ACCURATE WORD REPRESENTATIONS WITH UNIVERSAL VISUAL GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Word representation is a fundamental component in neural language understanding models. Recently, pre-trained language models (PrLMs) offer a new performant method of contextualized word representations by leveraging the sequence-level context for modeling. Although the PrLMs generally give more accurate contextualized word representations than non-contextualized models do, they are still subject to a sequence of text contexts without diverse hints for word representation from multimodality. This paper thus proposes a visual representation method to explicitly enhance conventional word embedding with multiple-aspect senses from visual guidance. In detail, we build a small-scale word-image dictionary from a multimodal seed dataset where each word corresponds to diverse related images. The texts and paired images are encoded in parallel, followed by an attention layer to integrate the multimodal representations. We show that the method substantially improves the accuracy of disambiguation. Experiments on 12 natural language understanding and machine translation tasks further verify the effectiveness and the generalization capability of the proposed approach.

1 INTRODUCTION

Learning word representations has been an active research field that has gained renewed popularities for decades (Mikolov et al., 2013; Radford et al., 2018). Word representations have been evolved from standard distributed representations (Brown et al., 1992; Mikolov et al., 2013; Pennington et al., 2014) to contextualized language representations from deep pre-trained representation models (PrLMs) (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019). The former static embeddings are commonly derived from distributed representations through capturing the local co-occurrence of words from large-scale unlabeled texts. In contrast, the later contextualized representations are mainly obtained by PrLMs. However, the contexts during language modeling are subject to the individual input text sequence, without diverse hints for word representation from multimodality. Polysemy is a common phenomenon in human language. As shown in Figure 1, word polysemy may result in completely different sentence meanings.[1] The difficulty lies in what kind of meanings the word expresses in each context, and it may be insufficient to obtain the exact representations from the sentence context alone. Therefore, in addition to the contextual word representation, we turn to model multiple-aspect representations for each word to refine more accurate sentence representations.

Recently, the study of multimodal representations has aroused great interest, which shows potential benefits to the language representations through a more comprehensive perception of the real world (Zhang et al., 2018). The major theme of multimodality is to use different modality sources such as texts and images to get a more distinguished representation, which would be beneficial than learning representations from only one modality (Lazaridou et al., 2015). Multimodal information provides more discriminative inputs such as image features than learning representations from only one modality, which is potential to help deal with the ambiguity of words in text-only modality. Most of previous works focused on jointly modeling images and texts (Su et al., 2019; Lu et al., 2019; Tan & Bansal, 2019; Li et al., 2019; Zhou et al., 2019; Sun et al., 2019; Kiela & Bottou, 2014; Silberer & Lapata, 2014; Zablocki et al., 2018; Wu et al., 2019). However, these studies rely on text-image

---

[1] There exists sentence-level syntactic ambiguity in languages as an example, *the girl saw the man on the hill with a telescope*, shown in Appendix A.2, which is not the focus in this work. We left it in future research.

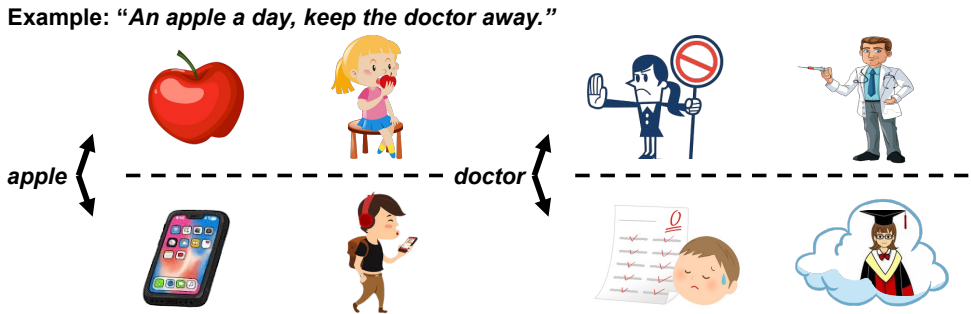**Example: "*An apple a day, keep the doctor away.*"**

Figure 1: Example of ambiguity. This sentence ambiguity which comes from polysemy ambiguities cannot be solved only by text context.

annotations as the paired input. They thus are retrained only in vision-language and multimodal tasks, such as image captioning and visual question answering. For natural language processing (NLP) tasks, most text is not distributed with visual information; therefore, it is essential to break the constraint of the annotation prerequisite when applying visual information to a wide range of mono-modal (e.g., text-only) tasks. To solve the bottlenecks of the high cost of manual image annotation, Zhang et al. (2020) proposed to build a lookup table from a multimodal dataset and then used the search-based method to retrieve images for each input sentence. However, the lookup table is constructed at sentence-level from *Multi30K* (Elliott et al., 2016), which would lead to the relatively limited coverage of the related images for each sentence, and suffers from noise as well. Different from the previous work that incorporates the visual modality as sentence-level guidance, we propose to leverage word-level multimodal assistance inspired by our observation in Figure 1, which offers more accurate means for disambiguation as word-level clues may conveniently help solve sentence-level ambiguity.[2] Meanwhile, multimodality derived from word representations can also be more helpful to provide as many images as possible to alleviate the scarcity of images retrieved for sentences. Furthermore, word-level operation enables the resultant model to benefit broad NLP tasks more than machine translation as previous work focuses on.

