A Visual Attention Grounding Neural Model for Multimodal Machine Translation

Mingyang Zhou **Runxiang Cheng** **Yong Jae Lee** **Zhou Yu**
Department of Computer Science
University of California, Davis
{minzhou, samcheng, yongjaelee, joyu}@ucdavis.edu

Abstract

We introduce a novel multimodal machine translation model that utilizes parallel visual and textual information. Our model jointly optimizes the learning of a shared visual-language embedding and a translator. The model leverages a visual attention grounding mechanism that links the visual semantics with the corresponding textual semantics. Our approach achieves competitive state-of-the-art results on the Multi30K and the Ambiguous COCO datasets. We also collected a new multilingual multimodal product description dataset to simulate a real-world international online shopping scenario. On this dataset, our visual attention grounding model outperforms other methods by a large margin.

1 Introduction

Multimodal machine translation is the problem of translating sentences paired with images into a different target language (Elliott et al., 2016). In this setting, translation is expected to be more accurate compared to purely text-based translation, as the visual context could help resolve ambiguous multi-sense words. Examples of real-world applications of multimodal (vision plus text) translation include translating multimedia news, web product information, and movie subtitles.

Several previous endeavours (Huang et al., 2016; Calixto et al., 2017a; Elliott and Kádár, 2017) have demonstrated improved translation quality when utilizing images. However, how to effectively integrate the visual information still remains a challenging problem. For instance, in the WMT 2017 multimodal machine translation challenge (Elliott et al., 2017), methods that incorporated visual information did not outperform pure text-based approaches with a big margin.

In this paper, we propose a new model called Visual Attention Grounding Neural Machine Translation (VAG-NMT) to leverage visual information more effectively. We train VAG-NMT with a multitask learning mechanism that simultaneously optimizes two objectives: (1) learning a translation model, and (2) constructing a vision-language joint semantic embedding. In this model, we develop a visual attention mechanism to learn an attention vector that values the words that have closer semantic relatedness with the visual context. The attention vector is then projected to the shared embedding space to initialize the translation decoder such that the source sentence words that are more related to the visual semantics have more influence during the decoding stage. When evaluated on the benchmark Multi30K and the Ambiguous COCO datasets, our VAG-NMT model demonstrates competitive performance compared to existing state-of-the-art multimodal machine translation systems.

Another important challenge for multimodal machine translation is the lack of a large-scale, realistic dataset. To our knowledge, the only existing benchmark is Multi30K (Elliott et al., 2016), which is based on an image captioning dataset, Flickr30K (Young et al., 2014) with manual German and French translations. There are roughly 30K parallel sentences, which is relatively small compared to text-only machine translation datasets that have millions of sentences such as WMT14 EN→DE. Therefore, we propose a new multimodal machine translation dataset called IKEA to simulate the real-world problem of translating product descriptions from one language to another. Our IKEA dataset is a collection of parallel English, French, and German product descriptions and images from IKEA's and UNIQLO's websites. We have included a total of 3,600 products so far and will include more in the future.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3643–3653
Brussels, Belgium, October 31 - November 4, 2018.

2 Related Work

In the machine translation literature, there are two major streams for integrating visual information: approaches that (1) employ separate attention for different (text and vision) modalities, and (2) fuse visual information into the NMT model as part of the input. The first line of work learns independent context vectors from a sequence of text encoder hidden states and a set of location-preserving visual features extracted from a pre-trained convnet, and both sets of attentions affect the decoder's translation (Calixto et al., 2017a; Helcl and Libovický, 2017). The second line of work instead extracts a global semantic feature and initializes either the NMT encoder or decoder to fuse the visual context (Calixto et al., 2017b; Ma et al., 2017). While both approaches demonstrate significant improvement over their Text-Only NMT baselines, they still perform worse than the best monomodal machine translation system from the WMT 2017 shared task (Zhang et al., 2017).

The model that performs best in the multimodal machine translation task employed image context in a different way. (Huang et al., 2016) combine region features extracted from a region-proposal network (Ren et al., 2015) with the word sequence feature as the input to the encoder, which leads to significant improvement over their NMT baseline. The best multimodal machine translation system in WMT 2017 (Caglayan et al., 2017) performs element-wise multiplication of the target language embedding with an affine transformation of the convnet image feature vector as the mixed input to the decoder. While this method outperforms all other methods in WMT 2017 shared task workshop, the advantage over the monomodal translation system is still minor.

