supervised machine translation. For machine translation, they achieved state-of-the-art performance on WMT'16 German to English and Romanian to English datasets. They also evaluate the cross-lingual performance on the XNLI dataset and found that it outperforms the state-of-the-art.

3.3.3. Multimodal Pre-training

Modeling the relation between vision and natural language is a challenging problem. Because it needs to understand both visual and language contexts and capture the alignment between them. There are many successful applications in the literature that are able to understand language and vision separately. In this section, the pre-trained models which combines visual and language modalities are examined and summarized in Table 3.1..

The first attempt for multimodal pretraining is VideoBERT [55]. The proposed approach consist of three parts; BERT model for language understanding, automatic speech recognition system to convert the audio in the videos into text and vector quantization for extracting the features from video. They cluster the video clips according to features and the loss is calculated using the cluster number of the masked video token. VideoBERT can be applied for various vision-language tasks including image and captioning and also achieved the state-of-the-results for video captioning task.

ViLBERT [56] (stands for Vision Language BERT) introduce a joint model based on BERT using aligned visual and language data. ViLBERT employs separate streams for each modality unlike VideoBERT. The streams later fused with an additional co-attention layer. In this way, each modality can be processed under its specificities. Additionally, the modalities can be combined in different representation levels. They use Conceptual Captions [2] dataset for pre-training and perform finetuning with four different visual-language downstream tasks. ViLBERT outperform state-of-the-art results on various downstream tasks including Visual Question Answering (VQA) and Visual Commonsense Reasoning (VCR).

LXMERT [57] introduce a pre-training model based on transformer architecture which contains three different encoders; language encoder, visual encoder and cross-modality encoder. LXMERT employs five different pre-training objectives to understand the alignments between linguistic and visual contents. Similar to ViLBERT, different streams are used for language and vision modalities. In this way, model can deduce the masked tokens using unmasked tokens from same modality or the corresponding modality. Authors combined different datasets for pretraining including image question answering and image captioning datasets. The final pretraining data consist of 9m image-language pairs. The sentences are splitted into subwords using WordPiece tokenizer [58]. For each image, object level features are extracted using a pretrained detection network. The final object representation is obtained from the region feature and the position embeddings. For pretraining objective, they used masked cross-modality language model and masked object detection tasks. They performed finetuning for VQA with only small modifications to the network and achieved state-of-the-art overall accuracy for two VQA datasets. They also evaluated the model on visual reasoning task and showed significant increase on model generalizability.

VisualBERT [59] modified the original BERT training objectives to conform both visual and textual input. They used a large scale image captioning dataset; MSCOCO [60] for pre-training and employs two different training procedures; task agnostic pretraining and task specific pretraining. Task agnostic pretraining includes masked language modeling and sentence image prediction objectives. However, task specific pretraining only use masked language modeling to accommodate model into domain of the downstream task. VisualBERT showed powerful performance for various vision-language tasks. The ablation studies proved that proposed model's attention mechanism is able to catch information between vision and language which is interpretable.

Unicoder-VL [61] is designed to learn language and vision representations jointly using Transformer architecture. The multimodal models mentioned before calculates KL divergence between real and predicted object label distributions for the visual pre-training objective. However, Unicoder-VL directly predicts the object labels. They used pooled ROI features from pretrained Faster R-CNN network and box coordinates to encode the position

which are fed to separate fully connected layers. Lastly, the outputs of fully connected layers are summed up to produce final object embedding. They trained the model with 3.8m image/text pairs which are the combination of two vision/language datasets including Conceptual Captions. Unicoder-VL evaluate the performance of the proposed embeddings for two downstream tasks; image to text and text to image retrieval. They obtained state-of-the-art performance for both of the tasks.

VL-BERT [62] aims to generate representations for vision language for vision language tasks using a single Transformer. Images are represented as ROI features extracted from Fast R-CNN [63] which are fed to the network together with the text input. Different from mentioned models, Fast R-CNN weights are also updated in the training. They calculate the position embedding using box coordinates of the region. In addition, VL-BERT adopts segment embedding to separate the input modalities. The network is trained with large scale datasets; Conceptual Captions [2], BooksCorpus [39] and English Wikipedia data. They finetuned VL-BERT on VQA which performed better than concurrent works except LXMERT [57]. The reason is LXMERT is pre-trained with various VQA datasets. VL-BERT also shows better performance than the other works in VCR task.

UNITER (UNiversal Image-TExt Representation) [64] created multimodal embeddings trained with four different large-scale vision-language datasets which is based Transformers. They designed four different training objectives to jointly train vision and language. The difference that distinguish this model from previous is that they adopted conditional masking instead of the random masking. UNITER also benefit from a Word Region Alignment (WRA) which pushes the language vision pairs to be aligned. They demonstrated that both conditional masking and WRA strategy and improves the pre-training.

Model	Stream Type	Dataset	Pre-training Objective	Downstream Tasks
VideoBERT	single-stream	Cooking312K	- Masked Language Modeling - Masked Image RoI Prediction - Image-Sentence Alignment	- Video Captioning - Action Classification
CBT	single-stream	Cooking312K	- Masked Language Modeling - Masked Visual-Feature Regression - Sentence-Image Alignment	- Video Captioning - Action Anticipation
ViLBERT	two-stream	Conceptual Captions	- Masked Language Modeling - Masked Visual-Feature Classification - Sentence-Image Alignment	- Visual Question Answering - Image Retrieval - Visual Commonsense Reasoning - Grounding Referring Expressions
LXMERT	two-stream	MS COCO VQA v2.0 GQA VG-QA	- Masked Language Modeling - Masked Visual-Feature Classification - Masked Visual-Feature Regression - Sentence-Image Alignment	- Visual Question Answering - Natural Language Visual Reasoning
VisualBERT	single-stream	MS COCO	- Masked Language Modeling - Sentence-Image Alignment	- Visual Question Answering - Natural Language Visual Reasoning - Visual Commonsense Reasoning
Unicoder-VL	single-stream	Conceptual Captions	- Masked Language Modeling - Masked Visual-Feature Classification - Sentence-Image Alignment	- Image-Text Retrieval
VL-BERT	single-stream	Conceptual Captions Book Corpus Wikipedia (English)	- Masked Language Modeling - Masked Visual-Feature Classification	- Visual Question Answering - Grounding Referring Expressions - Visual Commonsense Reasoning
UNITER	single-stream	Conceptual Captions Visual Genome MS COCO SBU Captions	- Masked Language Modeling - Masked Region Modeling - Word Region Alignment - Image-Text Matching	- Visual Question Answering - Natural Language Visual Reasoning - Image-Text Retrieval - Visual Commonsense Reasoning

Table 3.1. The summarization of the models proposed for multimodal pretraining

4. Cross-lingual Multimodal Pretraining

4.1. Textual Representation

The first step for developing a pre-training network is constructing the vocabulary from training data. The regular vocabulary consists of the unique words in the dataset, but the size of the vocabulary can be gigantic depending on the training data. Furthermore, for the vocabulary containing the most frequent words, the words that do not exist in vocabulary are represented as unknown words. Recently, almost every NLP application has built vocabulary from smaller units such as subwords to address these problems.

One of the most popular approaches for subword creation is Byte Pair Encoding (BPE) [65]. The BPE model uses an iterative method to create a subset of words in the training set. The words in the education set are first separated into the characters that are the smallest part of a word. Each parsed character is considered a symbol, and the co-occurrence of these binary symbols is calculated. A symbol that has two characters is created by combining the two most common characters. This merging process is repeated iteratively for all symbols until a certain number of sub words are obtained. The main purpose of this process is to finally convert the most common character n-grams in the training set into a single symbol, or in other words, a subword.

Another commonly used approach in subword tokenization is called SentencePiece (SPM) [66] which has been specifically developed for text processing. SPM is a deterministic approach which is based on both BPE approach and unigram language modeling. Unlike BPE, SPM can work directly on raw data that are not tokenized or processed at all. The first step of the SPM approach is transforming all characters in the input data into unicode which eliminates the language dependency.

We used BPE approach to extract subwords and created a shared vocabulary from Conceptual Captions using both English and German sentences and limit the vocabulary to 50k subwords. We also used the same subword dictionary to tokenize Multi30k [3] that is used in finetuning phase.

4.2. Object Representation

It is crucial to obtain good representation of the image in order to learn rich multimodal embeddings. Most of the recent works have applied a pre-trained object detection network to extract region representations.

LXMERT [57] extracts 2048 dimensional ROI features from a Faster R-CNN [67] network that is pretrained on Visual Genome Dataset [68]. VL-BERT [62] follows the same approach, but instead of a feature extractor, it also updates the parameter of the Faster R-CNN network during pre-training. ViLBERT [56] obtains mean-pooled convolutional features from the

same Faster R-CNN network. For selecting the ROIs, the authors defined a threshold value for the proposal scores and selected 10 to 36 region proposals.

