GOOD FOR MISCONCEIVED REASONS: REVISITING NEURAL MULTIMODAL MACHINE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A neural multimodal machine translation (MMT) system is one that aims to perform better translation by extending conventional text-only translation models with multimodal information. Many recent studies report improvements when equipping their models with the multimodal module, despite the controversy whether such improvements indeed come from the multimodal part. We revisit the recent development of neural multimodal machine translation by proposing two *interpretable* MMT models that achieve new state-of-the-art results on the standard Multi30k dataset. To our surprise, however, while we observe similar gains as in the recent developed multimodal-integrated models, our models learn to *ignore* the multimodal information. Upon further investigation, we discover that the improvements bought about by the multimodal models over text-only counterpart are in fact results of the regularization effect. We report our empirical findings which express the importance of MMT models' interpretability and set new paradigms for future MMT research.

1 INTRODUCTION

Since statistical methods and machine learning algorithms were first introduced to the field of natural language processing (NLP) in the early 1990s, most researchers hold the belief that, with appropriate regularization, adding more features to NLP models rarely harms performance. Nevertheless, due to the recent prevalence of *deep* neural network models, applying proper regularization has been far more difficult than tuning a simple λ hyperparameter, which controls the importance of the regularization term in the learning objective[1]. For this reason, while it is often tempting to conclude the gain of a model comes from some exciting features newly introduced, we should also realize that in the context of neural models, this gain may well be due to the fact that the baseline isn't properly regularized. The additional features merely serve as a noise that regularizes the neural model (Goodfellow et al., 2016; Bishop, 1995).

From this perspective, we revisit the recent development of neural multimodal machine translation (MMT). MMT aims at designing better translation systems by extending conventional text-only translation models to take into account multimodal information. Many researches (Calixto et al., 2017; Helcl et al., 2018; Ive et al., 2019; Lin et al., 2020; Yin et al., 2020) report improvements when models are equipped with multimodal information, while others (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018) argue that visual features did not seem to help reliably. In particular, Elliott (2018) conduct thorough experiments to demonstrate that MMT models can translate without significant performance losses even in the presence of features derived from unrelated images. A more recent study (Caglayan et al., 2019), however, shows that under limited textual context (e.g., noun words are masked), models are capable of leveraging the visual input to generate better translations. But it remains unclear where the gains of MMT methods come from when the textual context is complete. Despite those concerns, the field appears to be progressing steadily, albeit slowly in recent years.

In this paper we propose two interpretable models taking advantage of multimodal information. Instead of directly infusing visual features into the model, we design learnable components, allowing the model to voluntarily decide the usefulness of the visual features and reinforce their effects when

[1]Most commonly, quadratic or "l_2" regularization (Krogh & Hertz, 1992) is used: $R(\mathbf{w}) = \lambda \sum_i w_i^2$.

they are helpful. Our models achieve new state-of-the-art results on the standard multi30k (Elliott et al., 2016) dataset. In spite of that, both models learned to ignore the multimodal information to our surprise. Our further analysis suggests that the improvements in fact come from a regularization effect that is similar to the addition of random noise (Bishop, 1995) and weight decay (Hanson & Pratt, 1989). The additional information is treated as noise signal that lower the model's trust on the hidden representations generated from the neural encoders, resulting in a more robust network that has lower generalization error.

Our contributions are threefold. First, we revisit neural models for the popular task of multimodal machine translation. Our findings stress the importance of MMT models' interpretability. Second, for the MMT task, we provide a strong text-only baseline implementation as well as two models with interpretable visual modules that replicate similar gains as reported in previous works. In particular, our MMT models focus on the *interpretability* of how multimodal information is fused into the translation models. Our methods can serve as standard baselines for future interpretable MMT studies. Third, our analysis suggests that the improvements of the multimodal models over text-only counterparts is a result of better regularization effect.

2 BACKGROUND

One can broadly categorize MMT systems into two types: (1) Conventional MMT, where there is gold alignment between the source (target) sentence pair and a relevant image and (2) Retrieval-based MMT, where there is an image corpus from which the system needs to retrieve relevant images as additional clues to assist translation.

Conventional MMT Most MMT systems require MMT datasets of images with bilingual annotations for both training and inference. Many early attempts use a pre-trained model (e.g., ResNet (He et al., 2016)) to encode images into feature vectors. This image/visual representation is then either used in the initialization of the decoder's and/or the encoder's hidden states (Elliott et al., 2015; Libovický & Helcl, 2017; Calixto et al., 2016) or appended/prepended to word embeddings as additional input tokens (Huang et al., 2016; Calixto & Liu, 2017). Recent works (Calixto et al., 2017; Libovický et al., 2018; Zhou et al., 2018; Ive et al., 2019; Lin et al., 2020) use attention mechanism to select relevant visual features and generate a visual-aware representation for the decoder. While there are more works on engineering decoders, encoder-based approaches are relatively less explored. To this end, Yao & Wan (2020) and Yin et al. (2020) propose to replace the vanilla Transformer encoder with a stack of multi-modal encoder layers.

Retrieval-based MMT The effectiveness of conventional MMT relies on the availability of images with bilingual annotations, which limits its applicability. To address this issue, Zhang et al. (2020) integrate a retrieval component into MMT. They use term-frequency-inverse-document-frequency (TF-IDF) to build a token-to-image lookup table. With this table, images that share similar topics with a source sentence are retrieved as relevant images. This creates image-bilingual-annotation instances for training. Retrieval-based models have been shown to improve performance across a variety of NLP tasks besides MMT, such as open-domain question answering (Guu et al., 2020; d'Autume et al., 2019), fact checking (Thorne et al., 2018), dialogue (Weston et al., 2018), language modeling (Khandelwal et al., 2019), question generation (Lewis et al., 2020), and translation (Gu et al., 2018).