In this paper, we propose a visual representation method to explicitly enhance each word with multiple-aspect senses by the visual modality. In detail, based on the universal visual representation method (Zhang et al., 2020), we build a word-image dictionary from the small-scale multimodal dataset Multi30K where one word corresponds to diverse images,[3] which can further be used for text-only tasks. During training and decoding processes, multiple and diverse images corresponding to the input word, which represent the multiple-aspect senses of the word, will be retrieved from the dictionary and then be encoded as image representations by a pre-trained ResNet (He et al., 2016). The texts and paired images are encoded in parallel, followed by an attention layer to fuse those representations. Then, the fused representations are passed to downstream task-specific layers for prediction. Experiments on various natural language understanding and machine translation tasks verified the effectiveness and the generalization capability of the proposed approach.

2 BACKGROUND

2.1 LANGUAGE REPRESENTATION

Training machines to comprehend human language requires comprehensive and accurate modeling of natural language texts, in which the fundamental component is language representation. Empirically, distributed representations (Brown et al., 1992; Mikolov et al., 2013; Pennington et al., 2014) have been widely applied in diverse NLP tasks, which benefit from the ability to capture the local co-occurrence of words from the large-scale unlabeled text. Recently dominant pre-trained

---

[2]Instead of a linguistic term, "disambiguation" in this paper means a broad kind of practice for more distinguishable representation from diverse input sources.

[3]We describe our method by regarding the processing unit as word though this method can also be applied to a subword-based sentence for which the subword is considered to be the processing unit.

contextualized language models such as ELMo (Peters et al., 2018), GPT (Radford et al., 2018), and BERT (Devlin et al., 2019) focus on learning context-dependent representations by taking into account the context for each occurrence of a given word (Peters et al., 2018) The major technical improvement over the traditional embeddings of these newly proposed language models is that they focus on extracting context-sensitive features, whose contextualized embedding for each word will be different according to the sentence. When integrating these contextual word embeddings with existing task-specific architectures, ELMo helps boost several major NLP benchmarks (Peters et al., 2018), including question answering on SQuAD, sentiment analysis (Socher et al., 2013), and named entity recognition (Sang & De Meulder, 2003). The latest evaluation shows that BERT especially shows effectiveness in language understanding tasks on GLUE, MultiNLI, and SQuAD (Devlin et al., 2019). Nevertheless, the word embeddings learned through these approaches are completely based on the textual corpus. In practice, the meaning of the word itself would involve multiple senses and naturally leads to ambiguity. In this work, we focus on more accurate word representation with effective disambiguation for pre-trained contextualized language models.

2.2 MULTIMODAL PERCEPTION

As a special kind of *language* shared by people worldwide, visual modality may help machines have a more comprehensive perception of the real world. Recently, there has been a great deal of interest in integrating image presentations into pre-trained Transformer architectures (Su et al., 2019; Lu et al., 2019; Tan & Bansal, 2019; Li et al., 2019; Zhou et al., 2019; Sun et al., 2019). The common strategy is to take a Transformer model (Vaswani et al., 2017), such as BERT, as the backbone and learn joint representations of vision and language in a pre-training manner inspired by the mask mechanism in pre-trained language models (Devlin et al., 2019). All these works require the annotation of task-dependent sentence-image pairs, which are limited to vision-language tasks, such as image captioning and visual question answering. Notably, Glyce (Wu et al., 2019) proposed incorporating glyph vectors for Chinese character representations. However, it can only be used for Chinese and only involves single image enhancement. Zhang et al. (2020) proposed using multiple images for neural machine translation (NMT), based on a text-image lookup table trained over a sentence-image pair corpus. However, the number of images is fixed because of the lack of similarity measurement in the simple lookup method, so that it would possibly suffers from the noise of irrelevant images. Compared with Zhang et al. (2020), this paper mainly differs by both sides of motivation and technique.

1) The lookup table constructed at sentence-level as that in Zhang et al. (2020) would lead to the relatively limited coverage of the related images for each sentence, and suffers from noise as well. Different from the previous work that incorporates the visual modality as sentence-level guidance, we propose to leverage word-level multimodal assistance, which offers more accurate means for disambiguation as word-level clues may conveniently help solve sentence-level ambiguity. Meanwhile, multimodality derived from word representations can also be more helpful to provide as many images as possible to alleviate the scarcity of images retrieved for sentences.

2) Technically, we extend the word embeddings with discriminative image features, enabling the resultant model to benefit broad NLP tasks more than machine translation as previous work focuses on. Intuitively, our method is more fine-grained which can result in adequate word-image overlap. According to the analysis in Section 5, we see that in most of our datasets, over 50% tokens have paired images, effectively addressing the drawback of sentence-level method in Zhang et al. (2020) as it is hard to guarantee adequate images for each sentence. Further, the results in Table 3 show that our method is indeed better than the previous methods, especially in the small-scale datasets, indicating that our method would be useful for low-resource scenarios.