The proposed visual context grounding process in our model is closely related to the literature on multimodal shared space learning. (Kiros et al., 2014) propose a neural language model to learn a visual-semantic embedding space by optimizing a ranking objective, where the distributed representation helps generate image captions. (Karpathy and Li, 2014) densely align different objects in the image with their corresponding text captions in the shared space, which further improves the quality of the generated caption. In later work, multimodal shared space learning was extended to multimodal multilingual shared space learning. (Calixto et al., 2017c) learn a multi-modal multilingual shared space through optimization of a modified pairwise contrastive function, where the extra multilingual signal in the shared space leads to improvements in image-sentence ranking and semantic textual similarity task. (Gella et al., 2017) extend the work from (Calixto et al., 2017c) by using the image as the pivot point to learn the multilingual multimodal shared space, which does not require large parallel corpora during training. Finally, (Elliott and Kádár, 2017) is the first to integrate the idea of multimodal shared space learning to help multimodal machine translation. Their multi-task architecture called "imagination" shares an encoder between a primary task of the classical encoder-decoder NMT and an auxiliary task of visual feature reconstruction.

Our VAG-NMT mechanism is inspired by (Elliott and Kádár, 2017), but has important differences. First, we modify the auxiliary task as a visual-text shared space learning objective instead of the simple image reconstruction objective. Second, we create a visual-text attention mechanism that captures the words that share a strong semantic meaning with the image, where the grounded visual-context has more impact on the translation. We show that these enhancements lead to improvement in multimodal translation quality over (Elliott and Kádár, 2017).

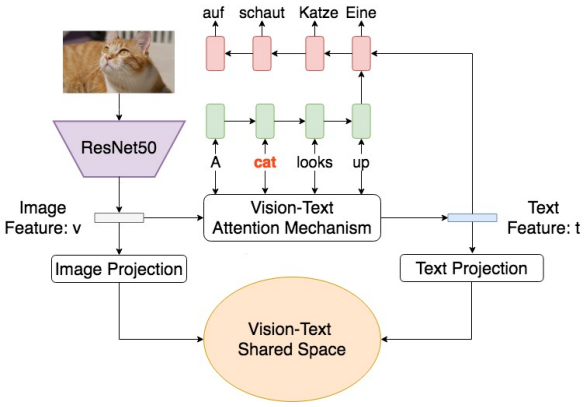

Figure 1: An overview of the VAG-NMT structure

3 Visual Attention Grounding NMT

Given a set of parallel sentences in language X and Y, and a set of corresponding images V paired with each sentence pair, the model aims to translate sentences $\{x_i\}_{i=1}^{N} \in X$ in language X to sentences $\{y_i\}_{i=1}^{N} \in Y$ in language Y with the assistance of images $\{v_i\}_{i=1}^{N} \in V$.

We treat the problem of multimodal machine

translation as a joint optimization of two tasks: (1) learning a robust translation model and (2) constructing a visual-language shared embedding that grounds the visual semantics with text. Figure 1 shows an overview of our VAG-NMT model. We adopt a state-of-the-art attention-based sequence-to-sequence structure (Bahdanau et al., 2014) for translation. For the joint embedding, we obtain the text representation using a weighted sum of hidden states from the encoder of the sequence-to-sequence model and we obtain the image representation from a pre-trained convnet. We learn the weights using a visual attention mechanism, which represents the semantic relatedness between the image and each word in the encoded text. We learn the shared space with a ranking loss and the translation model with a cross entropy loss.

The joint objective function is defined as:

$$J(\theta_T, \phi_V) = \alpha J_T(\theta_T) + (1-\alpha) J_V(\phi_V), \quad (1)$$

where J_T is the objective function for the sequence-to-sequence model, J_V is the objective function for joint embedding learning, θ_T are the parameters in the translation model, and ϕ_V are the parameters for the shared vision-language embedding learning, and α determines the contribution of the machine translation loss versus the visual grounding loss. Both J_T and J_V share the parameters of the encoder from the neural machine translation model. We describe details of the two objective functions in Section 3.2.