In a similar way to the literature, we extracted the image features using a Faster R-CNN network that is pre-trained on Open Images dataset [69]. We also performed mean pooling to convolution features resulting in 1586 dimensional ROI embedding which is projected into the embedding size of the network with a linear layer. For the sake of simplicity, we selected 30 ROI features for each image that had the highest scores.

Object features are projected into embedding size and fed into the network as visual words. We also applied a special regional encoding strategy to encode the order of object proposals. For this purpose, bounding box coordinates of the objects are also projected into the embedding size. At the end, bounding box coordinates and object features are summed to obtain final object representation.

4.3. Model

In this section, we proposed a cross-lingual multimodal to create general language and visual embeddings that can be used in not only language/vision problems including image captioning, multimodal machine translation but also other translation-related tasks such as machine translation. In addition, word embeddings obtained with the help of visual information carry richer information than those obtained using only textual information.

We developed a cross-lingual multimodal model based on XLM [1] that is cross-lingual model trained in both a supervised and unsupervised manner. XLM employs three different learning strategies. The first approach is termed Casual Language Modeling that is based on typical language modeling for monolingual setting. The authors also employ Masked Language Modeling (MLM) trained on multiple languages in a streaming manner. All words in the input data are treated as a single stream with a length 256. During the training, 15% of the text stream is sampled and 80% of the tokens in the stream are switched with masked token [*MASK*].

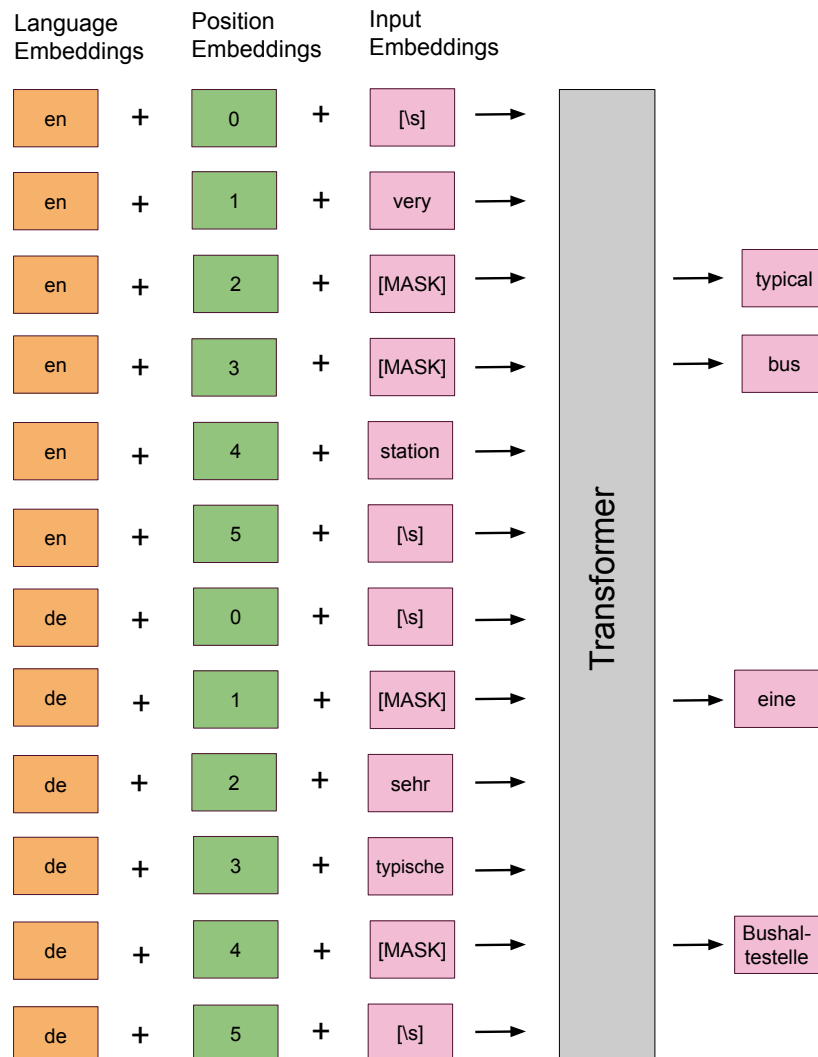


Figure 4.1. The TLM objective of XLM model. The words are randomly masked for both languages.

The final objective, which is the inspiration of this work is Translation Language Modeling (TLM) illustrated in Figure 4.1.. In this approach, text streams from different languages are concatenated into a single stream. TLM leverages parallel data, in contrast to other previously mentioned objectives. In TLM, a fixed portion of the both languages is randomly masked during training and promotes the alignment between two languages while predicting the masked token. In other words, the model learns to attend to German language tokens while predicting a token in English. Inspired from XLM, we developed a training objective called Visual Translation Language Modeling (vTLM) to learn rich cross-lingual multimodal

representations.

4.3.1. Visual Translation Language Modeling (vTLM)

TLM objective applies parallel pairs from different languages to learn cross-lingual representations. Visual Translation Language Modeling (vTLM) extended the TLM objective by adding an another modality; image. vTLM simply includes the visual input together with the parallel data which is illustrated in Figure 4.2.. Each image in the input data is represented with ROI features that can also be interpreted as visual words.

For masking visual words, VL-BERT [62] and LXMERT [57] zero out 15% of the ROI features. On the other hand, vTLM masks 15% percent of the ROI features and are replaced with the [MASK] token. In pre-training, vTLM uses the same [MASK] token for both visual and textual tokens. 15% of the textual tokens are also masked randomly.

For the textual part of the input stream, vTLM follows the MLM objective where the model learns to predict masked tokens. In the TLM objective, a softmax classifier with cross-entropy loss is used for masked token prediction. vTLM also follows the same approach for the textual part of the input. Similar to this idea, vTLM predicts the object label for masked ROI feature and calculates cross-entropy loss for label prediction task. In other words, vTLM employs two separate cross-entropy losses for visual and textual parts of the input stream. For final loss, the outputs of these cross-entropy loss functions are summed together.

4.4. Pre-training Settings

To train our model, we used a small version of the original XLM [1] due to an insufficiency of computational power. We set the embedding dimension to 512, which is half of the original XLM setting. We used 6 layers and 8 attention heads with the Adam optimizer. We trained our network with batch size of 64 and learning rate 0.0001. The network is trained with Conceptual Captions [2] dataset which is originally contains 3.3m image/caption pairs.