3 METHOD

In this section we introduce two new MMT models: (1) *Gated Fusion* for conventional MMT and (2) *Retrieval-augmented MMT* (RMMT) for retrieval-based MMT. Our design philosophy is that the models should learn, in an interpretable manner, to which degree multimodal information should be used.

We start with the simple yet effective *Gated Fusion* model that translates a source sentence x into a target y with the help of a gold-standard image z associated to x and y. The model learns a gating vector $\vec{\lambda} \in [0, 1]^d$ that controls the amount of visual information that is blended into the textual

representation. The learned $\vec{\lambda}$ makes the fusion process interpretable: a larger gating value $\lambda \in \vec{\lambda}$ indicates that the model exploits more visual context in the translation.

For RMMT, we augment a conventional MMT model with a learnable retriever, which learns to retrieve images that are semantically-relevant to an input sentence x. We jointly train the retriever with the rest of the model in an end-to-end manner such that the model *learns to retrieve* images that are highly influential in the translation.

3.1 GATED FUSION MMT

Given a source sentence x of length T, we first feed it into a vanilla Transformer (Vaswani et al., 2017) encoder to obtain a textual representation $\mathbf{H}_{\text{text}} \in \mathbb{R}^{T \times d}$. The encoder contains L identical layers, each with a multi-head self-attention sublayer followed by a position-wise, fully connected feed-forward sublayer. Layers and sublayers are chained using residual connection (He et al., 2016) and layer normalization (Ba et al., 2016). For the image z associated with x, we use a ResNet-50 CNN (He et al., 2016) trained on ImageNet (Russakovsky et al., 2015) to extract a 2048-dimensional average pooled visual representation, which is then projected and broadcasted to the same dimension as $\mathbf{H}_{\text{text}}$.

$$\text{Embed}_{\text{image}}(z) = \mathbf{W}_{\text{z}}\ \text{ResNet}_{\text{pool}}(z)\,. \tag{1}$$

We now devise a gating mechanism to control the fusion of $\mathbf{H}_{\text{text}}$ and $\text{Embed}_{\text{image}}(z)$. The gating vector $\vec{\lambda} \in [0,1]^d$ is computed as follow:

$$\vec{\lambda} = \text{sigmoid}\left(\mathbf{W}_{\lambda}\ \text{Embed}_{\text{image}}(z) + \mathbf{U}_{\lambda}\mathbf{H}_{\text{text}}\right), \tag{2}$$

where $\mathbf{W}_{\lambda}$ and $\mathbf{U}_{\lambda}$ are model parameters. Note that this gating mechanism has been a building block for many recent MMT systems (Zhang et al., 2020; Lin et al., 2020; Yin et al., 2020). We are, however, the first to focus on its interpretable factors. Finally, we fuse the textual and visual representations:

$$\mathbf{H} = \mathbf{H}_{\text{text}} + \vec{\lambda}\,\text{Embed}_{\text{image}}(z). \tag{3}$$

The output $\mathbf{H}$ is fed into the decoder directly for translation as in vanilla Transformer.

3.2 RETRIEVAL-AUGMENTED MMT (RMMT)

On the basis of Gated Fusion, we introduce a retrieval component that retrieves images from a pool of candidates $\mathcal{Z}$ using input sentence x. Each retrieved image z is then treated as a latent variable and we marginalize over all $z \in \mathcal{Z}$ to obtain the overall likelihood of generating the target sentence y, yielding:

$$p(y|x) = \sum_{z \in \mathcal{Z}} p_{\eta}(z \mid x) \prod_{i}^{N} p_{\theta}\left(y_i \mid x, z, y_{<i}\right). \tag{4}$$

RMMT consists of two sequential components: (1) an image retriever $p_{\eta}(z|x)$ with parameters η that returns probability distributions over candidate images given a query x. (2) a multi-modal translator $\prod_i^N p_{\theta}(y_i|x, z, y_{<i})$ that generates each y_i condition on the input sentence x, the image z returned by the retriever, and the previous generated tokens $y_{<i}$. Ideally, we can implement this multimodal translator (denote as $p(y|x, z)$) using any existing MMT models. Without loss of generality, here we use the Gated Fusion model as an example.

Image Retriever The retriever is defined as a dense inner product model:

$$p(z \mid x) = \frac{\exp f(x, z)}{\sum_{z'} \exp f\left(x, z'\right)}, \tag{5}$$

$$f(x, z) = \text{Embed}_{\text{text}}(x)^{\top}\ \text{Embed}_{\text{image}}(z), \tag{6}$$

where $\text{Embed}_{\text{text}}(x)$ and $\text{Embed}_{\text{image}}(z)$ are functions that map x and z respectively to $d-$dimensional vectors. The relevance score $f(x, z)$ between x and z is defined as the inner product of the vector embeddings. The retrieval distribution is the softmax over all relevance scores.

We obtain $\text{Embed}_{\text{image}}(z)$ as in Equation 1. For $\text{Embed}_{\text{text}}(x)$, we implement it using BERT (Devlin et al., 2018) as below:

$$\text{Embed}_{\text{text}}(x) = \mathbf{W}_{\text{text}}\ \text{BERT}_{\text{CLS}}(x)\,. \tag{7}$$

Following standard practices, we use a pre-trained BERT model[2] to obtain the "pooled" representation of the sequence (denoted as $\text{BERT}_{\text{CLS}}(x)$). $\mathbf{W}_{\text{text}}$ is a projection matrix.