As a result, the major contributions of this work are as follows:

1) Instead of retrieving images for each sentence sparsely, this work focuses on word-level modeling, by enriching word representations with related images as fine-grained visual hints.

2) The proposed approach does not rely on large-scale aligned sentence-image corpus and can achieve modest results with only a few seed images.

3) Our method is light-weight and is not limited to specific tasks; it is generally motivated to apply visual guidance to a wide range of NLP tasks and can easily be applied to other NLP models.

3 MODEL

In this section, we will first introduce the word-image dictionary for word-image(s) retrieval, and then elaborate the proposed visual representation method to enhance word representations with visual guidance.

3.1 WORD-IMAGE DICTIONARY

During the preprocessing, we set up a dictionary for subsequent word-to-images queries inspired by Zhang et al. (2020). Specifically, as described in Algorithm 1, the default input setting of the multimodal dataset (i.e., *Multi30K*) is a sentence-image pair denoted as $(X_i, e_i)_{i \in \{1,2,...,k\}} \in (S, E)$, where $S = \{X_1, X_2, \dots X_k\}$ is the set of input sentences and $E = \{e_1, e_2, \dots, e_k\}$ is the set of paired images. Both sets have the same length k. The input sentence X_i is first filtered by a stop word list[4] and is further segmented to subwords by a specified tokenizer that depends on downstream tasks, which ensures the token overlap with task datasets to the maximum extent. We denote the formed token sequence of X_i as $T_i = \{t_1^i, t_2^i, \dots, t_n^i\}$. Then, for each subword token $t_j^i, j \in \{1, 2, \dots, n\}$, we map the paired image e_i to t_j^i in the dictionary $\mathcal{D}$. At the end of processing the whole multimodal corpus, we form a word-image dictionary where each subword contains diverse images sorted by the number of image occurrence, which are later converted to the image representation as visual hints.[5]

Algorithm 1 Word-image Dictionary Conversion Algorithm

Require: Input sentence set $S = \{X_1, X_2, \dots X_k\}$ and paired image set $E = \{e_1, e_2, \dots, e_k\}$
Ensure: Word-image dictionary $\mathcal{D}$ where each word is associated with a group of images
 1: **procedure** $\mathcal{D}$
 2:     **for** each sentence-image pair $(X_i, e_i)_{i \in \{1,2,...,k\}} \in (S, E)$ **do**
 3:         Filter stop words in the sentence X_i
 4:         Segment the sentence by a specified tokenizer that depends on downstream tasks
 5:         The formed subword set is denoted as $T_i = \{t_1^i, t_2^i, \dots, t_n^i\}$
 6:         **for** each word $(t_j^i)_{j \in \{1,2,...,n\}} \in T_i$ **do**
 7:             Add e_i to the corresponding meaning terms $\mathcal{D}[t_j^i]$ for word t_j^i
 8: **return** Word-image dictionary $\mathcal{D}$

3.2 INTEGRATION OF THE WORD AND IMAGE REPRESENTATIONS

Figure 2 overviews the framework of our visual representation method. First, we employ the same preprocessing method as making the word-image dictionary, that is to filter stop words and tokenize the input sentence X into a sequence of subwords $T = \{t_1, t_2, \dots, t_n\}$. The consistency of preprocessing method guarantees the maximum word overlap between the task datasets and the word-image dictionary. After that, the subword sequence T is fed into a specified text encoder that depends on downstream tasks to learn the source text representation:

$$T = \text{Tokenizer}(\text{Filter}(X)), \boldsymbol{H} = \text{Encoder}(T), \tag{1}$$

where $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \dots, \boldsymbol{h}_n\}$ is the text representation of length n. Synchronously, to obtain the corresponding image representation, we look up the word-image dictionary to retrieve top m images for each subword in T as the input to the image encoder. The image encoder is composed of a pre-trained ResNet and a pooling layer:

$$I = \mathcal{D}[T], M = \text{Pooling}(\text{ResNet}(I)), \tag{2}$$

where $I = \{I_1, I_2, \dots, I_n\}$ in which $I_i = \{img_1^i, img_2^i, \dots, img_m^i\}$ are the sequence of retrieved images for the i-th token, $\mathcal{D}$ is the word-image dictionary and $M = \{M_1, M_2, \dots, M_n\} \in \mathbb{R}^{n \times m \times 2048}$ is the source image representation of dimension 2048 as fined-grained visual guidance. Then, in aim of extracting the region of interest from image representation, on the top of a linear

---
[4]https://github.com/stopwords-iso/stopwords-en.