3.1 Encoder

We first encode an n-length source sentence $\{x\}$, as a sequence of tokens $x = \{x_1, x_2, \ldots, x_n\}$, with a bidirectional GRU (Schuster and Paliwal, 1997; Cho et al., 2014). Each token is represented by a one-hot vector, which is then mapped into an embedding e_i through a pre-trained embedding matrix. The bidirectional GRU processes the embedding tokens in two directions: left-to-right (forward) and right-to-left (backward). At every time step, the encoder's GRU cell generates two corresponding hidden state vectors: $\overrightarrow{h_i} = \overrightarrow{GRU}(h_{i-1}, e_i)$ and $\overleftarrow{h_i} = \overleftarrow{GRU}(h_{i-1}, e_i)$. The two hidden state vectors are then concatenated together to serve as the encoder hidden state vector of the source token at step i: $h_i = [\overleftarrow{h_i}, \overrightarrow{h_i}]$.

3.2 Shared embedding objective

After encoding the source sentence, we project both the image and text into the shared space to find a good distributed representation that can capture the semantic meaning across the two modalities. Previous work has shown that learning a multimodal representation is effective for grounding knowledge between two modalities (Kiros et al., 2014; Chrupala et al., 2015). Therefore, we expect the shared encoder between the two objectives to facilitate the integration of the two modalities and positively influence translation during decoding.

To project the image and the source sentence to a shared space, we obtain the visual embedding (v) from the pool5 layer of ResNet50 (He et al., 2015a) pre-trained on ImageNet classification (Russakovsky et al., 2015), and the source sentence embedding using the weighted sum of encoder hidden state vectors ($\{h_i\}$) to represent the entire source sentence (t). We project each $\{h_i\}$ to the shared space through an embedding layer. As different words in the source sentence will have different importance, we employ a visual-language attention mechanism—inspired by the attention mechanism applied in sequence-to-sequence models (Bahdanau et al., 2014)—to emphasize words that have the stronger semantic connection with the image. For example, the highlighted word "cat" in the source sentence in Fig. 1 has the more semantic connection with the image.

Specifically, we produce a set of weights $\beta = \{\beta_1, \beta_2, \ldots, \beta_n\}$ with our visual-attention mechanism, where the attention weight β_i for the i'th word is computed as:

$$\beta_i = \frac{\exp(z_i)}{\sum_{l=1}^{N} \exp(z_l)}, \quad (2)$$

and $z_i = \tanh(W_v v) \cdot \tanh(W_h h_i)$ is computed by taking the dot product between the transformed encoder hidden state vector h_i and the transformed image feature vector v, and W_v and W_h are the association transformation parameters.

Then, we can get a weighted sum of the encoder hidden state vectors $t = \sum_{i=1}^{n} \beta_i h_i$ to represent the semantic meaning of the entire source sentence. The next step is to project the source sentence feature vector t and the image feature vector v into the same shared space. The projected vector for text is: $t_{emb} = \tanh(W_{t_{emb}} t + b_{t_{emb}})$ and the projected vector for image is: $v_{emb} = \tanh(W_{v_{emb}} v + b_{v_{emb}})$.

We follow previous work on visual semantic embedding (Kiros et al., 2014) to minimize a pairwise ranking loss to learn the shared space:

$$J_V(\phi_V) = \sum_p \sum_k \max\{0, \gamma - s(v_p, t_p) + s(v_p, t_{k\neq p})\} + \sum_k \sum_p \max\{0, \gamma - s(t_k, v_k) + s(t_k, v_{p\neq k})\}, \quad (3)$$

where γ is a margin, and s is the cosine distance between two vectors in the shared space. k and p are the indexes of the images and text sentences, respectively. $t_{k\neq p}$ and $v_{p\neq k}$ are the contrastive examples with respect to the selected image and the selected source text, respectively. When the loss decreases, the distance between a paired image and sentence will drop while the distance between an unpaired image and sentence will increase.

In addition to grounding the visual context into the shared encoder through the multimodal shared space learning, we also initialize the decoder with the learned attention vector t such that the words that have more relatedness with the visual semantics will have more impact during the decoding (translation) stage. However, we may not want to solely rely on only a few most important words. Thus, to produce the initial hidden state of the decoder, we take a weighted average of the attention vector t and the mean of encoder hidden states:

$$s_0 = \tanh(W_{init}(\lambda t + (1-\lambda)\frac{1}{N}\sum_i^N h_i)), \quad (4)$$

where λ determines the contribution from each vector. Through our experiments, we find the best value for λ is 0.5.