Decoding The likelihood $p(y|x)$ estimated in RMMT (Equation 4) cannot be solved with a single beam search pass. Instead, we use the "Fast Decoding" algorithm proposed in Lewis et al. (2020) to approximate $\max_y p(y \mid x)$. In particular, we run beam search for each candidate image $z_j \in \{z_1, \ldots, z_K\}$, scoring each hypothesis using $p_\theta(y_i \mid x, z_j, y_{<i})$. This yields K set of hypotheses of which some might not have appeared in the search space of all images. We make a further approximation that $p_\theta(y \mid x, z_j) = 0$, where y was not generated during beam search from x, z_j. Finally, we multiply each score $p_\theta(y_i \mid x, z_j, y_{<i})$ with $p(z_j \mid x)$ and then sum up the probabilities across K beams to obtain the marginalized probability $p_\theta(y_i \mid x, y_{<i})$.

3.3 TRAINING

We train both Gated Fusion and RMMT by maximizing the log-likelihood, $\log p(y|x)$. For Gated Fusion, we follow standard sequence to sequence (Bahdanau et al., 2014) training protocols. For RMMT, since both the image retriever and multimodal translator are differentiable neural networks, we can jointly optimize them (as defined in Equation 4) using stochastic gradient descent. The key computational challenge is that the marginal probability $p(y|x)$ involves a summation over all images z in the image corpus $\mathcal{Z}$. We approximate this by summing over the top k images with the highest probability under $p_\eta(z \mid x)$, assuming that most images are irrelevant to x and thus giving negligible probabilities. The whole retrieval process is in-memory.

What does the retriever learn? Since there is no direct supervision for the retriever in training (i.e. we do not provide ground truth images to supervise the training of retriever), it is not obvious how the training objective encourages meaningful retrievals. Here, we provide mathematical guarantees about how RMMT rewards retrievals that improve translation performance. Recall that $f(x, z)$ (see Equation 6) is the relevance score that retriever assigns to images z, given input sequence x. We can see how a single step of gradient descent during training alters $f(x, z)$ by analyzing the gradient with respect to the parameters of the image retriever, η:

$$\begin{aligned} \nabla \log p(y \mid x) &= \sum_{z \in \mathcal{Z}} r(z) \nabla f(x, z), \\ r(z) &= \left[\frac{p(y \mid z, x)}{p(y \mid x)} - 1 \right] p(z \mid x). \end{aligned} \tag{8}$$

For each image z, the change of retriever's relevance score $f(x, z)$ is associated with $r(z)$ — increasing (decreasing) if $r(z)$ is positive (negative). The multiplier $r(z)$ is positive if and only if $p(y|x, z) > p(y|x)$. The term $p(y|x, z)$ is the probability of generating the correct sequence y conditioning on image z, while $p(y|x)$ is the expected value of $p(y|x, z)$ when conditioning on randomly sampled image from $p(z|x)$. Therefore, whenever the model performs better than expected, a ***helpful*** image z will receive a positive update (i.e $f(x, z) = f(x, z) + \text{learning rate} * \nabla f(x, z)$, since $\nabla f(x, z)$ is positive, the relevance score will increase). Through this training procedure, the retriever learns to focus on images that can improve translation performance. We refer readers to Guu et al. (2020) for a detailed derivation of Equation 8.

Retriever warm-start At the beginning of training, the retrieved images z will likely be unrelated to x since the retriever does not yet have good representations for $\text{Embed}_{\text{text}}(x)$ and $\text{Embed}_{\text{image}}(z)$. This starts a vicious cycle where the multimodal translator learns to ignore the retrieved *irrelevant* images and thus the retriever does not receive a meaningful gradient and cannot improve. To avoid this issue, we pre-train the retriever on a Text-to-Image-Retrieval task (Frome et al., 2013). The goal of the pre-training is to learn a common space for image and text such that we can measure the visual-semantic similarity by the inner product of two vectors.

[2]Here we use bert-base-uncased version from HuggingFace (Wolf et al., 2019).

4 EXPERIMENT

4.1 DATASET

We evaluate our methods on the widely-used MMT datasets: Multi30k (Elliott et al., 2016).

The Multi30k dataset is an extension of Flickr30k (Plummer et al., 2015). Flickr30k contains 31,014 images sourced from online photo-sharing websites. Each image is paired with five English descriptions. Researchers create Multi30k by sampling one image-caption pair for each image from Flickr30k and crowd-sourcing German/French translation for the English caption. We follow a standard split of 29,000 instances for training, 1,014 for validation, and 1,000 for testing (Test2016). We also report results on the 2017 test set (Test2017) with extra 1,000 instances and the MSCOCO test set that has 461 more challenging out-of-domain instances that contain ambiguous verbs.

We use the official preprocessed version of Multi30k[3] and merge the source and target sentences to build a joint vocabulary. We then apply the byte pair encoding (BPE) algorithm (Sennrich et al., 2016) with 10,000 merging operations to segment words into subwords, following standard practice. Resulting in a vocabulary of 9,712 (9,544) tokens for En-De (En-Fr).

Retriever pre-train. We pre-train the retriever on a subset of Flickr30k that has overlapping instances with Multi30k removed. We use Multi30k's validation set to evaluate the retriever. We measure the performance by recall-at-K ($R@K$) defined as the fraction of queries for which the correct images are retrieved in the closest K images to the query. The pre-trained retriever registers an $R@1$ of 22.8% and $R@5$ of 39.6%.