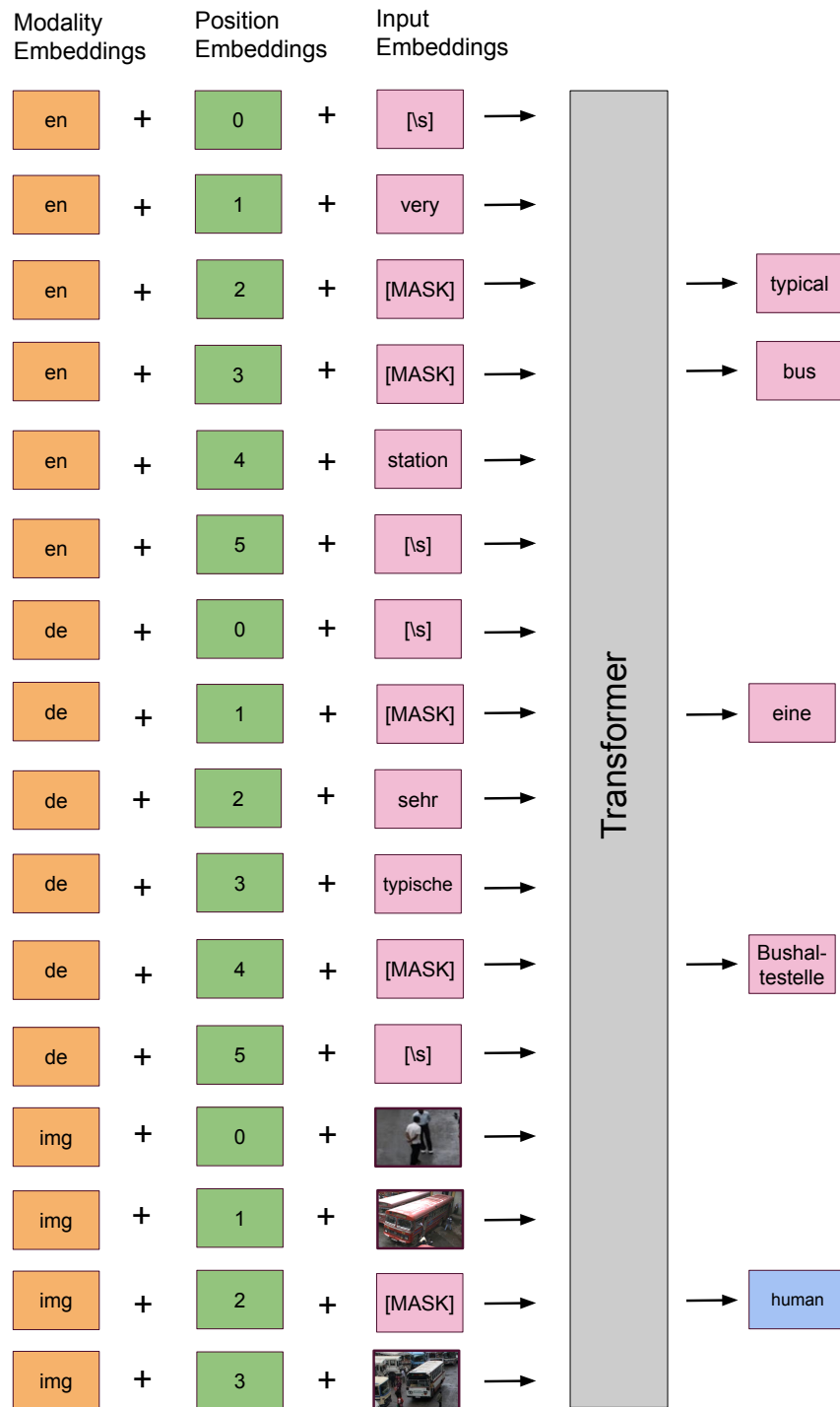


Figure 4.2. The proposed vTLM objective.

However, some of the image links were broken or cannot be downloaded. Therefore, the final dataset contains 3.1m pairs. We also set the epoch size to 300k which are sampled from the entire dataset. For early stopping, we end the training if validation accuracy of MLM is not changed in 25 epochs. Whereas most of the mentioned works in the literature such as ViBERT [56] initialize the network with pre-trained BERT embeddings, we initialize our network with random weights.

For comparison, pre-training was carried out with the TLM objective of XLM model with the same training configuration of the vTLM objective. English/German automatic translation pairs of the Conceptual Captions dataset are used for pre-training the XLM model.

4.5. Downstream Tasks

4.5.1. Machine Translation

vTLM and TLM models trained with the same configuration are finetuned for the machine translation task on the Multi30k dataset. The purpose of choosing this task is to observe that word representations created using visual information are more successful and semantically richer than those created using only textual information. To adapt the proposed model to the MT problem, the pre-training encoder was also used as a decoder. We initialized both encoder and decoder weights from pre-trained encoder. In the original adaption of XLM model to the MT task, the attention layers between the encoder and the decoder were randomly initialized. However, we realized that random initialization of these layers affected the finetuning performance badly and the finetuned model failed to produce meaningful translations in early epochs. For this reason, unlike the original XLM model, we set the weights of attention layers between the encoder and decoder using the encoder attention weights, and thus, the model started to produce correct translations even in the first epochs during the finetuning phase. We used a dropout of 0.2, an attention dropout of 0.1 and a learning rate of 0.0001. We finetuned both models to maximum 80 epochs and performed early-stopping according to the validation BLEU score.

4.5.2. Multimodal Machine Translation

We evaluated the performance of the vTLM model on MMT task. We used the same approach with MT only with the addition of object proposals. For input sequence, we concatenated the input sentence and object features extracted from the Faster-RCNN [68] network. We used the top 36 object proposals which is the same as for the pre-training. We used the same initialization on parameter set with the MT experiments.

5. Results and Analysis

We performed pre-training experiments for both TLM and vTLM objectives and evaluated the performance of pretrained models with two different downstream tasks: Machine Translation and Multimodal Machine Translation. We used the same setting for all experiments to compare different pre-training strategies. In addition to traditional performance metrics such as BLEU, we also used MLT accuracy which measures the correctness of the translated ambiguous words.

5.1. Datasets

5.1.1. Conceptual Captions

Conceptual Captions [2] is a large-scale image/language dataset collected using alt-text descriptions of the images on the Internet. The final dataset contains 3.3m image/description pairs in total. The authors build an automated process to create nice and clean captions from alt-text descriptions. This process employs a pipeline proposed by [70], which consists of extract, filter and process steps. In contrast to the MSCOCO [60], Conceptual Captions contains data from various sources since they are collected on the Internet. Raw descriptions contain many people/location names and it is, therefore, more difficult to learn captioning model. The authors used Google’s Natural Language APIs to locate named-entities and they replaced these with corresponding hypernym words using Google’s Knowledge Graph API.

They also removed numbers, dates, and units from descriptions and created a clean, learnable captions. Example descriptions are shown in Table 5.1..



Alt-text	Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.	A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.
Conceptual Captions	pop artist performs at the festival in a city.	a worker helps to clear the debris.

Figure 5.1. Samples taken from the Conceptual Captions dataset. Alt-text descriptions are the raw descriptions collected from Internet. Conceptual Captions descriptions are clean and fluent.

There are several large-scale image captioning and visual question answering datasets in the literature, although all of these datasets specialized for one language, particularly English. In this work, we extended the Conceptual Captions to another language— German. The reason behind the choice of the German language is that the recent translation systems have shown great success for English to German translation task. English and German belong to the same language family and share the same grammatical rules which makes it easier to translate one to another. For translation, we used fairseq [9] toolkit with a pretrained translation model proposed by [71] which is the best performing network on the WMT19 English-German translation task even outperforms human performance. The example translations can be seen in Figure 5.2..

Each caption in Conceptual Captions is translated into German language to develop a large-scale multimodal multilingual dataset. The German translations are not preprocessed or validated; they are exactly used as they are obtained from the translation model. Therefore,

the English-German translations are not perfectly aligned with each other and may contain translation errors.

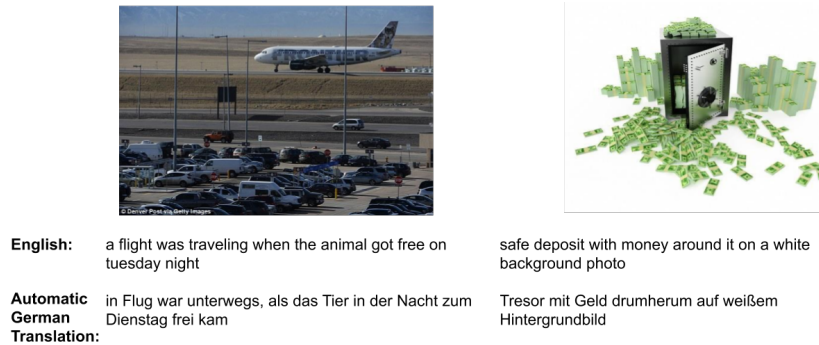


Figure 5.2. Example sentences from Conceptual Captions [2] and their automatic translations

5.1.2. Multi30K

Multi30K [3] is a multimodal machine translation dataset originated from Flickr30K [20]. The Flickr30k descriptions are collected with crowd-sourcing process for each image and contains 31k image/caption pairs. Human translators translated each description in the Flickr30K dataset into German in order to develop a parallel dataset. Human translators did not see any image related to the description. Later on, Multi30k is expanded with other languages; French and Czech. Different from German descriptions, images corresponding to the the source description are shown to French [72] and Czech [73] translators which yields image-aware translations.

5.2. Evaluation Metrics

Human evaluations of MT systems show several prospects of translation, including sufficiency, constancy, and eloquence of the translation [74] but most of the time human evaluation approaches are rather costly [74]. Furthermore, these processes can take significant amount of time to be completed which creates the need for automatic evaluation of these systems. To solve this problem, several methods such as BLEU [75], METEOR [76] has



- English:** a group of men are loading cotton onto a truck
- German:** un groupe d'hommes chargent du coton dans un camion
- French:** eine gruppe von männern lädt baumwolle auf einen lastwagen
- Czech:** skupina mužů nakládá bavlnu na nákladňák .

Figure 5.3. Example captions from Multi30k dataset [3].

been proposed. In this section, we will give details about the evaluation metrics that we used to evaluate our MT and MMT systems.

5.2.1. BLEU

Bilingual Evaluation Understudy (BLEU) [75] is the most popular MT performance metric that propose distinguishes a good quality translation. This evaluation metric can be used in many systems evaluated by comparing the source and target text data such as video/image captioning, MT, MMT and question answering.