3.3 Translation objective

During the decoding stage, at each time step j, the decoder generates a decoder hidden state s_j from a conditional GRU cell (Sennrich et al., 2017) whose input is the previously generated translation token y_{j-1}, the previous decoder hidden state s_{j-1}, and the context vector c_j at the current time step:

$$s_j = \text{cGRU}(s_{j-1}, y_{j-1}, c_j) \quad (5)$$

The context vector c_j is a weighted sum of the encoder hidden state vectors, and captures the relevant source words that the decoder should focus on when generating the current translated token y_j. The weight associated with each encoder hidden state is determined by a feed-forward network. From the hidden state s_j we can predict the conditional distribution of the next token y_j with a fully-connected layer W_o given the previous token's language embedding e_{j-1}, the current hidden state d_j and the context vector for current step c_j:

$$p(y_j|y_{j-1}, x) = \text{softmax}(W_o o_t), \quad (6)$$

where $o_t = \tanh(W_e e_{j-1} + W_d d_j + W_c c_j)$. The three inputs are transformed with W_e, W_d, and W_c, respectively and then summed before being fed into the output layer.

We train the translation objective by optimizing a cross entropy loss function:

$$J_T(\theta_T) = -\sum_j \log p(y_j|y_{j-1}, x) \quad (7)$$

By optimizing the objective of the translation and the multimodal shared space learning tasks jointly along with the visual-language attention mechanism, we can simultaneously learn a general mapping between the linguistic signals in two languages and grounding of relevant visual content in the text to improve the translation.

4 IKEA Dataset

Previous available multimodal machine translation models are only tested on image caption datasets, we, therefore, propose a new dataset, IKEA, that has the real-world application of international online shopping. We create the dataset by crawling commercial products' descriptions and images from IKEA and UNIQLO websites. There are 3,600 products and we plan to expand the data in the future. Each sample is composed of the web-crawled English description of a product, an image of the product, and the web-crawled German or French description of the product.

Different than the image caption datasets, the German or French sentences in the IKEA dataset is not an exact parallel translation of its English sentence. Commercial descriptions in different languages can be vastly different in surface form but still keep the core semantic meaning. We think IKEA is a good data set to simulate real-world multimodal translation problems. The sentence in the IKEA dataset contains 60-70 words on average, which is five times longer than the average text length in Multi30K (Elliott et al., 2016). These sentences not only describe the visual appearance of the product, but also the product usage. Therefore, capturing the connection between

	English → German		English → French	
Method	BLEU	METEOR	BLEU	METEOR
Imagination (Elliott and Kádár, 2017)	30.2	51.2	N/A	N/A
LIUMCVC (Caglayan et al., 2017)	31.1 ± 0.7	52.2 ± 0.4	52.7 ± 0.9	69.5 ± 0.7
Text-Only NMT	$\mathbf{31.6 \pm 0.5}$	52.2 ± 0.3	53.5 ± 0.7	70.0 ± 0.7
VAG-NMT	$\mathbf{31.6 \pm 0.3}$	52.2 ± 0.3	$\mathbf{53.8 \pm 0.3}$	$\mathbf{70.3 \pm 0.5}$

Table 1: Translation results on the Multi30K dataset

	English → German		English → French	
Method	BLEU	METEOR	BLEU	METEOR
Imagination (Elliott and Kádár, 2017)	28.0	**48.1**	N/A	N/A
LIUMCVC (Caglayan et al., 2017)	27.1 ± 0.9	47.2 ± 0.6	43.5 ± 1.2	63.2 ± 0.9
Text-Only NMT	27.9 ± 0.6	47.8 ± 0.6	44.6 ± 0.6	64.2 ± 0.5
VAG-NMT	$\mathbf{28.3 \pm 0.6}$	48.0 ± 0.5	$\mathbf{45.0 \pm 0.4}$	$\mathbf{64.7 \pm 0.4}$

Table 2: Translation results on the Ambiguous COCO dataset

visual semantics and the text is more challenging on this dataset. The dataset statistics and an example of the IKEA dataset is in Appendix A.

5 Experiments and Results

5.1 Datasets

We evaluate our proposed model on three datasets: Multi30K (Elliott et al., 2016), Ambiguous COCO (Elliott et al., 2017), and our newly-collected IKEA dataset. The Multi30K dataset is the largest existing human-labeled dataset for multimodal machine translation. It consists of 31,014 images, where each image is annotated with an English caption and manual translations of image captions in German and French. There are 29,000 instances for training, 1,014 instances for validation, and 1,000 for testing. Additionally, we also evaluate our model on the Ambiguous COCO Dataset collected in the WMT2017 multimodal machine translation challenge (Elliott et al., 2017). It contains 461 images from the MSCOCO dataset (Lin et al., 2014), whose captions contain verbs with ambiguous meanings.

