# DO WE REALLY NEED UNIVERSAL VISUAL REPRESENTATION IN MULTIMODAL NMT ?

**Xin Cheng**
Wangxuan Institute of Computer Technology, Peking University
Center for Data Science, AAIS, Peking University
chengxin1998@stu.pku.edu.cn

## ABSTRACT

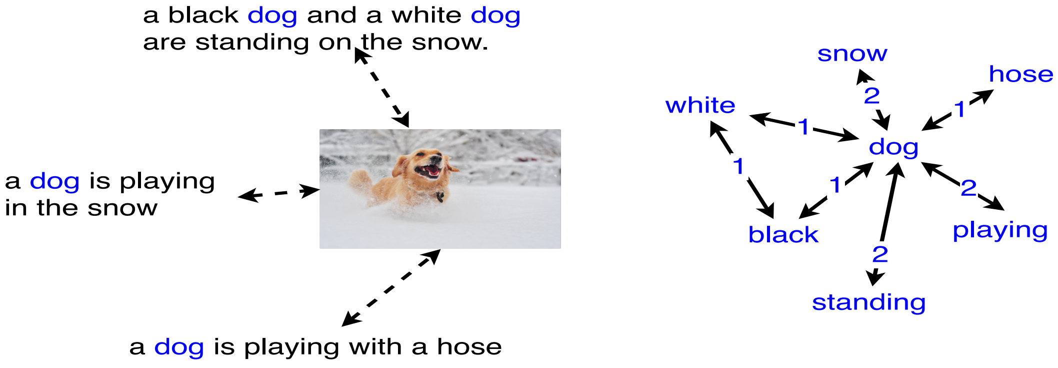
Multimodal Machine Translation (MMT) aims to introduce information from other modality, generally static images, to improve the translation quality. Because of the difficulty of manual annotation of relevant image for every translation pairs, previous works mainly verify their methods on MMT specific dataset (*Multi30k*), which is of tiny size (29k) compared with other translation tasks and unable to benefit text-only setting. To overcome this bottleneck, Universal Visual Representation NMT (Zhang et al., 2019) retrieves visual representation from image-paired *Multi30k* dataset and applies those visual features in text-only NMT, observing significant improvement over baseline models. In this paper, however, we would demonstrate that it is not **Visual** but **Universal** that truly matters behind the scene. Specifically, we propose a Universal Textual Representation model without introducing any visual information and achieve better results with less parameters. Our code is open sourced at https://github.com/Hannibal046/UTR_NMT.

## 1 INTRODUCTION

Multimodal machine translation (MMT) is a novel machine translation (MT) task which aims at designing better translation systems using context from an additional modality, usually images (Barrault et al., 2018). Current works focus on the dataset named *Multi30k* (Elliott et al., 2016), a multilingual extension of *Flickr30k* dataset with translations of the English image descriptions into different languages. Despite the promising results MMT systems have achieved (Calixto et al., 2017; Yao & Wan, 2020; Delbrouck & Dupont, 2017), the effectiveness heavily relies on the availablity of bilingual parallel sentence pairs with manual image annotation, which hinders the image applicability to the machine translation. As a result, the visual information is only applied to the translation task over a small and specific dataset, not feasible for text-only and low-resource NMT.

This paper, Zhang et al. (2019), presents a universal visual representation (UVR) method relying only on image-monolingual annotations instead of the existing approach that depends on the image-bilingual annotations, thus breaking the bottleneck of using visual information in NMT. In detail, UVR transform the existing sentence-image pairs into a topic-image lookup table from a small-scale multimodal data set *Multi30k*. During the training and decoding process, a group of images with a similar topic to the source sentence will be retrieved from the topic-image lookup table learned by the term frequency-inverse document frequency, and thus is encoded as image representations by a pretrained ResNet. UVR model then use a attention layer to fuse the image representations and the original source sentence representations as input to the decoder for predicting target translations. In particular, the proposed approach can be integrated into the text-only NMT model without annotating large-scale bilingual parallel corpora. The experimental results in text-only translation tasks and MMT task verifies the effectiveness of UVR model.