[5]Examples of paired images for tokens are shown in the Appendix A.1.
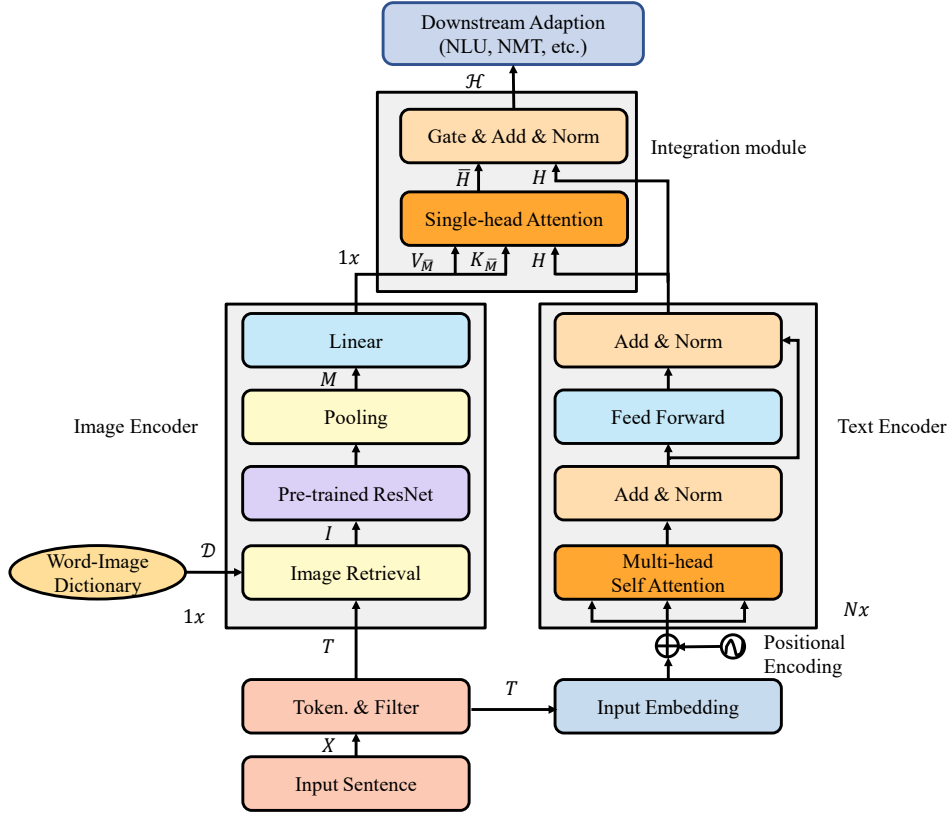
Figure 2: Overview of the framework of our proposed method.

transformation which reduces the dimension of image representation to that of text representation, we apply an attention mechanism following Zhang et al. (2020) to model the interactions of the text and image representations, which takes text representation as query and image representation as key and value:

$$\overline{M} = \text{Linear}(M), \overline{h}_i = \text{ATT}_{\overline{M}_i}(h_i, K_{\overline{M}_i}, V_{\overline{M}_i}), \tag{3}$$

where $(h_i, \overline{M}_i)_{i \in \{1,2,...,n\}} \in zip(H, \overline{M})$ are respectively the source text, image representation of subword t_i, and $\{K_{\overline{M}_i}, V_{\overline{M}_i}\}$ are packed from the image representation $\overline{M}_i$. Then, we have $\overline{H} = \{\overline{h}_1, \overline{h}_2, \ldots, \overline{h}_n\}$ as the text-conditioned image representation. Finally, we apply a gated aggregation method to fuse text representation and image representation in concern of learning the weights of the two independently:

$$\lambda = \sigma(\text{Linear}(\text{Concat}(h_i, \overline{h}_i))), \mathcal{H}_i = (1 - \lambda)h_i + \lambda\overline{h}_i, \tag{4}$$

where Concat, Linear and σ are respectively a concatenation layer, a linear transformer and a logistic sigmoid function. $\lambda \in [0, 1]$ is to weight the expected importance of image representation for the source token. The joint representation $\mathcal{H} = \{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_n\}$ will be fed into downstream tasks for prediction.

3.3 APPLICATION IN SPECIFIC NLP TASKS

As a general approach, the visual guidance can be easily applied to standard NLP models. Here, we introduce the specific implementation parts of our proposed method for downstream tasks by taking natural language understanding (NLU) and NMT tasks as examples. For NLU, the baseline model is BERT (Devlin et al., 2019), and we apply the BERT Tokenizer for subword segmentation. The pooled representation of $\mathcal{H}$ will be fed to a feed-forward layer to make the prediction, which follows the same downstream procedure as BERT. For NMT, the text encoder is designated as a

self-attention based encoder with multiple layers (Vaswani et al., 2017), and the byte pair encoding algorithm is adopted for tokenization. The fused representation $\mathcal{H}$ will be directly fed to the decoder to predict the target translation.

4 EXPERIMENTS

In this section, we first introduce our evaluation tasks and model implementations. For the experiments, we start by presenting experiments on a disambiguation task, and then further conduct a wide range of evaluations on the major natural language understanding and translation tasks, involving 12 NLP benchmark datasets for natural language inference (NLI), semantic similarity, text classification, and machine translation. Part of the NLU tasks are available from the recently released GLUE benchmark (Wang et al., 2018), which is a collection of nine NLU tasks.

4.1 TASKS

4.1.1 NATURAL LANGUAGE UNDERSTANDING

The NLU task involves natural language inference, semantic similarity, and classification subtasks.

Natural Language Inference involves reading a pair of sentences and assessing the relationship between their meanings, such as entailment, neutral, or contradiction. We evaluated the proposed method on four diverse datasets: SNLI (Bowman et al., 2015), MNLI (Nangia et al., 2017), QNLI (Rajpurkar et al., 2016), and RTE (Bentivogli et al., 2009).