4.2 SETUP

We implement the proposed models and baselines with FairSeq toolkit (Ott et al., 2019). We experiment with different model sizes (*Base*, *Small*, and *Tiny*) as detailed in Appendix A. *Base* is a widely-used model configuration for Transformer in both text-only translation (Vaswani et al., 2017) and MMT (Grönroos et al., 2018; Ive et al., 2019). However, for small datasets like Multi30k, training such a big model (about 50 million parameters) has the potential of overfitting. We therefore turn to a relatively small model commonly used for low-resourced translation (Zhu et al., 2020), denoted as *Small*. In our preliminary study, we found that even a *Small* model can still overfit. We thus perform a grid search on Multi30k'En-De validation set and obtain a *Tiny* configuration that works surprisingly well. A *Tiny* model has 4 encoder layers and 4 decoder layers. The dimensions of input layer, output layer, and inner feed-forward layer are set to 128, 128, and 256, respectively. The number of attention heads is set to 4.

We use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.98$ for model optimization. We start training with a warm-up phase (2,000 steps) where we linearly increase the learning rate from 10^{-7} to 0.005. Thereafter we decay the learning rate proportional to the number of updates. Each training batch contains at most 4,096 source/target tokens. We set label smoothing weight to 0.1, dropout to 0.3. We follow Zhang et al. (2020) to early-stop the training if validation loss does not improve for ten epochs. We average the last ten checkpoints for inference as in Vaswani et al. (2017) and Wu et al. (2019). We perform beam search with beam size set to 5. Multi-bleu.perl was used to compute 4-gram BLEU scores for all test sets. All models are trained and evaluated on one single machine with two Titan P100 GPUs.

4.3 BASELINES

Except the text-only **Transformer** (Vaswani et al., 2017) baseline, we also compare to the following conventional MMT baselines:

Doubly-ATT (Calixto et al., 2017; Helcl et al., 2018; Arslan et al., 2018) A doubly-attentive transformer is an extension of text-only Transformer, whose vanilla decoder is replaced with a multimodal decoder. An extra visual attention sublayer is inserted between the decoder's source-target attention sublayer and feed-forward sublayer.

Imagination The Imagination architecture, initially proposed by Elliott & Kádár (2017) and later integrated into Transformer by Helcl et al. (2018), attempts to leverage the benefits of multi-tasking

[3]https://github.com/multi30k/dataset

#	Model	En→De				En→Fr			
		#Params	Test2016	Test2017	MSCOCO	#Params	Test2016	Test2017	MSCOCO
	Text-only Transformer								
1	**Transformer**-Base	49.1M	38.33	31.36	27.54	49.0M	60.60	53.16	42.83
2	**Transformer**-Small	36.5M	39.68	32.99	28.50	36.4M	61.31	53.85	44.03
3	**Transformer**-Tiny	2.6M	41.02	33.36	29.88	2.6M	61.80	53.46	44.52
	Existing MMT Systems								
4	**GMNMT**♠	4.0M	39.8	32.2	28.7	-	60.9	53.9	-
5	**DCCN**♠	17.1M	39.7	31.0	26.7	16.9M	61.2	54.3	**45.4**
6	**Doubly-ATT**♠	3.2M	41.45	**33.95**	29.63	3.2M	61.99	53.72	45.16
7	**Imagination**♠	7.0M	41.31	32.89	29.90	6.9M	61.90	54.07	44.81
8	**UVR-NMT**♢	2.9M	40.79	32.16	29.02	2.9M	61.00	53.20	43.71
	Our MMT Systems								
9	**Balanced Fusion**♠	2.8M	39.48	30.39	27.51	2.8M	57.36	49.86	42.03
10	**Gated Fusion**♠	2.9M	41.96	33.59	29.04	2.8M	61.45	54.03	43.77
11	**RMMT-Static**♢	2.9M	**42.12**$^{\ddagger *}$	33.14	**30.53**$^{\dagger *}$	2.9M	**62.12**	**55.08**$^{\ddagger **}$	45.10†
12	**RMMT**♢	3.3M	42.00	33.82	30.56	3.2M	61.72	54.0	45.24

Table 1: BLEU scores on Multi30k. Results in row 4 and 5 are taken from the original papers. ♠ indicates conventional MMT models, while ♢ refer to retrieval-based models. Without further specified, all our implementations are based on the *Tiny* configuration. ‡/†: significantly better than Transformer-Tiny ($p < 0.01/0.05$), **/*: significantly better than Imagination ($p < 0.01/0.05$)

to improve MMT. They propose to share the sentence encoder between the translation task and an auxiliary visual reconstruction task, which encourages the visual grounding for translation.

Balanced Fusion In comparison to Gated Fusion, we experimented on a non-parametric fusion method that directly adds up visual and textual representations.

For retrieval-based MMT, we consider the following baselines:

UVR-NMT (Zhang et al., 2020). This method enables MMT to be applied to large-scale text-only NMT through a token-to-image lookup. They use an attention layer with a gated weighting to fuse the visual representations and the textual representations.

RMMT-static To evaluate the effectiveness of our methods against UVR-NMT, we propose a variant called RMMT-static. In particular, we do not update our retriever during training and degenerate the learning objective to the one used in UVR-NMT. We also use the same fusion component as in UVR-NMT. For both RMMT and RMMT-static, we use top-5 retrieved images following UVR-NMT.

To ensure fair comparison and minimize training/environmental differences, all the above baselines are implemented by ourselves based on FairSeq and with the same set of hyper-parameters. We also consider two more recent methods: **GMNMT** (Yin et al., 2020) and **DCCN** (Lin et al., 2020), we directly use the results reported in their paper.

5 RESULTS

Table 1 shows the results on Multi30k. Our reproduction of text-only Transformer-*Base* (row 1) matches the BLEU scores reported in Grönroos et al. (2018); Ive et al. (2019). We further test the models with fewer parameters, Transformer-*Small* and Transformer-*Tiny*. We observe clear improvements with less parameters. This supports our speculation that the large number of parameters at the scale of Transformer-*Base* leads to overfitting. Transformer-*Tiny*, whose parameters are about $1/20$ in number of Transformer-*Base*, is more robust and efficient in all our test sets. We therefore use it as the default basis for all our MMT systems in the following discussion.