5.2 Setting

We pre-process all English, French, and German sentences by normalizing the punctuation, lower casing, and tokenizing with the Moses toolkit. A Byte-Pair-Encoding (BPE) (Sennrich et al., 2015) operation with 10K merge operations is learned from the pre-processed data and then applied to segment words. We restore the original words by concatenating the subwords segmented by BPE in post-processing. During training, we apply early stopping if there is no improvement in BLEU score on validation data for 10 validation steps. We apply beam search decoding to generate translation with beam size equal to 12. We evaluate the performance of all models using BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014). The setting used in IKEA dataset is the same as Multi30K, except that we lower the default batch size from 32 to 12; since IKEA dataset has long sentences and large variance in sentence length, we use smaller batches to make the training more stable. Full details of our hyperparameter choices can be found in Appendix B. We run all models five times with different random seeds and report the mean and standard deviation.

5.3 Results

We compare the performance of our model against the state-of-the-art multimodal machine translation approaches and the text-only baseline. The idea of our model is inspired by the "Imagination" model (Elliott and Kádár, 2017), which unlike our models, simply averages the encoder hidden states for visual grounding learning. As "Imagination" does not report its performance on Multi30K 2017 and Ambiguous COCO in its original paper, we directly use their result reported in the WMT 2017 shared task as a comparison. LIUMCVC is the best multimodal machine translation model in WMT 2017 multimodal machine translation challenge and exploits visual information with several different methods. We always compare our VAG-NMT with the method that has been reported to

	English → German		English → French	
Method	BLEU	METEOR	BLEU	METEOR
LIUMCVC-Multi	59.9 ± 1.9	63.8 ± 0.4	58.4 ± 1.6	64.6 ± 1.8
Text-Only NMT	61.9 ± 0.9	65.6 ± 0.9	65.2 ± 0.7	$\mathbf{69.0 \pm 0.2}$
VAG-NMT	$\mathbf{63.5 \pm 1.2}$	$\mathbf{65.7 \pm 0.1}$	$\mathbf{65.8 \pm 1.2}$	68.9 ± 1.4

Table 3: Translation results on the IKEA dataset

have the best performance on each dataset.

Our VAG-NMT surpasses the results of the "Imagination" model and the LIUMCVC's model by a noticeable margin in terms of BLEU score on both the Multi30K dataset (Table 1) and the Ambiguous COCO dataset (Table 2). The METEOR score of our VAG-NMT is slightly worse than that of "Imagination" for English -> German on Ambiguous COCO Dataset. This is likely because the "Imagination" result was produced by ensembling the result of multiple runs, which typically leads to 1-2 higher BLEU and METEOR points compared to a single run. Thus, we expect our VAG-NMT to outperform the "Imagination" baseline if we also use an ensemble.

We observe that our multimodal VAG-NMT model has equal or slightly better result compared to the text-only neural machine translation model on the Multi30K dataset. On the Ambiguous COCO dataset, our VAG-NMT demonstrates clearer improvement over this text-only baseline. We suspect this is because Multi30K does not have many cases where images can help improve translation quality, as most of the image captions are short and simple. In contrast, Ambiguous COCO was purposely curated such that the verbs in the captions can have ambiguous meaning. Thus, visual context will play a more important role in Ambiguous COCO; namely, to help clarify the sense of the source text and guide the translation to select the correct word in the target language.

We then evaluate all models on the IKEA dataset. Table 3 shows the results. Our VAG-NMT has a higher value in BLEU and a comparable value in METEOR compared to the Text-only NMT baseline. Our VAG-NMT outperforms LIUMCVC's multimodal system by a large margin, which shows that the LIUMCVC's multimodal's good performance on Multi30K does not generalize to this real-world product dataset. The good performance may come from the visual attention mechanism that learns to focus on text segments that are related to the images. Such attention therefore teaches the decoder to apply the visual context to translate those words. This learned attention is especially useful for datasets with long sentences that have irrelevant text information with respect to the image.

(a) *a cyclist is wearing a helmet*

(b) *a black dog and his favorite toys.*

Figure 2: Top five images retrieved using the given caption. The original corresponding image of the caption is highlighted with a red bounding box.