However, since neural model is especially capable of modeling the co-occurrence among different inputs (Wiedemann et al., 2019) and the UVR model exactly does the same thing by explicitly connecting different samples via image features as shown in Figure 1(a). This motivates us to question about whether the improvement given by UVR comes from the visual features or the co-occurrence between similar translation samples? Concretely, we present a Universal Texual


a black dog and a white dog
are standing on the snow.
a dog is playing
in the snow
a dog is playing with a hose
snow
hose
white
dog
black
playing
standing
1
2
1
1
1
2
2

(a)

(b)

Figure 1: Topic-image lookup table and Keyword Net.

Representation (UTR) model relying only on text-based corpus and achieve better results than visual feature augmented NMT model with less trainable parameters.

## 2 Methodology

### 2.1 Universal Textual Representation

In this section, we will introduce the our Universal Textual Representation model. Generally, the default input setting of the UVR is a one sentence with multiple relevant images retrieved from topic-image lookup table according the topic word extracted from the sentence. Our base intuition is to transform the existing multiple images into multiple relevant keywords retrieved from a pre-constructed Keyword Net, which assumes the sentences with more similar keywords would be similar to each other. Consequently, a sentence can possess a group of keywords which share the most co-occurrence in the corpus, as shown in Figure 1(b).

**Keyword Net Conversion** To keep align with previous work Zhang et al. (2019), we adopt the same filtering method to extract the topic word of the sentence through the term frequency-inverse document frequency (TF-IDF) inspired by Chen et al. (2019). Specially, given a sentence $X = \{x_1, x_2, ...x_I\}$ of length $I$, $X$ is first filtered by a stopword list and the sentence is treated as a document $g$. We then compute TF-IDF $TI_{i,j}$ for each word $x_i$ in $g$:

$$TI_{i,j} = \frac{o_{i,j}}{\sum_k o_{k,j}} \times log \frac{|G|}{1 + |j : x_i \in g|} \tag{1}$$

where $o_{i,j}$ representes the number of occurrences of the word $x_i$ in the input sentence $g$, $|G|$ the total number of source language sentences in the training data, and $|j : x_i \in g|$ the number of source sentences including word $x_i$ in the training data. We then select top-$w$ high TF-IDF words $T = \{k_1, k_2, ...k_w\}$ as the representative words of the sentence $g$. After preprocessing the whole corpus, for each sentence we have a keyword list. And we construct a Keyword Net (KN) by making every keyword a node and connecting keywords that appear together in one sentence. The weight of each edge in the KN is the number of co-occurrence for two connected keywords in the training corpus. Different from UVR, whose topic-image lookup table is under the constraint of the scale of annotated sentence-pair dataset (i.e. *Multi30k*), our method is completely text-based and can generalize to massive unannotated raw text based on the assumption that the co-occurrence of keywords can be better captured with the increasement of training corpus.

**Keyword Retrieval** For the input sentence, we first obtain its keywords according to the text preprocessing method described above. Then we retrieve the neighbors of the keywords in the KN with their co-occurrence times and group all retrieved keywords together to form a keyword list $\mathcal{G}$. We observe that one keyword may appear multiple times with the given keyword extracted from the source sentence so it's obvious that the same co-occurrence pattern shows up multiple times in the training corpus which is useful for model to infer useful clues to translate current sentence. Then