5.1 CONVENTIONAL MMT SYSTEMS

Previous conventional MMT systems have demonstrated clear improvements over text-only baselines. For instance, Helcl et al. (2018) shows a 0.9 BLEU gain on Imagination model. While with an ever-strong text-only new baseline, we replicate similar, albeit smaller gains (see rows 6 and 7). Despite its simplicity, our Gated Fusion model (row 10) surpasses almost all baselines on En→De translation, and is on-par with the best on En→Fr translation. Comparing to Balanced Fusion (row

9), we can see that allowing the model to dynamically control the fusion of multimodal representation is critical to high performance. A natural question then arises: "How much visual context is exploited by Gated Fusion?"

To answer this question we focus on the interpretable component in Gated Fusion — the gating vector $\vec{\lambda} \in [0, 1]^d$ (see Equation 3). Intuitively, a larger gating value indicates that the model learns to depend more on visual context to perform better translation.

We propose to quantify the degree to which the visual context is exploited in Gated Fusion using *visual awareness* (denoted as $\Delta\vec{\lambda}$), which is defined as the percentage of non-zero values in $\vec{\lambda}$.

From Table 2, we observe that $\Delta\vec{\lambda}$ is generally small, suggesting that the model learns to discard visual context most of the time. Worse yet, non-zero gating values are within $[10^{-10}, 10^{-40}]$. Such negligibly small values indicate that even though the visual context is exploited by the model, it is orders-of-magnitude less important than previously thought.

5.2 Retrieval-based MMT

Our best-performing model is RMMT-static (Table 1, row 11), whose retriever is frozen and does not update during training. Although RMMT-static does not always use the gold-standard images associated with the sentence pairs, it still significantly outperforms conventional MMT models like Imagination. One plausible explanation is that conventional systems only rely on a single image for each sentence pair, while our model relies on top-k retrieved images, which injects more diversity into the translation.

	Visual awareness
Multi30k En→De	
Test2016	25.0%
Test2017	30.0%
MSCOCO	25.1%
Multi30k En→Fr	
Test2016	32.8%
Test2017	38.0%
MSCOCO	33.2%

Table 2: Visual awareness on Multi30k.

To our surprise, the RMMT model, in which the retriever is jointly optimized with the model, does not bring further improvement. Inspecting closer, we observe that RMMT's retriever has "collapsed" and learned to retrieve the same images regardless of the input. The retriever collapse issue (Lewis et al., 2020) is not an uncommon phenomenon in retrieval-augmented models. It usually happens when there is a *less-explicit requirement* for the retrieved items in the task, resulting in less informative gradients for the retriever.

We also apply RMMT and RMMT-static to text-only corpora, which is described in details in Appendix E.

5.3 Revisit the need for visual context in MMT

The phenomena of "zero-valued gating vector" in Gated Fusion and "retriever collapse" in RMMT suggest that visual context does not offer additional useful information in MMT, or at least the offering is much less than what was previously thought. Given that both previous papers and our experiments see similar gains of MMT systems when they are compared with text-only baselines, we want to investigate where these gains come from. Our hypothesis is that the improvements are due to some regularization effects, where the additional information is treated as random noise that lowers the model's trust on the hidden representations generated from the neural encoders, resulting in smaller network weights and a more robust network that has lower generalization error.

To verify our hypothesis, we conduct two sets of experiments based on two popular regularization methods: adding random noise (Bishop, 1995) and adding weight decay (Hanson & Pratt, 1989). Supported by these two experiments, we find that these regularization techniques achieve similar gains over the text-only baseline in incorporating the multimodal information.

5.3.1 Random Noise

In the random noise experiments we keep all hyper-parameters unchanged but replace visual representations extracted by ResNet (denoted as *ResNet features*) with randomly initialized vectors (denoted as *noise*). In other words, the noise was added once per training sequence as in Graves

#	Model	En→De			En→Fr		
		Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO
1	**Transformer**	41.02	33.36	29.88	61.80	53.46	44.52
2	**Doubly-ATT**	41.53(+0.08)	33.90(-0.05)	29.76(+0.15)	61.85(-0.35)	54.61(+0.46)	44.85(-0.80)
3	**Imagination**	41.20(-0.11)	33.32(+0.42)	29.92(+0.02)	61.28(-0.62)	53.74(-0.33)	44.79(-0.01)
4	**Gated Fusion**	41.53(-0.45)	33.52(-0.07)	29.87(+0.83)	61.58(+0.13)	54.21(+0.18)	44.86(+1.09)

Table 3: BLEU scores on Multi30k with randomly initialized visual representation. Numbers in parentheses indicate the relative improvement/deterioration compared with the original model with ResNet features.

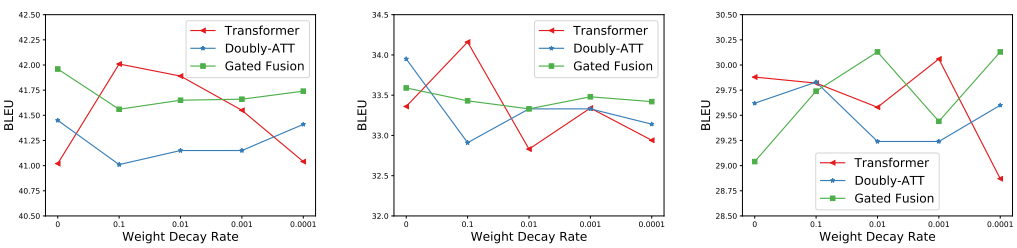
et al. (2013) rather than at every timestep. We ran each experiment three times with a different set of noise drawn from a Gaussian distribution whose mean and standard deviation are both set to 1. We report the average performance results, which are shown in Table 3. We observe that on En→De translation, using random noise (rows 2-4) can also achieve clear gains over the text-only baseline (row 1). On Gated Fusion (En→Fr), using random noise can even outperform the same model equipped with ResNet features.