Semantic Similarity aims to predict whether two sentences are semantically equivalent. Three datasets were used: Microsoft Paraphrase Corpus (MRPC) (Dolan & Brockett, 2005), Quora Question Pairs (QQP) dataset (Chen et al., 2018), and Semantic Textual Similarity benchmark (STS-B) (Cer et al., 2017).

Classification CoLA (Warstadt et al., 2018) is used to predict whether an English sentence is linguistically acceptable. SST-2 (Socher et al., 2013) provides a dataset for sentiment classification that needs to determine whether the sentiment of a sentence is positive or negative.

4.1.2 NEURAL MACHINE TRANSLATION

The proposed method was evaluated on three widely-used translation tasks, including Multi30K for WMT'16 and WMT'17 multimodal tasks, WMT'14 English-to-German (EN-DE) and WMT'16 English-to-Romanian (EN-RO) for text-only NMT, which are standard corpora for machine translation evaluation.

Multi30K contains 30K English→{German, French} parallel sentence pairs with visual annotations, which is an extension of Flickr30k (Brown et al., 2003). For WMT'16 and WMT'17 tasks, we have two Test sets, test2016 and test2017, with 1,000 pairs for each. For test2016, we report the results on English-Czech (EN-CS). We also report results on the MSCOCO testset of test2017 that has 461 more challenging out-of-domain instances that contain ambiguous verbs (Elliott et al., 2017).

WMT'14 EN-DE 4.43M bilingual sentence pairs of the WMT14 dataset were used as training data, including Common Crawl, News Commentary, and Europarl v7. The *newstest2013* and *newstest2014* datasets were used as the Dev set and Test set, respectively.

WMT'16 EN-RO we experimented with the officially provided parallel corpus: Europarl v7 and SETIMES2 from WMT'16 with 0.6M sentence pairs. We used *newsdev2016* as the Dev set and *newstest2016* as the Test set.

4.2 MODEL IMPLEMENTATION

Since our task involves text understanding and translation, there are two types of model architectures served as our baselines, the NLU model for language understanding and the NMT model for translation. According to our preliminary experiments, we set the default maximum number of images m for each word as 5. More detailed analysis of m is presented in Section 5. Besides standard NLU and

NMT baseline models as described below, we also compare to two baselines: 1) + *Random*: random baseline where each word is randomly paired with images; 2) + *Params*: the baseline that has the same number of parameters with our proposed method (+ *VG*) by replacing the image features with the default standard normal distribution as commonly used in the initialization of Embedding layer.[6]

4.2.1 NLU MODEL

For the NLU tasks, the baseline was BERT (Devlin et al., 2019).[7] We used the public pre-trained weights of BERT and followed the same fine-tuning procedure as BERT. We use BERT base as default.[8] The initial learning rate was set in the range {2e-5, 3e-5} with a warm-up rate of 0.1 and L2 weight decay of 0.01. The batch size was selected from {16, 24, 32}. The maximum number of epochs was set in the range [2, 5]. Text was tokenized using SentencePiece (Sennrich et al., 2016), with a maximum length of 128.[9] We report the average Dev set accuracy from 5 random runs.

4.2.2 NMT MODEL

Our NMT baseline was a text-only Transformer (Vaswani et al., 2017). We used six layers for both encoder and decoder. The number of dimensions of all input and output layers was set to 512 (the *base* setting in Vaswani et al. (2017)). The inner feed-forward neural network layer was set to 2048. The heads of all multi-head modules were set to eight in both encoder and decoder layers. The learning rate was varied under a warm-up strategy with 8,000 steps. In each training batch, a set of sentence pairs contained approximately 4096×4 source tokens and 4096×4 target tokens. For the Multi30K dataset, we trained the model up to 10,000 steps, and the training was early-stopped if *dev* set BLEU score did not improve for ten epochs. The dropout rate was 0.15. For the EN-DE and EN-RO tasks, the training steps are 200,000. The signtest (Collins et al., 2005) is a standard statistical-significance test. All experiments were conducted with *fairseq* (Ott et al., 2019).[10]

4.3 DISAMBIGUATION EXPERIMENT

A natural intuition of using visual clues for text representation is the advantage of alleviating the ambiguation of language. To evaluate the model performance for disambiguation, we use a dataset from the HVG (Parida et al., 2019), which served as a part of WAT'19 Multimodal Translation Task.[11] The dataset consists of a total of 31,525 randomly selected images from Visual Genome (Krishna et al., 2017) and a parallel image caption corpus in English-Hindi for selected image segments. The training part consists of $29K$ English and Hindi short captions of rectangular areas in photos of various scenes, and it is complemented by three test sets: development (Dev), evaluation (Test), and challenge test set (Challenge). The challenge test set was created by searching for (particularly) ambiguous English words based on the embedding similarity and manually selecting those where the image helps to resolve the ambiguity. We did not make any use of the images and use the same settings as the experiments on Multi30K.