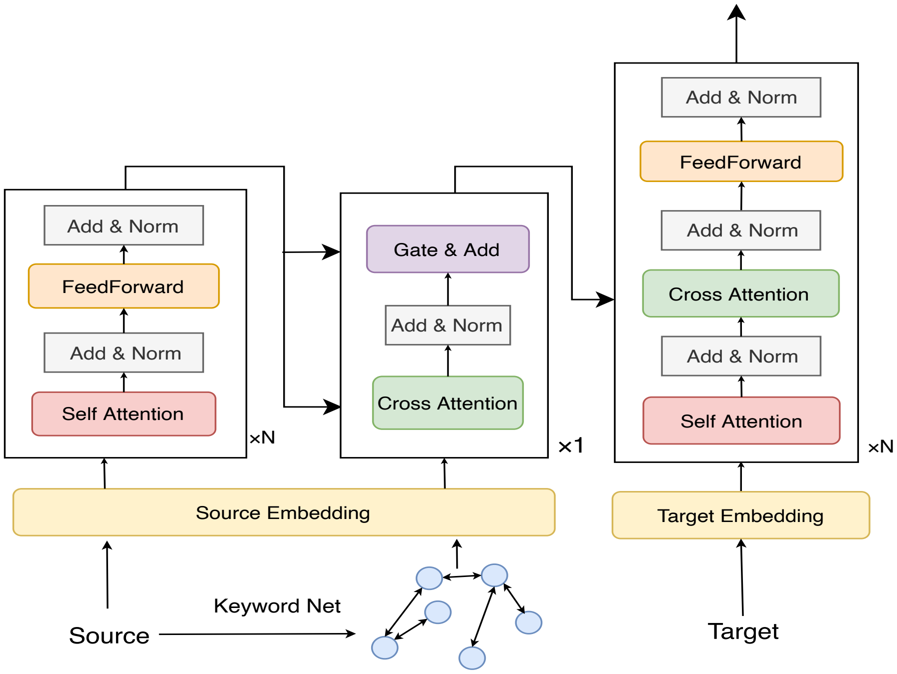


Figure 2: Overview of our model.

we sort the keyword list $\mathcal{G}$ according to the co-occurrence in a descending order and get the top-$m$ relevant keywords for each source sentence.

At test time, the process of getting images is done using the KN built by the training set, so we do not need to use the images from the dev and test sets in *Multi30k*. Intuitively, we do not require visual annotation, but solely rely on the co-occurrence of the keywords, which is more simple and can be easily generalize to larger dataset. In this way, we call our method Universal Textual Retrieval model.

### 2.2 NMT WITH UNIVERSAL TEXTUAL REPRESENTATION

For the model part, to give a fair comparison between the our method and UVR, we adopt the same transformer model (Vaswani et al., 2017) as in Zhang et al. (2019) except that our model don't use pretrained ResNet as visual embedding. Instead, we share the parameters of source sentence embedding and keyword embedding, leading to less parameters compared with UVR. Our model is shown in Figure 2.

## 3 EXPERIMENTS

### 3.1 DATA

The proposed model is evaluated in the *Multi30k* dataset, which is a standard corpus for MMT evaluation. The *Multi30k* contains 29K English→ {German,French,Czech} parallel sentence pairs. The 1014 English→ {German,French,Czech} sentence pairs are as dev set. The test sets are test2016, test2017-flickr, test2017-mscoco, test2018 with 1000 pairs for each. The details of the dataset is shown in Table 1.

### 3.2 SYSTEM SETTING

**Keyword Retrieval Implementation** To keep in line with previous work, we also used 29,000 English sentence from *Multi30k* to build our Keyword Net. We use the tokenized version of *Multi30k*

|  | #Train | #Dev | #Avg. src tokens | #Avg. trg tokens | #src vocab | #trg vocab |
|---|---|---|---|---|---|---|
| En→De | 29,000 | 1,014 | 12.97 | 12.10 | 5,922 | 7,859 |
| En→Cs | 29,000 | 1,014 | 12.97 | 10.50 | 5,922 | 10,000 |
| En→Fr | 29,000 | 1,014 | 12.97 | 13.99 | 5,9222 | 6,482 |

Table 1: Data statistics of *Multi30k*.