The initial task for a BLEU is to contrast unigram or n-grams of the competitor with the n-grams of the translation of reference and compute the number of matches which are position-independent. An increased number of matches shows that the candidate translation is more similar that the reference and better alignment.

$$\log BLEU = \min \left(1 - \frac{r}{h}, 0 \right) + \sum_{n=1}^N w_n \log p_n \quad (12)$$

The calculation of the BLEU score is formulated in Equation 12 where N is the n-gram number, h is the hypothesis sentence's length, r is the length of the reference corpus and w_n stands for the positive weights that are summed to 1.

5.2.2. METEOR

METEOR is an automatic evaluation metric that is predicated on an approach of unigram matching between the machine generated translations and reference translations. When matching unigrams of the alignment of the candidate and reference sentences, could be based on their surface, stemmed, and sentiment forms; besides, METEOR could be enlarge to comprise more advanced matching techniques. [76]. Essentially, METEOR calculates harmonic mean of recall and precision of the uni-grams matches. Recall is formulated as $R = m_{rh}/r_{total}$ and precision is $P = m_{rh}/h_{total}$ where m_{rh} is the count of matches between reference and hypothesis sentence, r_{total} is the count of unigrams in the reference set and h_{total} in the hypotheses set. To calculate n-gram matches, METEOR employs a penalty term p that groups the unigram matches to obtain longer matches. The final METEOR calculation is formulated in Equation 13.

$$M = \frac{10PR}{R + 9P}(1 - p) \quad (13)$$

5.2.3. Multimodal Lexical Translation Accuracy

Multimodal Lexical Translation (MLT) accuracy is a evaluation metric for the multimodal machine translation task. This method determines whether the ambiguous words in sentences are translated correctly. [6] identified ambiguous words in the Multi30k [3] dataset for this problem, and the translations corresponding to ambiguous words were performed by

Visual Content:



Ambiguous Word: big

Lexical Translation: große

Textual content: girl watching a
big wave heading towards her

Figure 5.4. An example from MLT [6] dataset

humans. The authors discovered that 1108 different English words are ambiguous in German or French. These words are in many sentences and in a total of 98,647 ambiguous word / sentence pairs. An example data from MLT dataset is shown in Figure 5.4..

For input word x and translation system T , the proposed MLT accuracy metric searches for the correct translation y of the ambiguous word in the output $T(x)$ from the translation system. Finally, the MLT accuracy is calculated by counting how many times the system predicted ambiguous words correctly. This metric is used not only to evaluate MMT performance, but also to the assess of standard MT systems.

5.3. Quantitative Results

We trained the XLM architecture for both TLM and vTLM objectives under same setting. In order to compare the performance of these models, we first performed finetuning on machine translation task. In this experiment, we investigated whether the word embeddings learned by

including visual information are richer than those learned by using the traditional language model.

In the pre-training phase, we used same parameter set for both TLM and MLM objective and validation MLM accuracy is used for early stopping. During the experiments, we realized that the word embedding obtained from the checkpoints where the validation MLM was highest performed poorly in finetuning. For this reason, we conducted our experiments with checkpoints that we received from an intermediate point where both models reached the similar MLM accuracy.

In Table 5.1., the experimental results are shown for downstream tasks. We performed machine translation experiments for both XLM-TLM (textual) and XLM-vTLM (textual + visual) models to observe whether word representations enriched by using visual information are of higher quality than those using only textual data. In machine translation experiments in Table 5.1., we showed that the model which is initialized using the weights of the visual-based XLM-vTLM model, obtain better results than the XLM-TLM initialization in the BLEU and METEOR metrics. We also achieved state-of-the-art results for MT and MMT tasks on Multi30k test2016 test set. We obtained 41.55 and 60.1 in BLEU and METEOR respectively for MMT tasks which improves the MT performance of textually-grounded model (XLM-TLM). This also shows that the image information is included in the model yields better translations. MT system initialized by our visually-grounded representations (XLM-vTLM) perform better than even existing MMT systems which includes visual content in the model.

We also evaluate the performance of the MT and MMT system using MLT accuracy [6] which a new evaluation metric measures the correct translation of ambiguous words. Experimental results showed that the achievements in the MLT accuracy are in contrast to BLEU and METEOR. The model initialized with the TLM weights obtained the highest accuracy. We investigated the dataset to understand the reason why TLM performs better in this metric and realized that the dataset [6] we used to calculate MLT accuracy involves only Multi30k vocabulary. When we examined the translations created by vTLM which are

Downstream Task	Model	BLEU	METEOR
Machine Translation	Doubly-att (TF) [77]	38.8	56.8
	nmtpy (RNN) [78]	38.9	58.4
	XLM-TLM (ours)	41.1	59.5
	XLM-vTLM (ours)	41.55	59.8
Multimodal Machine Translation	Trg-mul (RNN) [21]	37.8	57.7
	VMMT (RNN) [79]	37.5	56.0
	Deliberation Network (TF) [80]	38.0	55.6
	Graph-based (TF) [81]	39.8	57.6
	BN + Enc. Attention (RNN) [82]	40.5	57.9
	XLM-vTLM (ours)	41.8	60.1

Table 5.1. Experimental results of our models and recent state-of-the-art models on Multi30k test2016 dataset. There is no MT and MMT systems in the literature employs pre-training and fine-tuning on Multi30k dataset. Thus, the systems shared here does not use pre-training.

marked wrong, we found that it was not actually wrong. For source sentence "a motocross race with a lot of mud", the MT system based on vTLM generates the following translation; "einen motocrossennen mit viel schlamm" which is correct. However, MLT dataset searches the "matsch" word for lexical German translation of "race". In generated translation, "motocrossennen" also means "race" in English but MLT accuracy tagged this translation as incorrect. The reason is the MLT dataset does not contain similar words and dependent on the vocabulary of the Multi30k dataset. Therefore, MLT accuracy does not provide clear evaluation. For future work, we plan to extend the MLT dataset by using a similar words dictionary.

	Model	MLT Accuracy
Machine Translation	XLM-TLM	79.14
	XLM-vTLM	71.77
Multimodal Machine Translation	XLM-vTLM	72.22

Table 5.2. Experimental results on Multi30k test2017 dataset

5.4. Qualitative Results

In Table 5.3., we present the example English-German translation results for both TLM and vTLM models. The first column shows the English input sentence, and the second

and third columns show the English translations produced by TLM and vTLM models, respectively. Below each translation, there is an explanation translated back to English using Google Translate. In these results, samples differentiated by TLM and vTLM models were selected. In the selected examples, the TLM model failed to translate every word correctly, and there are missing words in the translations. In the translations obtained by the vTLM model, missing words are corrected, and some words are translated more accurately. We also demonstrated additional examples to compare MT and MMT systems 2.3.. We shared the English translation (ET) which is obtained using Google Translate. MT results are obtained without any visual information.

For selecting the weights for our pre-trained models, we first selected the best checkpoints according to the MLM accuracy. We realized that best checkpoints obtained from pre-trained models performs dramatically bad in finetuning. We observed that validation BLEU scores does not improve over 12 epochs while finetuning MT model which is initialized by the best checkpoint obtained from pre-trained TLM model. For this reason, we decided to use checkpoints taken from the intermediate steps of the pre-trained models in the finetuning experiments.

In Figure 5.5., we shared the BLEU scores for TLM and vTLM models on Multi30k validation set for each epoch in finetuning the MT system. We selected the checkpoints where both models obtained similar MLM accuracy, which is 80. Unlike the finetuning experiments with the best checkpoint of the pre-trained models, we observed that model initialized with vTLM showed better performance and obtained 26 BLEU on validation set even in first epoch finetuning which is almost 5 point higher than TLM. However, we noticed that vTLM converges slower than the TLM based model.

To examine the attention mechanism of the vTLM model, we extracted the attention weights in the pre-training phase which are shown in Figure 5.7., Figure 5.8.. For simplicity, we truncated the English sentence from the input sequence and only fed the German sentence and object regions as input the network and used the last Transformer layer.