Model	Dev	Test	Challenge
Baseline	47.32	39.61	20.66
Ours	48.56 (**+1.24**)	41.54 (**+1.93**)	22.86 (**+2.20**)

Table 1: Accuracy (here we use BLEU4 scores) of the multimodal disambiguation experiments on WAT'19 English to Hindi dataset.

We employ the above Transformer model as our baseline and strengthen it with visual guidance (described in Section 3). As the results shown in Table 1, we observe that our model works effectively

---

[6]https://pytorch.org/docs/stable/_modules/torch/nn/modules/sparse.html#Embedding

[7]https://github.com/huggingface/transformers.

[8]Detailed comparisons of different pre-trained weights are shown in the Appendix A.3.

[9]https://github.com/google/sentencepiece.

[10]https://github.com/pytorch/fairseq.

[11]http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019/index.html.

Method	Classification		Natural Language Inference				Semantic Similarity		
	CoLA (mc)	SST-2 (acc)	MNLI (acc)	QNLI (acc)	RTE (acc)	SNLI (acc)	MRPC (F1)	QQP (F1)	STS-B (pc)
BERT	57.3	92.6	84.6	90.8	66.4	90.6	88.9	87.2	89.4
BERT + Random	58.5	92.4	83.9	89.9	62.3	90.5	89.8	87.7	86.7
BERT + Params	57.6	92.3	83.8	89.8	61.7	90.4	89.3	87.6	86.8
BERT + VG	**59.4**++	**93.0**+	**84.7**	**91.1**	**69.0**++	90.7	**89.8**++	**88.2**++	**89.6**

Table 2: Results on GLUE benchmark. *mc* and *pc* denote the Matthews correlation and Pearson correlation, respectively. "++/+" indicate that the proposed method was significantly better than the corresponding baseline at significance level $p<0.01/0.05$.

Model	Multi30K 2016			Multi30K 2017 (flickr)		Multi30K 2017 (mscoco)		WMT'14	WMT'16
	EN-DE	EN-FR	EN-CS	EN-DE	EN-FR	EN-DE	EN-FR	EN-DE	EN-RO
Existing NMT systems									
Trans.	N/A	N/A	N/A	N/A	N/A	N/A	N/A	27.30	N/A
MMT	35.09	57.40	N/A	27.10	N/A	48.02	N/A	N/A	N/A
UVR	35.72	58.32	N/A	26.87	48.69	N/A	N/A	**28.14**	33.78
Our systems									
Trans.	36.02	57.88	30.08	28.26	49.66	27.16	41.73	27.31	32.66
Trans. + Random	36.16	58.87	29.79	28.36	50.62	27.12	41.64	27.22	32.46
Trans. + Rarams	36.06	57.62	28.56	28.06	50.24	26.98	42.14	27.16	32.51
Trans. + VG	**36.69**+	**59.47**++	**30.77**+	**30.74**++	**51.75**++	**27.53**+	**42.99**++	27.80+	**33.81**++

Table 3: BLEU scores on MMT and NMT tasks. Trans. is short for the transformer (Vaswani et al., 2017). The MMT and UVR are from Zhang et al. (2020). "++/+" after the BLEU score indicate that the proposed method was significantly better than the corresponding baseline Transformer at significance level $p<0.01/0.05$.

on the challenge disambiguation set, indicating that the visual information induced by retrieved images allows disambiguation of translation, which inspires us to apply visual modality as word-level auxiliary information for general language representation.

4.4 MAIN RESULTS

Tables 2-3 show the results for the 12 NLU and NMT tasks, respectively. According to the results, we make the following observations:

1) Table 2 shows that our method (+ VG) outperforms the baselines consistently, indicating that it is generally helpful for a wide range of NLU tasks.[12] The results verified the effectiveness of modeling visual information for language understanding.

2) Results in Table 3 show that our model also outperformed the Transformer baseline in both multimodal (Multi30K) and text-only machine translation (WMT) tasks. As seen, the proposed method significantly outperformed the baseline, demonstrating the effectiveness of modeling visual information for NMT. In particular, the effectiveness was adapted to the translation tasks of the different language pairs, which have different scales of training data, verifying that the proposed approach is a universal method for improving translation performance.

3) Our method introduced a few parameters over our baselines. Taking the NMT model as an example, the extra trainable parameter number is only +4.2M, which is around 6.3% of the baseline parameters as we used the fixed image embeddings from the pre-trained ResNet feature extractor. Besides, the training time was basically the same as the baseline model. Since both of the + *Random* and + *Param* have the same number of parameters with our proposed method + *VG*, our method still outperform those augmented baselines. The comparison indicates our method does not simply bene-