| System | Params | Test2016 |  |  | Test2017-C |  | Test2017-F |  | Test2018 |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | De | Cs | Fr | De | Fr | De | Fr | De | Cs | Fr |
| Transformer † | - | 35.59 | - | 57.88 | 26.31 | 48.55 | - | - | - | - | - |
| UVR † | - | 35.72 | - | 58.32 | 26.87 | 48.69 | - | - | - | - | - |
| Transformer | 51.20 M | 34.07 | 29.89 | 56.61 | 28.11 | **51.08** | 25.45 | 24.75 | 35.32 | 24.33 | **42.99** |
| UVR | 53.30 M | 35.93 | 30.91 | 56.52 | 29.09 | 49.82 | 25.94 | 24.96 | 34.94 | 25.25 | 42.24 |
| UVR(unfix) | 112.69 M | 35.29 | 28.91 | 57.43 | 29.45 | 50.58 | 24.48 | 23.49 | **36.64** | 25.51 | 42.14 |
| UVR(random) | 112.69 M | 35.42 | 28.62 | 54.98 | 28.63 | 47.09 | 25.78 | 23.77 | 35.82 | 24.56 | 39.87 |
| Ours | 52.51 M | **37.73** | **31.89** | **58.09** | **30.00** | 50.40 | **26.91** | **25.59** | 35.57 | **25.69** | 42.48 |

Table 2: BLEU scores on 10 test dataset of *Multi30k*. † denotes that the results was reported in the Zhang et al. (2019). UVR(unfix) means the image embedding layer initialized by ResNet is unfixed and can be updated during training. UVR(random) means we do not use pretrained embedding from ResNet but randomly initialize the image embedding layer. Considering that the parameters of the model is language-dependent because of embedding size, parameters of all models are calculated when De as target language.

and set the minimal frequency as 2 to construct our vocabulary for both source and target language. We selected top-8 high TF-IDF words, and the number of retrieved keywords was set 5. After prepocessing, the size of our KN is , and the average out degree of each node is .

**Baseline** Our baseline model is text-only transformer and UVR model. We used 6 layers for the encoder and the decoder. All other settings are the same as the Transformer Base model in Vaswani et al. (2017).

**Model Implementation** For simplicity, we use the tokenized version of *Multi30k* and use the space as the default tokenizer. We use Adam optimizer with noam scheduler. The learning rate was set 2e-6 with a 1200 warm-up steps across all experiments. The batch size of is 128, the epoch is 10 and all experiments are conducted on one RTX3090 Graphic Card. We adopt Sacrebleu[1] to calculate the BLEU score. For evaluation, we validate the model on the dev set after every epoch and choose the model checkpoint with highest BLEU score in dev set to do evaluation in the test set. During decoding, the beam size was set to 5. All models were implemented with *Huggingface/Transformer*[2] by ourselves.

### 3.3 RESULTS

Table 2 shows the results for *Multi30k*. Our implemented Transformer Base and UVR model show similar BLEU scores with Zhang et al. (2019). As can be seen, we have the following observation:

(1) The proposed Universal Textual Representation model get consistent improvement over Transformer Base model in terms of BLEU score which is non-trivial since we don't introduce any extra training data or content information from other modality.

(2) With less parameters, the UTR model achieves comparable even better results with UVR model across 10 test set which demonstrates the effectiveness of explicitly incorporating co-occurrence information into the NMT model. And this also verifies our assumption that the reason behind the improved performance of UVR doesn't lie in the injected visual information but benefits from explicitly modeling of co-occurrence between training samples. However, this also cast a new question: how can we simultaneously use textual and visual information in neural machine translation system ?

[1] `https://github.com/mjpost/sacrebleu`

[2] `https://github.com/huggingface/transformers`

| System | Test2016 | | | Test2017-C | | Test2017-F | | Test2018 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | De | Cs | Fr | De | Fr | De | Fr | De | Cs | Fr |
| Transformer | 34.07 | 29.89 | 56.61 | 28.11 | **51.08** | 25.45 | 24.75 | 35.32 | 24.33 | **42.99** |
| Ours | **37.73** | **31.89** | **58.09** | **30.00** | 50.40 | **26.91** | 25.59 | 35.57 | 25.69 | 42.48 |
| Transformer_Concat | 35.91 | 30.79 | 56.99 | 26.97 | 51.07 | 24.87 | **26.86** | **36.12** | **25.72** | 42.50 |

Table 3: BLEU score of our UTR model and transformer model with concatenated inputs.