Source Sentence	Translation	
	TLM	vTLM
a woman with brown hair sitting on a bench outside a cafe	eine frau mit braunen haaren sitzt drauen auf einer bank	eine braunhaarige frau sitzt drauen vor einem cafe auf einer bank
	ET: a woman with brown hair is sitting outside on a bench	ET: a brown-haired woman is sitting on a bench outside a cafe.
four people relaxing on a grassy hill overlooking a rocky valley	vier leute entspannen sich auf einem grasbewachsenen hugel mit blick auf ein steintal	vier personen entspannen sich auf einem grasbewachsenen hugel mit blick auf ein felsigen tal
	ET: four people relax on a grassy hill overlooking a stone valley	ET: four people relax on a grassyhill overlooking a rocky valley
a boy wearing red and white swimming trunks diving backwards in a beautiful pool	ein junge mit rot-weien badehose springt ruckwarts in einem wunderschonen pool ruckwarts	ein junge in einer rot-weien badehose springt ruckwarts in ein schones schwimmbecken
	ET: a boy in red and white swimming trunks jumps backwards in a beautiful pool	ET: a boy in red and white swimming trunks jumps backwards into a beautiful swimming pool.
a female police officer in a cap and navy uniform smiles while wearing sunglasses outside of a shop	eine polizistin mit kopfbedeckungund marineuniform lachelt lachelnd vor einem geschaft und lachelt	eine polizistin mit mutze und marineuniform lachelt wahrend sie eine sonnenbrille vor einem geschaft tragt
	ET: a policewoman with a hat and a marine uniform smiles and smiles in front of a shop	ET: a policewoman with a hat and navy uniform smiles while wearing sunglasses in front of a shop
a man dressed in black leather and a cowboy hat is walking around a renaissance festival	ein mann in schwarzer lederkleidung und cowboyhut lauft um ein renaissancestuck	ein mann in schwarzer lederkleidung und cowboyhut geht um ein renaissance-fest
	ET: a man in black leather clothes and a cowboy hat walks around a renaissance piece	ET: a man in black leather clothes and cowboy hat goes to a renaissance festival
a woman acts out a dramatic scene in public behind yellow caution tape	eine frau macht in der offentlichkeit hinter gelben warnbandern eine dramatisch szene	eine frau fuhr in der offentlichkeit eine dramatischer szene in einem offentlichen hinter gelbem absperband auf
	ET: a woman makes a dramatic scene in public behind yellow warning bands	ET: a woman performs a dramatic scene in public on a public behind a yellow barrier tape
two people are silhouetted against a lake reflecting a painted sky	zwei personen stehen gegen einen see und spiegeln einen bemalten himmel	zwei menschen stehen sich als silhouette an einen see und spiegelt sich einen bemalten himme
	ET: two people stand against a lake and reflect a painted sky	ET: two people stand in silhouette at a lake and a painted sky is reflected

Table 5.3. Example translations for produced from MT systems using TLM and vTLM models. We back-translated each translation to English using Google Translate which is shown under the translation (ET).

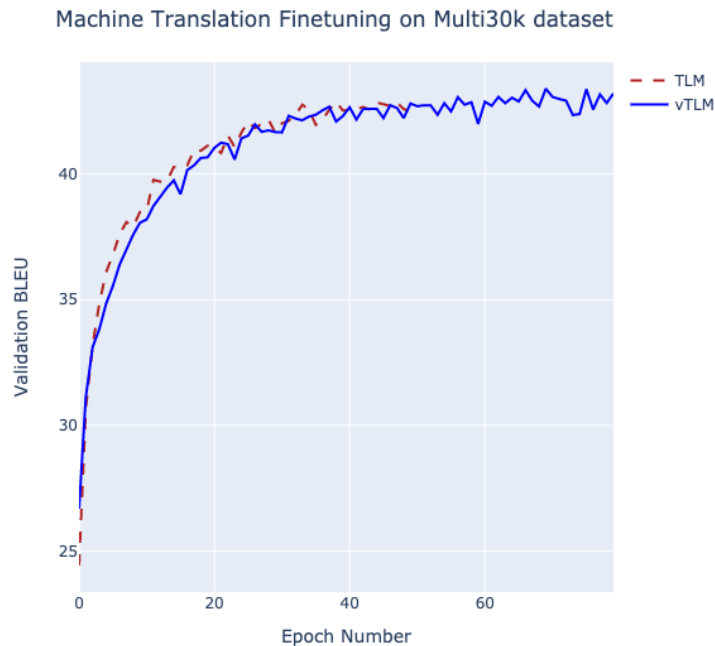


Figure 5.5. We showed the BLEU score on Multi30k validation set while finetuning for TLM and vTLM models

In Figure 5.7., we demonstrated the attention weights of the German sentence "ein junge springt auf seinen fußball spielenden bruder" and corresponding object proposals for each head. When we examine the heads with the masked token "fußball", we observed that majority of the heads are focused on random tokens and are not indicative. However, when looking at Head 6, we realized that the model directly attends to the object associated with the masked word "fußball". Also in Head 5, the model tends to attend to sport-related object regions such as football, sports-uniform and clothing. In this matter, [83] conducted various experiments to prove that the test performance does not significantly drops when some of the heads are removed. The authors also demonstrated that even some of the heads can be represented only one head. Therefore, most of the heads are redundant in our case.

In Figure 5.7., we show another attention visualization for the German sentence "ein junges mädchen versucht, eine ziege zu bürsten". We masked the word "ziege" which means "goat" into English and we examine whether model attends to animal-related objects or not. We




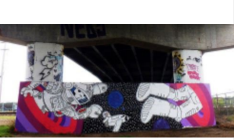

Image	Source Sentence	Ground Truth	MT output (TLM)	MT output (vTLM)	MMT (vTLM)
	four red pomegranates hang from a tree	vier rote granatäpfel hängen an einem baum	vier rote schleifen hängen von einem baum	vier rote helfer hängen von einem baum herab	vier rote äpfel hängen von einem baum
			ET: four red bows hangs from a tree	ET: four red helpers hang from a tree	four red apples hang from a tree
	a snow-covered mountain viewed from afar	ein schneebedeckter berg aus der ferne gesehen	ein schneebedeckter berg von weiter ferne betrachtet	ein schneebedeckten berg von weiter ferne	ein schneebedeckten berg betrachtet ferne
			ET: a snowy mountain viewed from afar	ET: a snow-capped mountain from afar	a snow-capped mountain looks afar
	a man is casting his line to fish	ein mann wirft zum angeln seine leine aus	ein mann wirft seine leine zum fischen	ein mann stellt seine leine zum fischen an	ein mann wirft seine leine zum fisch
			ET: a man throws his line for fishing	ET: a man puts his line on for fishing	a man throws his line to the fish
	graffiti of cartoon space scene on bridge underpass	graffiti einer cartoon-raumfahrtszene unter einer brückenunterführung	graffitis von cartoon-raumszene auf einer brücke	graffitis in einem cartoon-raum auf einer brücke aus brücke	graffitis im cartoon-raum auf einer unterführung
			ET: graffiti from cartoon room scene on a bridge.	ET: graffiti in a cartoon room on a bridge made of bridge	ET: graffiti in cartoon room on an underpass
	a man watches four defenseless cats	ein mann beobachtet vier wehrlose katzen .	ein mann beobachtet vier passanten	ein mann beobachtet vier lagerkatzen	ein mann beobachtet vier katzen
			ET: a man observes four passers-by	ET: a man watches four camp cats	ET: a man watches four cats

Figure 5.6. Example translations from MT and MMT systems for Multi30k test2017 dataset

observe that Head 3 has significantly high attention weights on "mule", "goat" and "dog" objects. Also in Head 8, the attention weights are higher in animal and human-related objects. Similar with previous example, most of the heads does not provide meaningful attention weights.

6. Conclusion

In this thesis, a new approach was developed for cross-lingual multimodal pre-training. We translated the large-scale image captioning dataset; Conceptual Captions [2] into German automatically using a state-of-the-art English to German Machine Translation model [71],

because a large-scale multilingual dataset was required during the pre-training phase of this model.

Our proposed model called Visual Translation Language Modeling (vTLM) is based on a cross-lingual model XLM [1]. XLM employs Transformer architecture, similar to other BERT-based [33] models. XLM proposed a pre-training objective called Translation Language Modeling (TLM) that takes sentences from multiple languages as an input and trains cross-lingual word embeddings. We extend the TLM objective by adding the object features extracted from corresponding images. Each image is represented by object proposals that can be interpreted as visual words. We extracted object proposals using a Masked Faster-RCNN [68] trained on OpenImages [69] dataset which consist of 600 object classes.

We trained XLM model for both TLM and vTLM objectives and compared the performance on two downstream tasks; MT and MMT. We used the Multi30k [3] dataset for downstream tasks which is an extension of Flickr30k [20] dataset with multiple languages. We showed that visually-grounded word representations (vTLM) performed better than the word representations only pre-trained with the textual data (TLM) in BLEU and METEOR. Experimental results showed that the proposed model outperforms all of the existing MT and MMT systems. Our MT system which is initialized with visually-grounded word representations (vTLM) also outperforms the existing MMT systems that directly used the visual content.

We compared the example translations of MT systems that are initialized with TLM and vTLM models' weights (Table 5.1.). It has been observed that the descriptions produced by the vTLM model contain more accurate words than those produced by TLM, and some words ignored by TLM are included in the vTLM translation output. In Appendix A, we also shared translation examples for both MT and MMT systems. In some examples, although MT systems using both TLM and vTLM are insufficient to make accurate translation, the MMT system has been able to produce correct explanation because it benefits from visual information directly.