Our adversarial evaluation shows that the added random noise is functioning just like the visual context. Injecting noise into a neural network makes it difficult for the network to fit individual data points precisely, and hence it reduces overfitting (Bishop et al., 1995). In terms of MMT, added noise or visual context make the translations of sentences in Multi30k, which are short and repetitive (Caglayan et al., 2019), more challenging.

5.3.2 Weight Decay

As for another controlled evaluation, we focus on weight decay. Weight decay has not been widely adopted in NMT because NMT datasets generally contain millions of sentences and are less likely to overfit. Even in low-resource translation (Wu et al., 2019), researchers are very cautious with the usage of weight decay (e.g., 0.0001). We experimented on text-only Transformer, existing MMT method (Doubly-ATT), and our Gated Fusion method.

We can see from Figures 1(a) and (b) that with a large penalty on model weights (0.1), a text-only Transformer (**red** line with triangle) can also exhibit performance that are comparable to or even superior than Dounly-ATT (**green** line with ■) and Gated Fusion (**blue** line with ★). This finding further supports our hypothesis that multimodal context serves as a kind of regularization. On the other hand, introducing more regularization into MMT models always deteriorates the performance on Test2016 and Test2017. One plausible explanation is that the visual context already serves as regularization in MMT, thus imposing more regularization will lead to under-fitting. On MSCOCO, we observe a similar trend for Transformer, but both MMT models show further improvement equipped with weight decay. This unusual behavior could be due to the fact that MSCOCO contains out-of-domain sentences with ambiguous verbs, thus injecting more regularization to force the network to trust the textual encoder's output less can further improve the model's performance.



(a) Test2016. (b) Test2017. (c) MSCOCO.

Figure 1: BLEU score curves on En→De translation with different weight decay rate. 0 indicates without weight decay.

6 CONCLUSION

In this paper we propose two novel interpretable models that exhibit new state-of-the-art performance on the widely adopted MMT datasets — Multi30k. Our analysis on the proposed models, as well as on other existing MMT systems, suggests that visual context helps MMT in the similar vein as regularization methods (e.g., weight decay). Those empirical findings, however, should not be understood as us denying the contribution of existing MMT methods. We believe that sophisticated MMT models are necessary for effective grounding of visual context into translation. Their effectiveness might be impaired by limitation in existing datasets (Caglayan et al., 2019). Our results emphasise the importance of interpretability in MMT research, and stress the need of new datasets to push forward the field.

REFERENCES

Hasan Sait Arslan, Mark Fishel, and Gholamreza Anbarjafari. Doubly attentive transformer machine translation. *arXiv preprint arXiv:1807.11605*, 2018.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *WMT18*, 2018.

Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

Christopher M Bishop et al. *Neural networks for pattern recognition*. Oxford university press, 1995.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. In *NAACL-HLT (1)*, 2019.

Iacer Calixto and Qun Liu. Sentence-level multilingual multi-modal embedding for natural language processing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pp. 139–148, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_020. URL https://doi.org/10.26615/978-954-452-049-6_020.

Iacer Calixto, Desmond Elliott, and Stella Frank. Dcu-uva multimodal mt system report. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 634–638, 2016.

Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1913–1924, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1175. URL https://www.aclweb.org/anthology/P17-1175.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. Episodic memory in lifelong language learning. *arXiv preprint arXiv:1906.01076*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Desmond Elliott. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2974–2978, 2018.

Desmond Elliott and Ákos Kádár. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 130–141, 2017.

Desmond Elliott, Stella Frank, and Eva Hasler. Multilingual image description with neural sequence models. *arXiv preprint arXiv:1510.04709*, 2015.

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. *arXiv preprint arXiv:1710.07177*, 2017.

Stella Frank, Desmond Elliott, and Lucia Specia. Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. *Natural Language Engineering*, 24(3):393–413, 2018.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129, 2013.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649. IEEE, 2013.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, et al. The memad submission to the wmt18 multimodal translation task. *arXiv preprint arXiv:1808.10802*, 2018.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. Search engine guided neural machine translation. In *AAAI*, pp. 5133–5140, 2018.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*, 2020.

Stephen José Hanson and Lorien Y Pratt. Comparing biases for minimal network construction with back-propagation. In *Advances in neural information processing systems*, pp. 177–185, 1989.

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.

Jindřich Helcl, Jindřich Libovickỳ, and Dušan Variš. Cuni system for the wmt18 multimodal translation task. *arXiv preprint arXiv:1811.04697*, 2018.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 639–645, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2360. URL `https://www.aclweb.org/anthology/W16-2360`.

Julia Ive, Pranava Madhyastha, and Lucia Specia. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6525–6538, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1653. URL `https://www.aclweb.org/anthology/P19-1653`.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2019.

Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pp. 950–957, 1992.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020.

Jindřich Libovický and Jindřich Helcl. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 196–202, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2031. URL `https://www.aclweb.org/anthology/P17-2031`.

Jindřich Libovický, Jindřich Helcl, and David Mareček. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 253–260, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6326. URL https://www.aclweb.org/anthology/W18-6326.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. Dynamic context-guided capsule network for multimodal machine translation. *arXiv preprint arXiv:2009.02016*, 2020.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

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. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

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.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 543–553, 2016.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. 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. 809–819, 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.

Jason Weston, Emily Dinan, and Alexander Miller. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pp. 87–92, 2018.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1901.10430.

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.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3025–3035, 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*, 2020. URL https://openreview.net/forum?id=Byl8hhNYPS.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3643–3653, 2018.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*, 2020.