# 4 ANALYSIS

## 4.1 WHY DOES KEYWORD NET WORK

The contribution of our Universal Textual Representation can be two folds: (1) Explicitly Modeling co-occurrence through keywords rather than images, which can be easily extended to massive unannotated corpus and have better generalization ability based on the assumption that the co-occurrence between words can be better captured with the rise of corpus size. (2) The UTR demonstrates that the key intuition behind Universal Visual Representation is **Universal** rather than **Visual**. According the Distributional Hypothesis, which states that words that occur in similar contexts tend to have similar meanings, the key insight of UTR is easy to follow: sentences that have more overlapping keywords tend to have similar meanings.

## 4.2 CORRELATION BETWEEN TRANSLATION MEMORIES

Translation memory (TM) is basically a database of segmented and paired source and target texts that translators can access in order to re-use previous translations while translating new texts (Christensen & Schjoldager, 2010). Explicitly retrieving similar translation memories (TM) from training set has been proved useful in text-based NMT system(Gu et al., 2018; Cai et al., 2021). Considering that UTR and TM-augmented NMT both retrieve useful hints from training corpus to boost the translation quality, it is interesting to verify the connection between these two methods. Following Xu et al. (2020), we directly concatenate the retrieved keywords to the source sentence with a [SEP] between them. The results is shown in table 3. As can be seen, the transformer_concate method surpasses base transformer in $6/10$ tasks

## 4.3 INFLUENCE OF THE THRESHOLD $w$ AND RETRIEVED NUMBER $m$

In KN construction process, the hyperparameter $w$ decides the maximal number of keywords one sentence could hold. Larger $w$ would include more nodes in the KN, but also bring more noise in the process of keyword retrieval. Hyperparameter $m$ indicates the total number of keywords fed into the model for one source sentence. Figure 4.3 shows the results of different $w$ and $m$ of translation direction En→ De, which is considered as a trade-off between relevance and informativeness when choosing different $w$ and $m$.

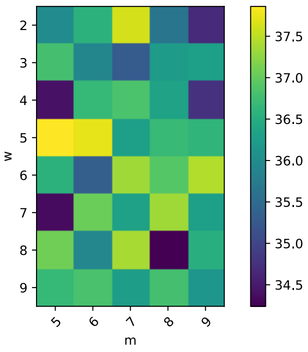


Figure 3: Test 2016

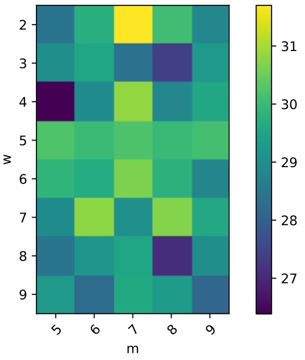


Figure 4: Test 2017

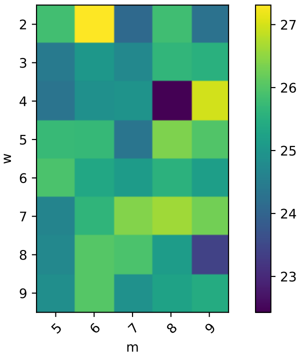


Figure 5: Test 2018

# References

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *THIRD CONFERENCE ON MACHINE TRANSLATION (WMT18)*, volume 2, pp. 308–327, 2018.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. Neural machine translation with monolingual translation memory. *arXiv preprint arXiv:2105.11269*, 2021.

Iacer Calixto, Qun Liu, and Nick Campbell. Incorporating global visual features into attention-based neural machine translation. *arXiv preprint arXiv:1701.06521*, 2017.

Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. Neural machine translation with sentence-level topic context. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):1970–1984, 2019.

Tina Paulsen Christensen and Anne Schjoldager. Translation-memory (tm) research: what do we know and how do we know it? *Hermes-Journal of Language and Communication in Business*, (44):89–101, 2010.

Jean-Benoit Delbrouck and Stéphane Dupont. An empirical study on the effectiveness of images in multimodal neural machine translation. *arXiv preprint arXiv:1707.00995*, 2017.

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

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. Search engine guided neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*, 2019.

Jitao Xu, Josep M Crego, and Jean Senellart. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1580–1590, 2020.

Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4346–4350, 2020.

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*, 2019.