5.4 Multimodal embedding results

To assess the learned joint embedding, we perform an image retrieval task evaluated using the Recall@K metric (Kiros et al., 2014) on the Multi30K dataset.

We project the image feature vectors $V = \{v_1, v_2, \ldots, v_n\}$ and their corresponding captions $S = \{s_1, s_2, \ldots, s_n\}$ into a shared space. We follow the experiments conducted by the previous visual semantic embedding work (Kiros et al., 2014), where for each embedded text vector, we find the closest image vectors around it based on the cosine similarity. Then we can compute the recall rate of the paired image in the top K nearest neighbors, which is also known as $R@K$ score. The shared space learned with VAG-NMT achieves 64% R@1, 88.6% R@5, and 93.8% R@10 on Multi30K, which demonstrates the good quality of the learned shared semantics. We also achieved 58.13% R@1, 87.38% R@5 and 93.74% R@10 on IKEA dataset; 41.35% R@1, 85.48% R@5 and 92.56% R@10 on COCO dataset. Besides the quantitative results, we also show several qualitative results in Figure 2. We show the top five images retrieved by the example captions.

a person is skiing or snowboarding down a mountainside .

a mountain climber trekking through the snow with a pick and a blue hat .

the snowboarder is jumping in the snow .

two women are water rafting .

three people in a blue raft on a river of brown water .

people in rafts watch as two men fall out of their own rafts .

Figure 3: Visual-text attention score on sample data from Multi30K. The first and second rows show the three closest images to the caption *a person is skiing or snowboarding down a mountainside* and *two woman are water rafting*, respectively. The original caption is listed under each image. We highlight the three words with highest attention in red.

The images share "cyclist", "helmet", and "dog" mentioned in the caption.

5.5 Human evaluation

We use Facebook to hire raters that speak both German and English to evaluate German translation quality. As Text-Only NMT has the highest BLEU results among all baseline models, we compare the translation quality between the Text-Only and the VAG-NMT on all three datasets. We randomly selected 100 examples for evaluation for each dataset. The raters are informed to focus more on semantic meaning than grammatical correctness when indicating the preference of the two translations. They are also given the option to choose a tie if they cannot decide. We hired two raters and the inter-annotator agreement is 0.82 in Cohen's Kappa.

We summarize the results in Table 4, where we list the number of times that raters prefer one translation over another or think they are the same quality. Our VAG-NMT performs better than Text-Only NMT on MSCOCO and IKEA dataset, which correlates with the automatic evaluation metrics. However, the result of VAG-NMT is slightly worse than the Text-Only NMT on the Multi30K test dataset. This also correlates with the result of automatic evaluation metrics.

Finally, we also compare the translation quality between LIUMCVC multimodal and VAG-NMT on 100 randomly selected examples from the IKEA dataset. VAG-NMT performs better than LIUMCVC multimodal. The raters prefer our VAG-NMT in 91 cases, LIUMCVC multimodal in 68 cases, and cannot tell in 47 cases.

	MSCOCO	Multi30K	IKEA
Text-Only NMT	76	**72**	75
VAG-NMT	**94**	71	**82**
Tie	30	57	43

Table 4: Human evaluation results

6 Discussion

To demonstrate the effectiveness of our visual attention mechanism, we show some qualitative examples in Figure 3. Each row contains three images that share similar semantic meaning, which are retrieved by querying the image caption using our learned shared space. The original caption of each image is shown below each image. We highlight the three words in each caption that have the highest weights assigned by the visual attention mechanism.

In the first row of Figure 3, the attention mechanism assigns high weights to the words "skiing", "snowboarding", and "snow". In the second row, it assigns high attention to "rafting" or "raft" for every caption of the three images. These examples demonstrate evidence that our attention mechanism learns to assign high weights to words that have corresponding visual semantics in the image.

We also find that our visual grounding attention captures the dependency between the words that

Source Caption: a tennis player is moving to the side and is gripping his racquet with both hands .

Text-Only NMT: ein tennisspieler bewegt sich um die seite und greift mit beiden händen an den boden .

VAG-NMT: ein tennisspieler bewegt sich zur seite und greift mit beiden händen den schläger .

Source Caption: three skiers skiing on a hill with two going down the hill and one moving up the hill .

Text-Only NMT: drei skifahrer fahren auf skiern einen hügel hinunter und eine person fährt den hügel hinunter .