A MODEL SIZE

Table 4 shows the configuration of different model sizes.

Model component	Base	Small	Tiny
Number of encoder/decoder layers	6	6	4
Input/Output layer dimension	512	512	128
Inner feed-forward layer dimension	2048	1024	256
Number of attention heads	8	4	4

Table 4: Model configurations for *Base*, *Small*, and *Tiny*.

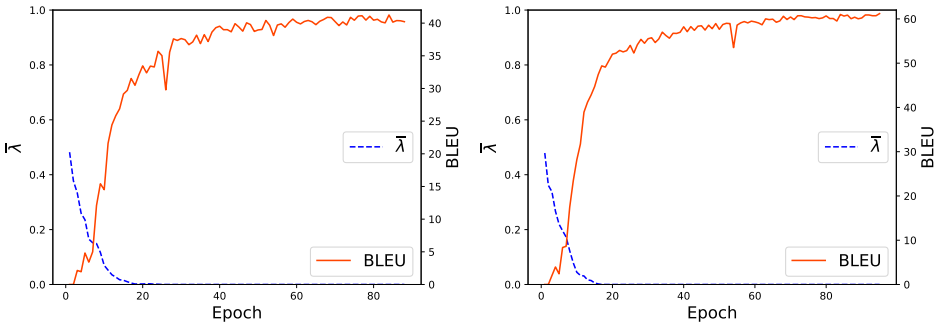
B RESULTS ON VATEX

To further support our findings, we replicate our experiments on a new large-scale dataset – VaTex. VaTex is a video-based MMT corpus that contains 129,955 English-Chinese sentence pairs for training, 15,000 sentence pairs for validation, and 30,000 sentence pairs for testing. Each pair of sentences is associated with a video clip. Since the testing set is not publicly available, we use half of the validation set for validating and the other half for testing. We apply the byte pair encoding algorithm on the lower-cased English sentences and split Chinese sentences into sequences of characters, resulting in a vocabulary of 17,216 English tokens and 3,384 Chinese tokens. We use the video features provided along with the VaTex dataset, in which each video is represented as $\mathbb{R}^{k*1024}$, where k is the number of segments. Since some MMT systems take a "global" visual feature as input, we use 3D-Max-Pooling to extract the pooled representation $\mathbb{R}^{1024}$ for each video.

Model	**BLEU**	**METEOR**
Transformer	35.82	59.02
+weight decay 0.1	36.32	59.38
+weight decay 0.01	36.07	59.14
+weight decay 0.001	35.92	59.22
Doubly-ATT	36.05	59.26
Imagination	36.25	59.26
Gated Fusion	36.06	59.34
RMMT-Static	36.35	59.44
RMMT	35.34	58.32

Table 5: Results on VaTex En-Zh translation.

The results are shown in Table 5. We observe that although most MMT systems show improvement over the Transformer baseline, the gains are quite marginal. Indicating that although image-based MMT models can be directly applied to video-based MMT, there is still room for improvement due to the challenge of video understanding. We also note that regularize the text-only Transformer with weight decay demonstrates similar gains as injecting video information into the models, which further supports our findings in Section 5.3.2.



(a) En-De. (b) En-Fr.

Figure 2: Training dynamic analysis on Test2016 of Multi30k En-de and En-Fr translation.

C Training Dynamic Analysis

All our investigations so far are conducted in a post-hoc manner (analyze after model converge). In this section, we dive deeply into our models' training dynamics to understand how they accommodate to the introduction of visual information during training. Recall that Gated Fusion contains a gating vector $\vec{\lambda} \in [0, 1]^d$ that controls the amount of visual information that is blended into the textual representation. We compute a corpus-level averaged gating weight $\overline{\lambda} \in [0, 1]$ to measure the degree visual information is exploited in Gated Fusion. Figure 2 shows how $\overline{\lambda}$ changes during training, starting from epoch 1.

Note that at the beginning of training, Gated Fusion uses a relatively high $\overline{\lambda}$ (> 0.5), but quickly decrease to ≈ 0.48 at the end of Epoch 1. We speculate that the model relies heavily on images at the beginning because visual features are extracted from pre-trained ResNet-50 and are more informative, whereas the textual encoder is randomly initialized. As the training continues, $\overline{\lambda}$ gradually decrease to near zero ($<$10e-10). This learning curve shows that Gated Fusion learns to translate without "seeing" images. Because with such a magnitude smaller weight, most (if not all) visual knowledge is discarded during fusion. However, although negligibly small, $\overline{\lambda}$ is never precisely zero (i.e. the model will not discard visual information completely). This motivates our conjecture that the visual information is functioning like some structured noise that slightly alters the model's hidden representation. We observe a similar trend in RMMT that the retriever's recall quickly decrease to near zero over time.

D Results on METEOR

We also report our results based on METEOR (Banerjee & Lavie, 2005), which consistently demonstrates higher correlation with human judgments than BLEU does in independent evaluations such as EMNLP WMT 2011 [4].

From Table 6, we can see that on En-Fr translation, MMT systems demonstrate similar improvements over text-only baselines in both METEOR and BLEU(see Table 1). On En-De translation, however, MMT systems are mostly on-par with Transformer-tiny on METEOR and do not show consistent gains as BLEU. We hypothesis the reason being that En-De train/valid/test2016 set is created in a *image-blind* fashion, in which the crowd-sourcing workers produce translations without seeing the images (Frank et al., 2018). Such that source sentence can already provide sufficient context for translation. When creating the En-Fr corpus, the image-blind issue is fixed (Elliott et al., 2017), thus images are perceived as "needed" in the translation for whatever reason. Although BLEU is unable to elicit this difference, evaluation based on METEOR captured it and confirmed previous research.