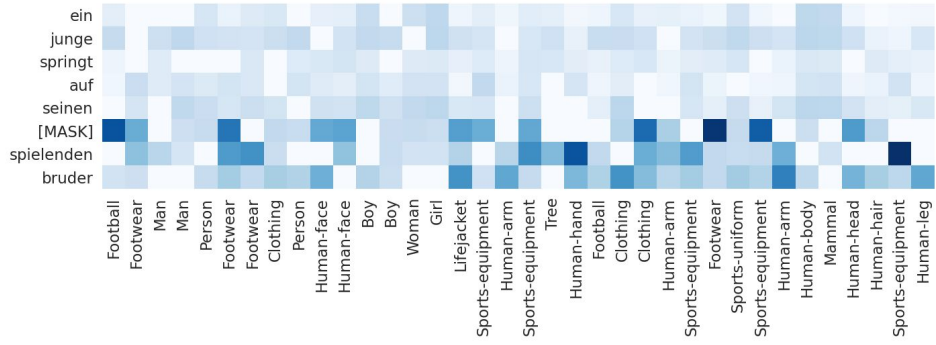
We also demonstrated that MMT improves the MT performance in BLEU and METEOR which confirms adding visual information increase the translation quality. Additionally, we

evaluated MT and MMT models using MLT accuracy which are in contrast to other reported evaluation metrics. When we analyzed the translations and ground truth, we observed that MLT dataset is created based on Multi30k dataset which is significantly smaller than the Conceptual Captions and most of the translated words does not exist in Multi30k vocabulary. In future work, we will expand the MLT dataset using a similar words dictionary and hopefully obtain more relevant results.

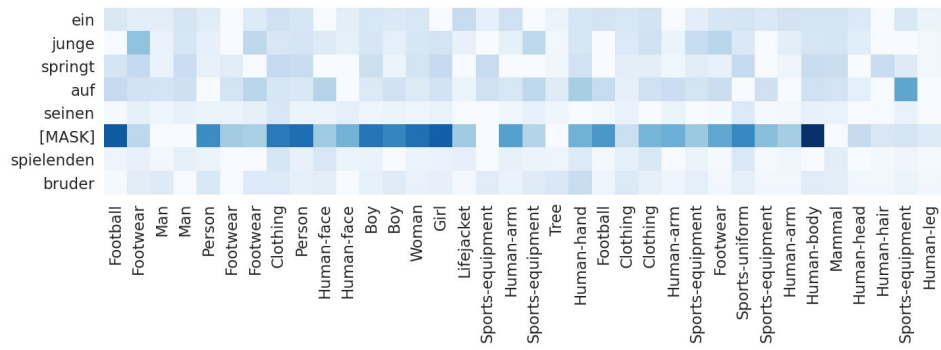
We visualized the attention weights for each head to demonstrate whether the proposed model focused on the relevant object regions. We observed that most of the heads operate randomly and are not indicative but some heads learn to attend related object regions. We also showed the MT finetuning performance in each epoch according to BLEU metric for both models initialized with TLM and vTLM. Accordingly, it was observed that the vTLM-based model not only achieved higher final BLEU score, but also performed significantly better than the TLM-based model in the early stages of the finetuning phase. Thus, we concluded that visually grounded word representations are more successful than textual ones even before finetuning.

As future work, we began to develop a new masking strategy to improve model's learning capacity. This strategy essentially masks the word corresponding to the relevant object region, forcing the model to attend that object region when generating the masked word. We assume that this strategy will improve the performance in downstream tasks and results richer word embeddings. We also plan to work on probing tasks to understand how model operates and how it uses the visual information.