VAG-NMT: drei skifahrer auf einem hügel fahren einen hügel hinunter und ein bewegt sich den hügel hinauf .

Source Caption: a blue , yellow and green surfboard sticking out of a sandy beach .

Text-Only NMT: ein blau , gelb und grünes surfbrett streckt aus einem sandstrand .

VAG-NMT: ein blau , gelb und grüner surfbrett springt aus einem sandstrand .

Figure 4: Translations generated by VAG-NMT and Text-Only NMT. VAG-NMT performs better in the first two examples, while Text-Only NMT performs better in the third example. We highlight the words that distinguish the two systems' results in red and blue. Red words are marked for better translation from VAG-NMT and blue words are marked for better translation from Text-Only NMT.

have strong visual semantic relatedness. For example, in Figure 3, words, such as "raft","river", and "water", with high attention appear in the image together. This shows that the visual dependence information is encoded into the weighted sum of attention vectors which is applied to initialize the translation decoder. When we apply the sequence-to-sequence model to translate a long sentence, the encoded visual dependence information strengthens the connection between the words with visual semantic relatedness. Such connections mitigate the problem of standard sequence-to-sequence models tending to forget distant history. This hypothesis is supported by the fact that our VAG-NMT outperforms all the other methods on the IKEA dataset which has long sentences.

Lastly, in Figure 4 we provide some qualitative comparisons between the translations from VAG-NMT and Text-Only NMT. In the first example, our VAG-NMT properly translates the word "racquet" to "den schläger", while the Text-Only NMT mistranslated it to "den boden" which means "ground" in English. We suspect the attention mechanism and visual shared space capture the visual dependence between the word "tennis" and "racquet". In the second example, our VAG-NMT model correctly translates the preposition "up" to "hinauf" but Text-Only NMT mistranslates it to "hinunter" which means "down" in English. We consistently observe that VAG-NMT translates prepositions better than Text-Only NMT. We think it is because the pre-trained convnet features captured the relative object position that leads to a better preposition choice. Finally, in the third example, we show a failure case where Text-Only NMT generates a better translation. Our VAG-NMT mistranslates the verb phrase "sticking out" to "springt aus" which means "jump out" in German, while Text-Only NMT translates to "streckt aus", which is correct. We find that VAG-NMT often makes mistakes when translating verbs. We think it is because the image vectors are pre-trained on an object classification task, which does not have any human action information.

7 Conclusion and Future Work

We proposed a visual grounding attention structure to take advantage of the visual information to perform machine translation. The visual attention mechanism and visual context grounding module help to integrate the visual content into the sequence-to-sequence model, which leads to better translation quality compared to the model with only text information. We achieved state-of-the-art results on the Multi30K and Ambiguous COCO dataset. We also proposed a new product dataset, IKEA, to simulate a real-world online product description translation challenge.

In the future, we will continue exploring different methods to ground the visual context into the translation model, such as learning a multimodal shared space across image, source language

text, as well as target language text. We also want to change the visual pre-training model from an image classification dataset to other datasets that have both objects and actions, to further improve translation performance.

Acknowledge

We would like to thank Daniel Boyla for providing insightful discussions to help with this research. We also want to thank Chunting Zhou and Ozan Caglayan for suggestions on machine translation model implementation. This work was supported in part by NSF CAREER IIS-1751206.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. *CoRR*, abs/1707.04481.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017a. Doubly-attentive decoder for multi-modal neural machine translation. *CoRR*, abs/1702.01287.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017b. Incorporating global visual features into attention-based neural machine translation. *CoRR*, abs/1701.06521.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017c. Multilingual multi-modal embeddings for natural language processing. *CoRR*, abs/1702.01101.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Grzegorz Chrupala, Ákos Kádár, and Afra Alishahi. 2015. Learning language through pictures. *CoRR*, abs/1506.03694.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *In Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. *CoRR*, abs/1710.07177.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *CoRR*, abs/1605.00459.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. *CoRR*, abs/1705.04350.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. *CoRR*, abs/1707.07601.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015a. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015b. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852.

Jindrich Helcl and Jindrich Libovický. 2017. CUNI system for the WMT17 multimodal translation task. *CoRR*, abs/1707.04550.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *WMT*.