We also repeat our experiments about random noise (Section 5.3.1) and weight decay (Section 5.3.2) using METEOR.

[4] http://statmt.org/wmt11/papers.html

#	Model	En→De				En→Fr			
		#Params	Test2016	Test2017	MSCOCO	#Params	Test2016	Test2017	MSCOCO
	Text-only Transformer								
1	**Transformer**-Base	49.1M	65.92	60.02	54.73	49.0M	80.09	74.93	68.57
2	**Transformer**-Small	36.5M	66.01	60.80	55.95	36.4M	80.71	75.74	69.10
3	**Transformer**-Tiny	2.6M	68.22	62.05	56.64	2.6M	81.02	75.62	69.43
	Existing MMT Systems								
4	**GMNMT**♠	4.0M	57.6	51.9	47.6	-	74.9	69.3	-
5	**DCCN**♠	17.1M	56.8	49.9	45.7	16.9M	76.4	70.3	65.0
6	**Doubly-ATT**♠	3.2M	68.04	61.83	56.21	3.2M	81.12	75.71	70.25
7	**Imagination**♠	7.0M	68.06	61.29	56.57	6.9M	81.2	76.03	70.35
	Our MMT Systems								
9	**Balanced Fusion**♠	2.8M	66.64	58.62	54.12	2.8M	78.16	72.90	67.71
10	**Gated Fusion**♠	2.9M	67.84	61.94	56.15	2.8M	80.97	76.34	70.51
11	**RMMT-Static**♢	2.9M	68.64	61.95	56.75	2.9M	81.44	76.88	70.34
12	**RMMT**♢	3.3M	68.75	62.07	56.35	3.2M	81.05	75.75	71.01

Table 6: METEOR scores on Multi30k. Results in row 4 and 5 are taken from the original papers. ♠ indicates conventional MMT models, while ♢ refers to retrieval-based models. Without further specification, all our implementations are based on the *Tiny* configuration.

#	Model	En→De			En→Fr		
		Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO
1	**Transformer**	68.22	62.05	56.64	81.02	75.62	69.43
2	**Doubly-ATT**	68.39(+0.35)	61.83(+0.0)	56.46(+0.25)	81.27(+0.15)	76.22(+0.51)	70.21(-0.04)
3	**Imagination**	67.93(-0.13)	61.84(+0.55)	56.49(-0.08)	80.75(-0.45)	76.57(+0.54)	69.88(-0.47)
4	**Gated Fusion**	68.25(+0.41)	61.5(-0.44)	55.93(-0.22)	81.22(+0.25)	76.01(-0.33)	70.33(-0.18)

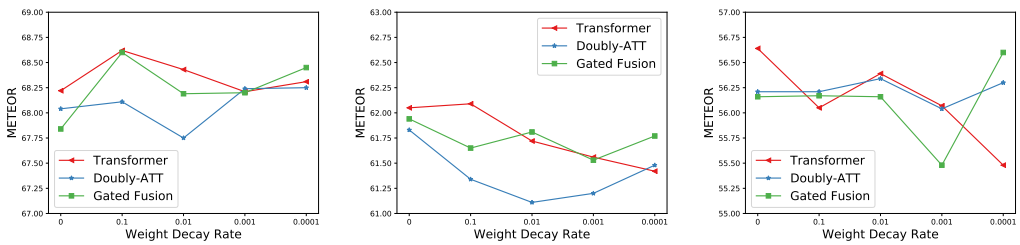
Table 7: METEOR scores on Multi30k with randomly initialized visual representation. Numbers in parentheses indicate the relative improvement/deterioration compared with the original model with ResNet features.

D.1 RADNDOM NOISE

From Table 7, we observe that Doubly-ATT equipped with random noise (row 2) is consistently on-par with (En→De) or better than (En→Fr) text-only baseline (row 1). On Doubly-ATT (En→De), using random noise can even outperform the same model equipped with ResNet features.

D.2 WEIGHT DECAY

We can see from Figures 3 (a) and (b) that with a large weight decay penalty (0.1), a text-only Transformer can also exhibit METEOR scores that are comparable to or even higher than MMT systems. These results are mostly consistent with those evaluated using BLEU (see Figure 1). And again, we observe very different model behaviors on the MSCOCO set, which we have discussed in Section 5.3.2.



(a) Test2016. (b) Test2017. (c) MSCOCO.

Figure 3: METEOR score curves on En→De translation with different weight decay rate. 0 indicates without weight decay.

Model	BLEU
Transformer-*Small*	28.62
+weight decay 0.0001	29.14
RMMT-Static-*Small*	29.08
RMMT-*Small*	28.57

Table 8: BLEU score on IWSLT'14 EN→DE translation.

All adversarial evaluation above complement our early findings illustrated using BLEU.

E RESULTS ON IWSLT'14

We also evaluate the retrieval-based model RMMT on text-only corpus — IWSLT'14. The IWSLT'14 dataset contains 160k bilingual sentence pairs for En-De translation task. Following the common practice, we lowercase all words, split 7k sentence pairs from the training dataset for validation and concatenate *dev2010*, *dev2012*, *tst2010*, *tst2011*, *tst2012* as the test set. The number of BPE operations is set to 20,000. We use the *Small* configuration in all our experiments. The dropout and label smoothing rate are set to 0.3 and 0.1, respectively. Since there is no images associated with IWSLT, we follow Zhang et al. (2020) and retrieve images from Multi30K corpus.

From Table 8, we see that Transformer without weight decay is marginally outperformed by RMMT-static, but achieves slightly higher BLEU scores when trained with a 0.0001 weight decay. Our visual context as regularization hypothesis also sheds light on why visual context is helpful on non-grounded low-resourced datasets like IWSLT'14.