---

[12]Since the test set of GLUE is not publicly available, we conducted the comparison with our baselines on the Dev set. More detailed results including the Test results are presented in the Appendix A.3.
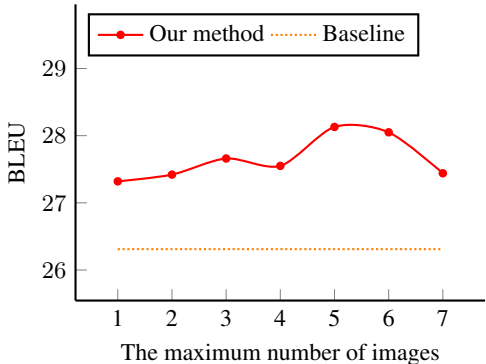
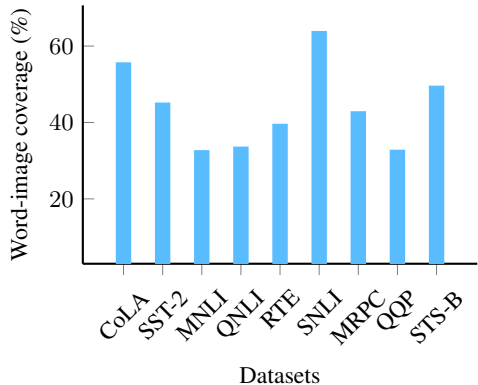Figure 4: Influence of the maximum number of images paired for each word on the Multi30K EN-DE Test2017.

Figure 5: Token overlap (ratios) of the task datasets and the seed Multi30K dataset.

fit from more parameters. Meanwhile, we notice that + *Random* showed slightly better performance than the baseline on the Multi30K dataset though it might involve some noise, which is reasonable due to the possible regularization effect and would alleviate over-fitting on the small-scale dataset (Brown et al., 2003; Noh et al., 2017; Brownlee, 2019).

5 ANALYSIS

The influence of numbers of paired images To investigate the influence of the maximum number of images m paired for each word, we constrained m in $\{1, 2, 3, 4, 5, 6, 7\}$ for experiments on the EN-DE Test2017 set, as shown in Figure 4. All models outperformed the baseline Transformer (base), indicating the effectiveness of visual guidance. As the number of images increases, the BLEU score generally showed an upward trend at the beginning from 27.32 to 28.13 and dropped sightly after reaching the peak when $m = 5$.

Image coverage analysis One possible drawback of the sentence-level method in Zhang et al. (2020) would suffer from relatively limited coverage of the related images because it is hard to guarantee adequate images for each sentence, which could be alleviated by our fine-grained word-level method. To investigate how many tokens in our tasks can be paired with related images, we calculate the token overlap ratio between the tokens from the multimodal seed dataset and the task datasets. Figure 5 depicts the statistics. We see that in most of our datasets, over 50% tokens have paired images, which indicates that our tasks could enjoy adequate overlap through our method.

The influence of using different images corpora We are interested in whether extra large-scale seed image data could render better model performance. Therefore, we evaluate the performance by further using the larger-scale MS COCO image caption dataset (Lin et al., 2014). The BLEU score of Multi30K EN-DE Test2017 is boosted from 28.13 to 28.36. We contemplate that additional data may further improve the performance, even that image-only data can also be annotated by image caption models and then employed to enhance the model capability, which is left for future work.

6 CONCLUSION

In this paper, we present a visual representation method to explicitly enhance conventional word embedding with multiple-aspect senses by the visual modality. Empirical studies on a range of NLP tasks verified the effectiveness, especially the advance for disambiguation. The implementation of our method in the existing deep learning NLP systems demonstrates its flexibility and versatility. In future work, we consider investigating non-parallel text and image data to improve the language representation ability of deep learning models.

REFERENCES

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The fifth pascal recognizing textual entailment challenge. In *ACL-PASCAL*, 2009.

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*, pp. 632–642, 2015. doi: 10.18653/v1/D15-1075.

Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480, 1992.

Warick M Brown, Tamás D Gedeon, and David I Groves. Use of noise to augment training data: a neural network method of mineral–potential mapping in regions of limited known deposit examples. *Natural Resources Research*, 12(2):141–152, 2003.

J Brownlee. Train neural networks with noise to reduce overfitting. *Machine Learning Mastery*, 2019.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. Quora question pairs, 2018.

Michael Collins, Philipp Koehn, and Ivona Kucerova. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, 2005.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, June 2019. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *IWP2005*, 2005.

Desmond Elliott, Stella Frank, Khalil Sima ' an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pp. 70–74, 2016.

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*, pp. 215–233, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Douwe Kiela and Léon Bottou. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 36–45, 2014. doi: 10.3115/v1/D14-1005.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Angeliki Lazaridou, Marco Baroni, et al. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 153–163, 2015.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4487–4496, 2019. doi: 10.18653/v1/P19-1441.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pp. 13–23, 2019.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel R Bowman. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. In *RepEval*, 2017.

Hyeonwoo Noh, Tackgeun You, Jonghwan Mun, and Bohyung Han. Regularizing deep neural networks by noise: Its interpretation and optimization. In *Advances in Neural Information Processing Systems*, pp. 5109–5118, 2017.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53, 2019.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. Hindi visual genome: A dataset for multi-modal english to hindi machine translation. *Computación y Sistemas*, 23(4), 2019.

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)*, 2014. doi: 10.3115/v1/D14-1162.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, 2018. doi: 10.18653/v1/N18-1202.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *Technical report*, 2018.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2016.

Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

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)*, pp. 1715–1725, 2016.

Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 721–732, 2014. doi: 10.3115/v1/P14-1068.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *arXiv preprint arXiv:1904.01766*, 2019.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5103–5114, 2019. doi: 10.18653/v1/D19-1514.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *2018 EMNLP Workshop BlackboxNLP*, 2018.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*, 2018.