Andrej Karpathy and Fei-Fei Li. 2014. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Mingbo Ma, Dapeng Li, Kai Zhao, and Liang Huang. 2017. OSU multimodal machine translation system report. *CoRR*, abs/1710.02718.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2012. Understanding the exploding gradient problem. *CoRR*, abs/1211.5063.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 91–99, Cambridge, MA, USA. MIT Press.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. *CoRR*, abs/1703.04357.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. 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.

Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, Graham Neubig, and Satoshi Nakamura. 2017. Nictnaist system for wmt17 multimodal translation task. In *WMT*.

A IKEA Dataset Stats and Examples

We summarize the statistics of our IKEA dataset in Figure 5, where we demonstrate the information about the number of tokens, the length of the product description and the vocabulary size. We provide one example of the IKEA dataset in Figure 6.

Pair	EN-DE		EN-FR	
Language	EN	DE	EN	FR
Tokens	256355	216892	239966	275251
Min length	6	6	6	6
Max length	343	324	334	469
Avg length	71.4	60.4	72.2	82.9
Std dev	46.3	39.1	47.2	54.7
Vocabulary	6601	10468	6442	7575

Figure 5: Statistic of the IKEA dataset

(a) *Product image*

the 4 large drawers on casters give you an extra storage space under the bed . adjustable bed sides allow you to use mattresses of different thicknesses . 17 slats of layer glued birch adjust to your body weight and increase the suppleness of the mattress . wipe clean using a damp cloth and a mild cleaner . wipe dry with a clean cloth.

(b) *Source description*

die 4 geräumigen schubladen auf rollen sorgen für zusätzlichen stauraum unter dem bett . durch verstellbare bettseiten können matratzen in verschiedenen stärken verwendet werden . mit feuchtem tuch (evtl . mit mildem reinigungsmittel) abwischen . mit trockenem tuch nachwischen .

(c) *Target description in German*

Figure 6: An example of product description and the corresponding translation in German from the **IKEA dataset**. Both descriptions provide an accurate caption for the commercial characteristics of the product in the image, but the details in the descriptions are different.

B Hyperparameter Settings

In this Appendix, we share details on the hyperparameter settings for our model and the training process. The word embedding size for both encoder and decoder are 256. The Encoder is a one-layer bidirectional recurrent neural network with Gated Recurrent Unit (GRU), which has a hidden size of 512. The decoder is a recurrent neural network with conditional GRU of the same hidden size. The visual representation is a 2048-dim vector extracted from the `pool5` layer of a pre-trained ResNet-50 network. The dimension of the shared visual-text semantic embedding space is 512. We set the decoder initialization weight

value λ to 0.5.

During training, we use Adam (Kingma and Ba, 2014) to optimize our model with a learning rate of $4e-4$ for German Dataset and $1e-3$ for French dataset. The batch size is 32. The total gradient norm is clipped to 1 (Pascanu et al., 2012). Dropout is applied at the embedding layer in the encoder, context vectors extracted from the encoder and the output layer of the decoder. For Multi30K German dataset the dropout probabilities are $(0.3, 0.5, 0.5)$ and for Multi30K French dataset the dropout probabilities are $(0.2, 0.4, 0.4)$. For the Multimodal shared space learning objective function, the margin size γ is set to 0.1. The objective split weight α is set to 0.99. We initialize the weights of all the parameters with the method introduced in (He et al., 2015b).

C Ablation Analysis on Visual-Text Attention

We conducted an ablation test to further evaluate the effectiveness of our visual-text attention mechanism. We created two comparison experiments where we reduced the impact of visual-text attention with different design options. In the first experiment, we remove the visual-attention mechanism in our pipeline and simply use the mean of the encoder hidden states to learn the shared embedding space. In the second experiment, we initialize the decoder with just the mean of encoder hidden states without the weighted sum of encoder states using the learned visual-text attention scores.

We run both experiments on Multi30K German Dataset five times and demonstrate the results in table 5. As can be seen, the performance of the changed translation model is obviously worse than the full VAG-NMT in both experiments. This observation suggests that the visual-attention mechanism is critical in improving the translation performance. The model improvement comes from the attention mechanism influencing the model's objective function and decoder's initialization.

	English → German	
Method	BLEU	METEOR
-attention in shared embedding	30.5 ± 0.6	51.7 ± 0.4
-attention in initialization	30.8 ± 0.8	51.9 ± 0.5
VAG-NMT	$\mathbf{31.6 \pm 0.6}$	$\mathbf{52.2 \pm 0.3}$

Table 5: Ablation analysis on visual-text attention mechanism in the Multi30K German dataset.