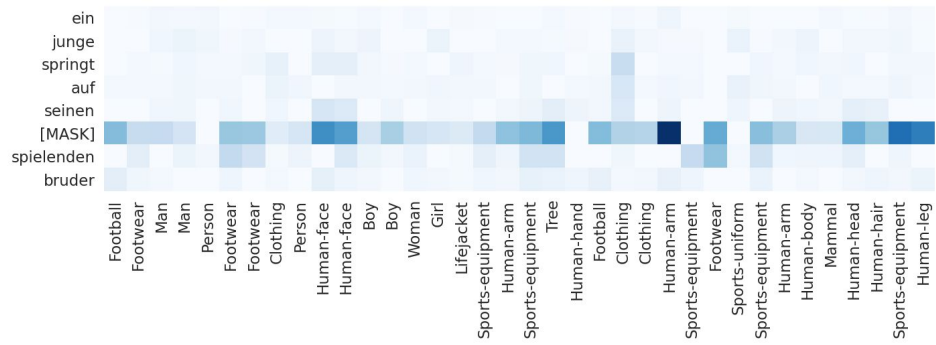
Head #1



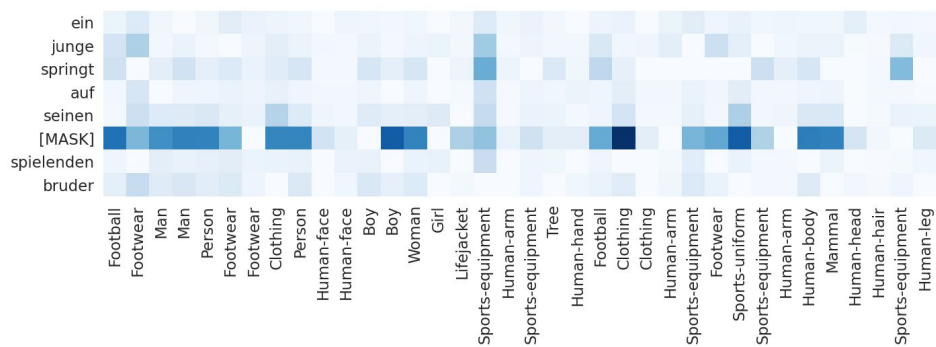
Head #2



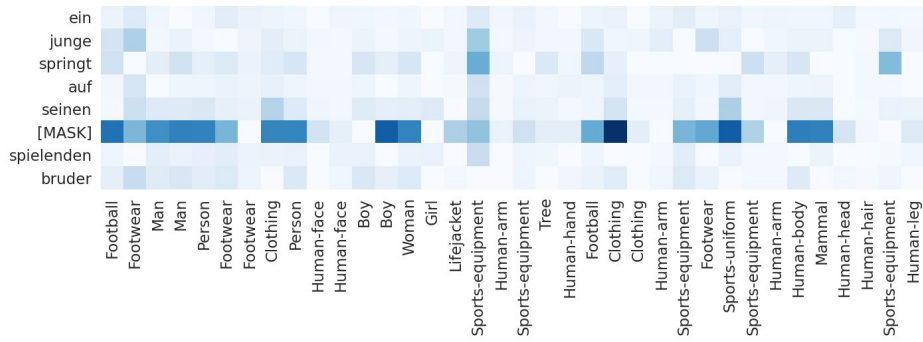
Head #3



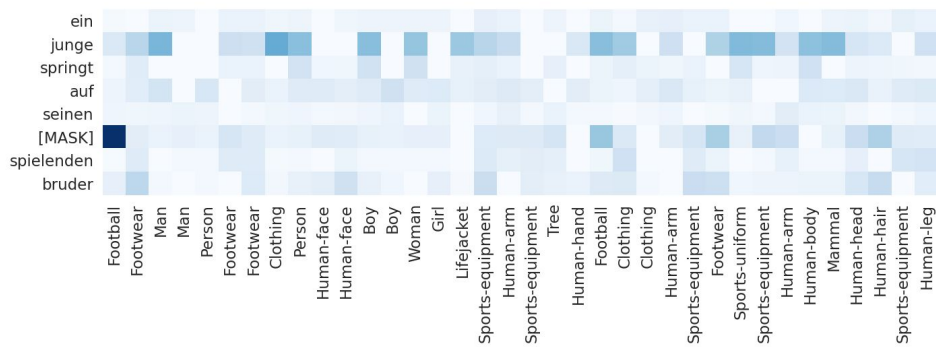
Head #4



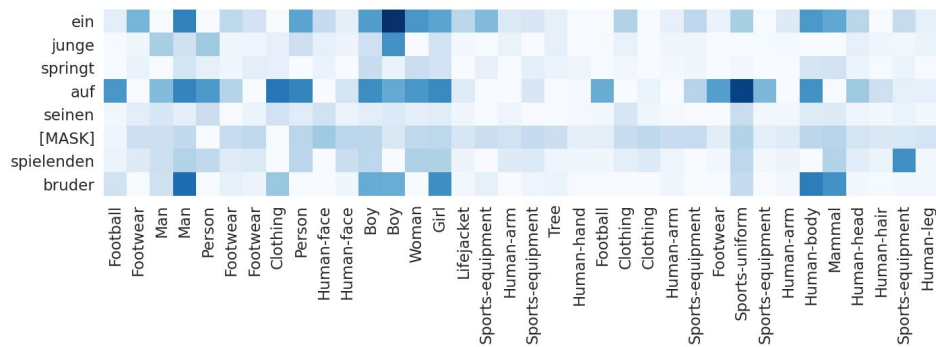
Head #5



Head #6



Head #7



Head #8

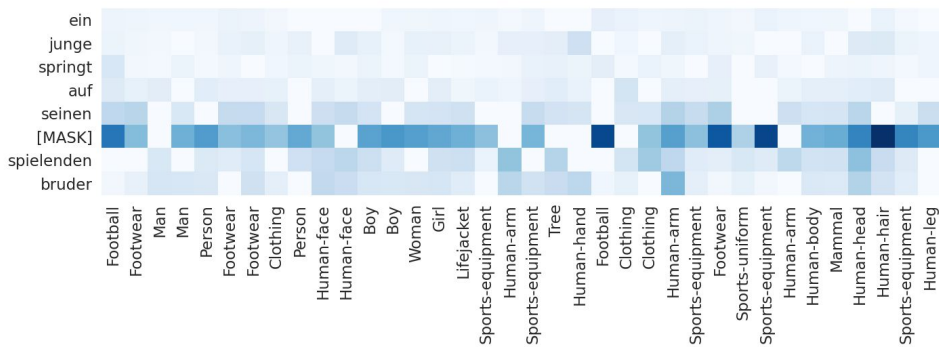
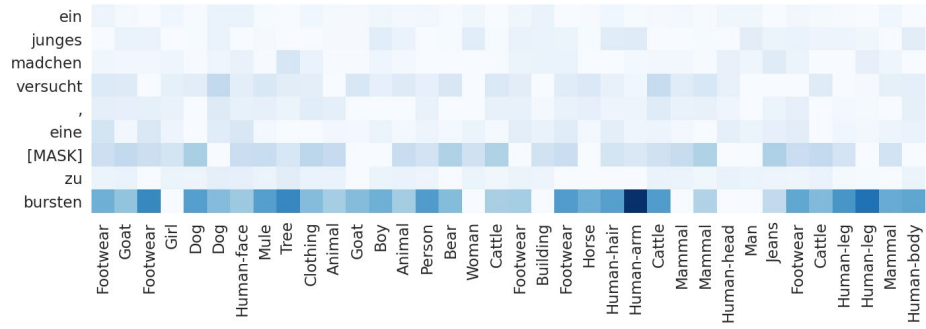
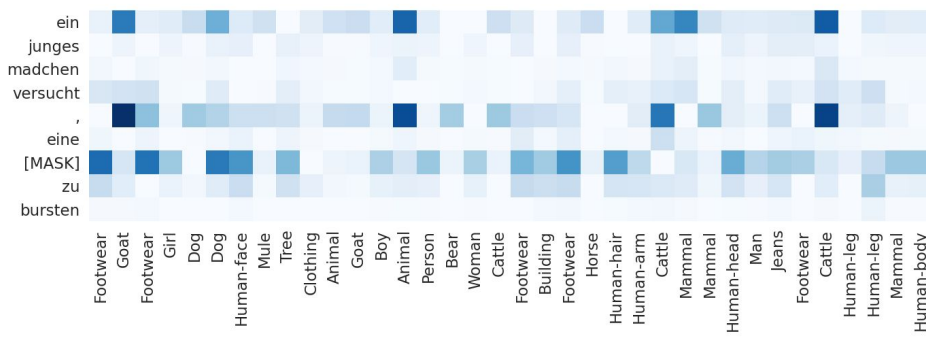


Figure 5.7. We visualize the attention weights for each head in pre-training for a German sentence and object regions. The rows contains German tokens and columns are named as the object label. We replaced the token "fußball" with [MASK] token to observe model's attention on visual input.