Wei Wu, Yuxian Meng, Qinghong Han, Muyu Li, Xiaoya Li, Jie Mei, Ping Nie, Xiaofei Sun, and Jiwei Li. Glyce: Glyph-vectors for chinese character representations. *arXiv preprint arXiv:1901.10125*, 2019.

Eloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. Learning multi-modal word representation grounded in visual context. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 5626–5633, 2018.

Kun Zhang, Guangyi Lv, Le Wu, Enhong Chen, Qi Liu, Han Wu, and Fangzhao Wu. Image-enhanced multi-level sentence representation net for natural language inference. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 747–756. IEEE, 2018.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=Byl8hhNYPS`.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019.

A APPENDIX

A.1 RETRIEVED IMAGE EXAMPLES

Figure 5 shows examples to interpret the image retrieval process intuitively, where the words in bold face contain various meanings, for example:
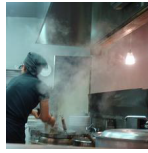**glass**: a drinking container or an optical instrument; **cook**: an action or a profession;
bank: financial establishment or the land alongside or sloping down to a river or lake;
body: physical structure of a person or a body of water;
guitar: different shapes and colors;
shirt: different shapes and colors.

a chef with **glasses cooks** french bread in a giant oven .
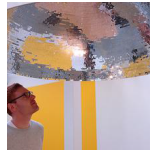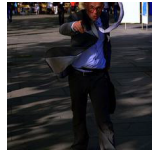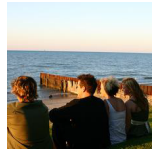
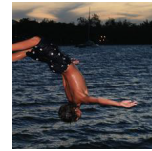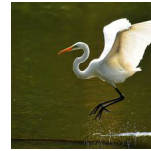**glass**   **cook**

a small group of people are sitting on the **bank** of a large **body** of water in the grass .

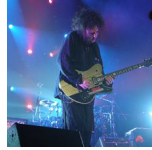**bank:**   **body:**

a man in **green** holds a **guitar** while the other man observes his **shirt** .

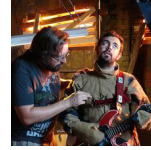**guitar**   **shirt**

Figure 5: Retrieved image examples.

A.2 GRAMMATICAL AMBIGUITY

Grammatically ambiguous sentences are not covered by our model, such as *the girl saw the man on the hill with a telescope* shown in Figure 6 and 7.

Figure 6: **the girl** saw the man **on the hill with a telescope**

Figure 7: the girl saw **the man on the hill with a telescope**

A.3 GLUE RESULTS

Table 4 shows the complete Dev and Test results for the GLUE benchmark.

Method	Classification		Natural Language Inference				Semantic Similarity		
	CoLA	SST-2	MNLI	QNLI	RTE	SNLI	MRPC	QQP	STS-B
	(mc)	(acc)	(acc)	(acc)	(acc)	(acc)	(F1)	(F1)	(pc)
Dev set results for Comparison									
BERT_{LARGE}	60.6	93.2	86.6	92.3	70.4	91.0	88.0	88.0	90.0
MT-DNN	63.5	**94.3**	87.1	92.9	83.4	92.2	87.5	**89.2**	90.7
BERT_{BASE}	57.3	92.6	84.6	90.8	66.4	90.6	88.9	87.2	89.4
BERT_{BASE} + VG	59.4	93.0	84.7	91.1	69.0	90.7	89.8	88.2	89.6
BERT_{LARGE}	60.8	93.3	86.3	92.4	71.1	91.3	89.5	88.0	89.5
BERT_{LARGE} + VG	62.8	93.6	86.6	92.5	73.6	91.5	90.5	88.5	90.1
BERT_{WWM}	63.6	93.6	87.2	93.6	77.3	92.1	90.8	88.8	90.5
BERT_{WWM} + VG	**64.9**	93.9	**87.4**	**93.9**	**78.5**	**92.2**	**90.9**	88.9	**91.4**
Test set results for single model with standard single-task training									
GPT	45.4	91.3	82.1	88.1	56.0	89.9	82.3	70.3	82.0
GPT on STILTs	47.2	93.1	80.8	87.2	69.1	-	87.7	70.1	85.3
BERT	60.5	94.9	86.7	92.7	70.1	-	89.3	72.1	87.6
MT-DNN	61.5	**95.6**	86.7	-	75.5	91.6	**90.0**	72.4	88.3
BERT_{BASE} + VG	50.7	93.1	84.3	90.5	66.7	91.0	84.9	71.0	85.8
BERT_{LARGE} + VG	57.4	94.5	85.4	92.9	70.1	91.1	88.1	71.4	87.3
BERT_{WWM} + VG	**61.6**	94.9	**87.1**	**94.0**	**78.6**	**91.7**	**90.6**	**72.6**	**88.8**

Table 4: Results on GLUE benchmark. The public results are from GPT (Radford et al., 2018), BERT (Devlin et al., 2019), MT-DNN (Liu et al., 2019). *mc* and *pc* denote the Matthews correlation and Pearson correlation, respectively.