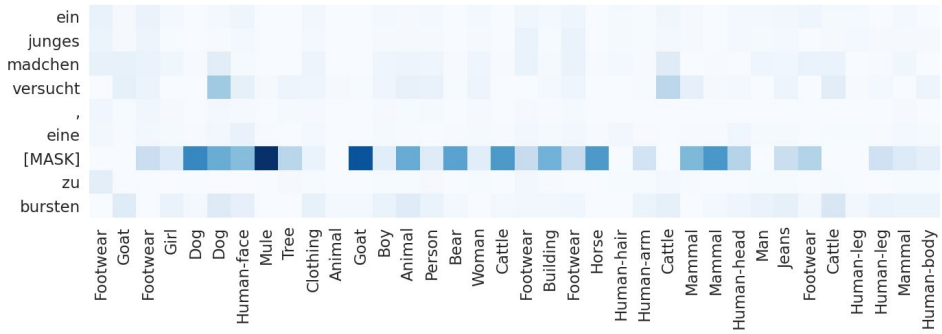
Head #1



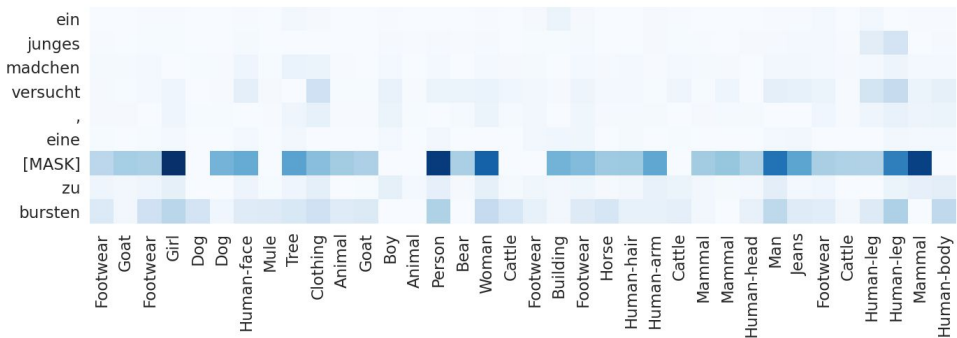
Head #2



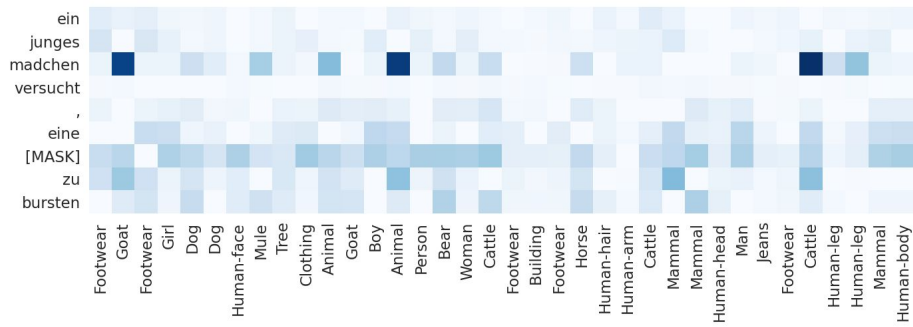
Head #3



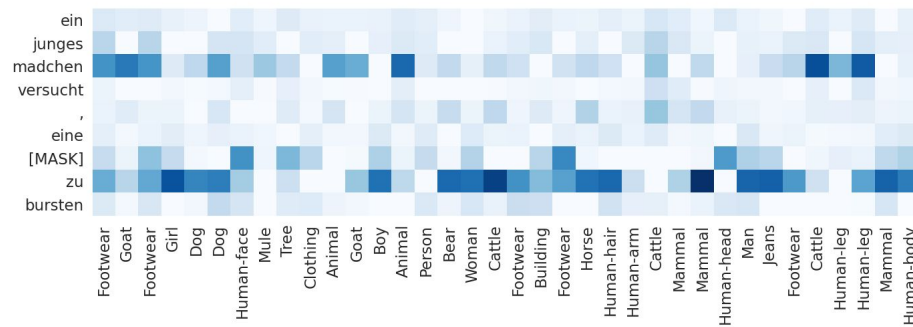
Head #4



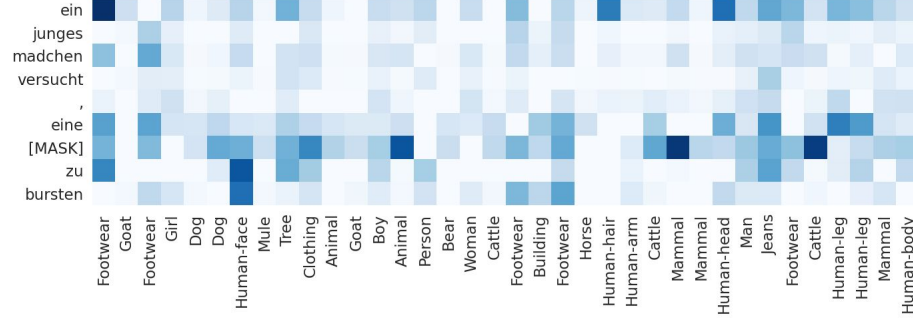
Head #5



Head #6



Head #7



Head #8

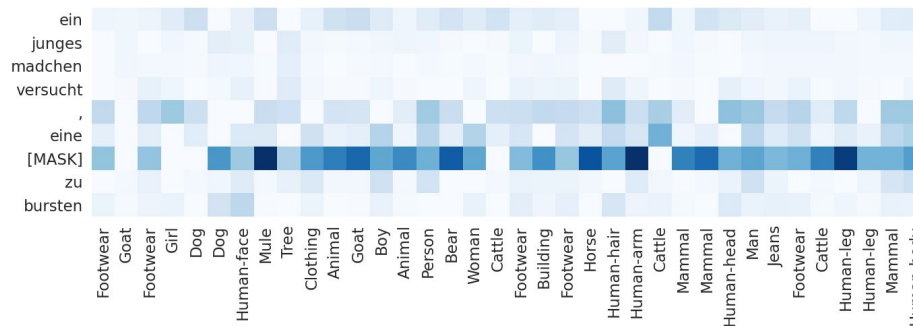


Figure 5.8. We visualize the attention weights for each head in pre-training for a German sentence and object regions. The rows contains German tokens and columns are named as the object label. We replaced the token "ziege" with [MASK] which means "goat" in English.

REFERENCES

- [1] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291, **2019**.
- [2] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565. Association for Computational Linguistics, Melbourne, Australia, **2018**. doi: 10.18653/v1/P18-1238.
- [3] Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *CoRR*, abs/1605.00459, **2016**.
- [4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, **2017**.
- [5] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *CoRR*, abs/1907.10529, **2019**.
- [6] Chiraag Lala and Lucia Specia. Multimodal Lexical Translation. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, H el ene Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan, **2018**. ISBN 979-10-95546-00-9.
- [7] Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. Stochastic answer networks for machine reading comprehension, **2017**.

- [8] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, **2017**. doi:10.18653/v1/p17-1152.
- [9] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*. **2019**.
- [10] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, **1994**.
- [11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, **1997**. doi:10.1162/neco.1997.9.8.1735.
- [12] KyungHyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259, **2014**.
- [13] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, **2015**.
- [14] Robert B. Allen. Several studies on natural language and back-propagation. **1987**.
- [15] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *EMNLP*. **2013**.
- [16] Ozan Çağlayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost Weijer. Lium-cvc submissions for wmt18 multimodal translation task. **2018**.

- [17] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553. Association for Computational Linguistics, Berlin, Germany, **2016**. doi:10.18653/v1/W16-2346.
- [18] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233. Association for Computational Linguistics, Copenhagen, Denmark, **2017**. doi:10.18653/v1/W17-4718.
- [19] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323. Association for Computational Linguistics, Belgium, Brussels, **2018**. doi:10.18653/v1/W18-6402.
- [20] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, **2014**. doi:10.1162/tacl_a_00166.
- [21] Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439. Association for Computational Linguistics, Copenhagen, Denmark, **2017**. doi:10.18653/v1/W17-4746.

- [22] Jindřich Helcl and Jindřich Libovický. CUNI system for the WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 450–457. Association for Computational Linguistics, Copenhagen, Denmark, **2017**. doi:10.18653/v1/W17-4749.
- [23] Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 603–611. Association for Computational Linguistics, Belgium, Brussels, **2018**. doi:10.18653/v1/W18-6439.
- [24] Jindřich Helcl, Jindřich Libovický, and booktitle =. CUNI system for the WMT18 multimodal translation task.
- [25] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. **2009**.
- [26] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398, **2011**.
- [27] Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. Efficient estimation of word representations in vector space. pages 1–12. **2013**.
- [28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, **2013**.
- [29] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on*

- Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, Doha, Qatar, **2014**. doi: 10.3115/v1/D14-1162.
- [30] Oren Melamud, Jacob Goldberger, and Ido Dagan. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61. Association for Computational Linguistics, Berlin, Germany, **2016**. doi: 10.18653/v1/K16-1006.
- [31] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14*, page II–1188–II–1196. JMLR.org, **2014**.
- [32] Wilson L. Taylor. "cloze procedure": a new tool for measuring readability. *Journalism Mass Communication Quarterly*, 30:415–433, **1953**.
- [33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, **2018**.
- [34] Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. *CoRR*, abs/1511.01432, **2015**.
- [35] Prajit Ramachandran, Peter J. Liu, and Quoc V. Le. Unsupervised pretraining for sequence to sequence learning. *CoRR*, abs/1611.02683, **2016**.
- [36] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, **2018**.
- [37] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language*

- Processing*, pages 632–642. Association for Computational Linguistics, Lisbon, Portugal, **2015**. doi:10.18653/v1/D15-1075.
- [38] Alec Radford. Improving language understanding by generative pre-training. **2018**.
- [39] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724, **2015**.
- [40] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kociský, and Phil Blunsom. Reasoning about entailment with neural attention. *CoRR*, abs/1509.06664, **2015**.
- [41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. **2019**.
- [42] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics, Berlin, Germany, **2016**. doi:10.18653/v1/P16-1162.
- [43] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461, **2018**.
- [44] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, **2019**.
- [45] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223, **2019**.

- [46] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE 2.0: A continual pre-training framework for language understanding. *CoRR*, abs/1907.12412, **2019**.
- [47] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, **2019**.
- [48] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, **2019**.
- [49] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860, **2019**.
- [50] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *CoRR*, abs/1905.03197, **2019**.
- [51] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, **2020**.
- [52] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: masked sequence to sequence pre-training for language generation. *CoRR*, abs/1905.02450, **2019**.
- [53] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: evaluating cross-lingual sentence representations. *CoRR*, abs/1809.05053, **2018**.
- [54] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464, **2018**.

- [55] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *CoRR*, abs/1904.01766, **2019**.
- [56] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, **2019**.
- [57] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers, **2019**.
- [58] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, **2016**.
- [59] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language, **2019**.
- [60] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, **2015**.
- [61] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training, **2019**.
- [62] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VI-bert: Pre-training of generic visual-linguistic representations, **2019**.
- [63] R. Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448. **2015**.

- [64] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, **2019**.
- [65] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909, **2015**.
- [66] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71. Association for Computational Linguistics, Brussels, Belgium, **2018**. doi:10.18653/v1/D18-2012.
- [67] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, **2015**.
- [68] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and VQA. *CoRR*, abs/1707.07998, **2017**.
- [69] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Tom Duerig, and Vittorio Ferrari. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982, **2018**.
- [70] Craig Chambers, Ashish Raniwala, Frances Perry, Stephen Adams, Robert R. Henry, Robert Bradshaw, and Nathan Weizenbaum. Flumejava: Easy, efficient data-parallel pipelines. *SIGPLAN Not.*, 45(6):363–375, **2010**. ISSN 0362-1340. doi:10.1145/1809028.1806638.

- [71] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook fair’s WMT19 news translation task submission. *CoRR*, abs/1907.06616, **2019**.
- [72] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233. Association for Computational Linguistics, Copenhagen, Denmark, **2017**. doi:10.18653/v1/W17-4718.
- [73] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323. Association for Computational Linguistics, Belgium, Brussels, **2018**. doi:10.18653/v1/W18-6402.
- [74] Margaret King, Eduard Hovy, John White, Benjamin K T’sou, and Yusoff Zaharin. Mt evaluation. In *Proceedings of the Machine Translation Summit VII*, page 1. **1999**.
- [75] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, **2002**. doi:10.3115/1073083.1073135.
- [76] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Association for Computational Linguistics, Ann Arbor, Michigan, **2005**.

- [77] Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. *CoRR*, abs/1702.01287, **2017**.
- [78] Ozan Caglayan. *Multimodal Machine Translation*. Theses, Université du Maine, **2019**.
- [79] Iacer Calixto, Miguel Rios, and Wilker Aziz. Latent visual cues for neural machine translation. *CoRR*, abs/1811.00357, **2018**.
- [80] Julia Ive, Pranava Madhyastha, and Lucia Specia. Distilling translations with visual awareness. *CoRR*, abs/1906.07701, **2019**.
- [81] Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3035. Association for Computational Linguistics, Online, **2020**. doi:10.18653/v1/2020.acl-main.273.
- [82] Jean-Benoit Delbrouck and Stéphane Dupont. Modulating and attending the source image during encoding improves multimodal translation. *CoRR*, abs/1712.03449, **2017**.
- [83] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *CoRR*, abs/1905.10650, **2019**.