VisDNMT: Improving Neural Machine Translation via Visual Knowledge Distillation

Anonymous ACL submission

Abstract

Multi-modal machine translation (MMT) is the research field that aims to improve neural machine translation (NMT) models with visual knowledge. While existing MMT systems achieve promising performance over text-only NMT methods, they typically require paired text and image as input, which limits their applicability to general translation tasks. To benefit general translation with visual knowledge, we propose *VisDNMT*, which distills visual knowledge from a pre-trained multilingual visual-language model to help translation. In particular, we train a transformer-based model jointly with a standard cross-entropy loss for translation and a knowledge distillation (KD) objective that aligns its language embedding with vision contextualized language embedding of the teacher model. VisDNMT achieves consistently higher gains over text-only NMT baselines, compared to state-of-art methods on rich and sparse visually grounded text.

1 Introduction

Existing multi-modal machine translation research explores improving translation with an auxiliary image input. However, the need for images limits their real-world application. This has motivated recent work on improving translation with visual knowledge by retrieving relevant images as an alternative image input for MMT models. (Zhang et al., 2020; Fang and Feng, 2023). While the retrieval-based MMT demonstrate improved performance over text-only methods, the overall advantages of these approaches are still marginal due to the requirement of an image database for retrieval. For instance, sentence-level image retrieval approach in (Zhang et al., 2020; Tang et al., 2022) suffers from poorly matched images when visual context in source text is sparse. The fine-grained noun-phrase to image-region retrieval approach in (Fang and Feng, 2022) ignores other potential visual-language interactions like activity-based scene and verbs. Lastly, existing retrieval-based methods focus on capturing connection between source text and visual knowledge, but ignore visual-knowledge grounding during translated text generation.

To address these limitations, we propose a simple framework, *VisDNMT*, which dynamically incorporates visual knowledge into translation by distilling visually grounded language representation from a pre-trained model as shown in figure 1. We jointly optimize a transformer architecture with two objectives: (1) translating text from source to target language. (2) distilling from a pre-trained multi-modal model to enrich language representations with visual knowledge. We use MCLIP(Chen et al., 2023), a multilingual variant of CLIP(Radford et al., 2021a), as the teacher model for KD. As CLIP learns cross-modal alignment from image-text pairs via contrastive learning, it captures enriched visual-language interactions for a large vocabulary. We also introduce a simple bi-directional translation (BT) training curriculum to enhance visual grounding of the target language representation. Unlike conventional translation that only translate from source to target, we optimize the model to translate the sentence in both directions (i.e source→target) along with the KD objective. Experiments on both caption(rich visual grounding) and news text translation(sparse visual grounding) corpus show that VisDMT achieves consistently higher gains over the text-only NMT compared to the previous methods.

2 Related Work

Multi-modal Machine Translation Multi-modal machine translation (MMT) research aims to enhance NMT models with additional visual knowledge (Bahdanau et al., 2015; Huang et al., 2016; Caglayan et al., 2017; Calixto and Liu, 2017; Zhou et al., 2018; Yao and Wan, 2020; Yin et al., 2020; Caglayan et al., 2021). The

majority of existing MMT systems require an image as the auxiliary input, restricting their ability to benefit general translation tasks when tte paired image is not available. To overcome this bottleneck, Zhang et al. (2020); Tang et al. (2022) propose to retrieve images for the source sentence to provide related visual context based on the topic words. (Fang and Feng, 2022) further improve the quality of retrieved visual context by mapping noun phrases in source text to image region features. However, the retrieval based methods are prune by the limited cross-modal grounding that mainly focuses on the alignment of noun-phrases to objects. We propose enhancing NMT models via knowledge distillation from a large pre-trained vision-language model, where the teacher model has learnt dynamic interactions between modalities.

Knowledge Distillation Knowledge Distillation (KD) is the process of transferring knowledge from a teacher model to a student model, which has been studied in a wide range of research topics (Kim and Rush, 2016; He et al., 2019; Chebotar and Waters, 2016). Our work is highly related to the KD research that aims to transfer knowledge across different modalities (Tang et al., 2021; Tuong Do, 2019; Tian et al., 2020). We transfer knowledge of a multilingual vision-language pretrained teacher model to student model for machine translation.

3 VisDNMT

VisDNMT is jointly optimized for two tasks: (1) translating text $\{x_i\}_{i=1}^N$ in language X to target text $\{y_i\}_{i=1}^N$ in language Y. (2) Distilling the language representation of x_i and y_i from the pre-trained multilingual vision language model. Figure 1 illustrates the overall architecture of our approach.

3.1 Teacher Model

For the teacher model, we use large pre-trained multimodal MCLIP (Carlsson et al., 2022). MCLIP extends the CLIP architecture (Radford et al., 2021b) to train a shared embedding space between a multilingual text-encoder and a visual encoder via contrastive learning objective among large scale image-text pairs in various languages. The text-encoder of MCLIP is initialized from the weights of MBERT, a multilingual variant of the BERT language model (Devlin et al., 2019). Multilingual CLIP is trained on 40k sentences for each language from the combined descriptions of GCC (Sharma et al., 2018) + MSCOCO (Lin et al., 2014) + VizWiz datasets (Gurari et al., 2020; Simons et al., 2020) and their translations in 69 languages. After pre-training, MCLIP obtains visual knowledge enriched text representation for all languages.

3.2 Student Model

For the student model, we adopt a transformer based encoder-decoder architecture for translation. Unlike the conventional bilingual translation model that adopts different embedding layer for various languages, we use a pre-trained shared embedding initialized from the embedding layer of M-BERT (Devlin et al., 2019) to extract language feature for all languages. The student model is then optimized with a weighted sum of translation loss and the knowledge distillation loss.

$$\mathcal{L} = \mathcal{L}_{\text{trans}} + w * \mathcal{L}_{\text{KD}} \tag{1}$$

3.3 Knowledge Distillation Objectives

To distill knowledge from MCLIP, we align the visually grounded text representations from MCLIP with text representations from the student model. This is done using the knowledge distillation objective: **Neuron Selectivity Transfer (NST)** (Huang and Wang, 2017). The NST objective is used to align the activation patterns of teacher and student neurons using squared maximum mean discrepancy ($\mathbf{MMD^2}$) between the student and teacher neurons. This mimics the neuron selectivity patterns of the teacher model to the student model whose architecture could be much smaller than the teacher.

The $\mathcal{L}_{\text{KD}}$ KD loss is given as $\mathbf{MMD^2}$ between the student and teacher probability distribution calculated using the following equation with kernel trick as in Tang et al. (2021):

$$\mathcal{L}_{\text{KD}} = \frac{1}{d^2}\sum_{i=1}^{d}\sum_{i'=1}^{d} k(\mathbf{s_i}, \mathbf{s_{i'}}) - \frac{2}{d^2}\sum_{i=1}^{d}\sum_{j=1}^{d} k(\mathbf{s_i}, \mathbf{t_j}) + \frac{1}{d^2}\sum_{j=1}^{d}\sum_{j'=1}^{d} k(\mathbf{t_j}, \mathbf{t_{j'}})$$

where d is the dimension of the student and teacher neurons, k is the kernel function, t_i is the teacher s_j is the student neuron distribution. We use polynomial kernel as provided in Tang et al. (2021)'s implementation $k(\mathbf{s}, \mathbf{t}) = (\mathbf{s}^\top \mathbf{t} + c)^d$; $c = 0$; $d = 2$

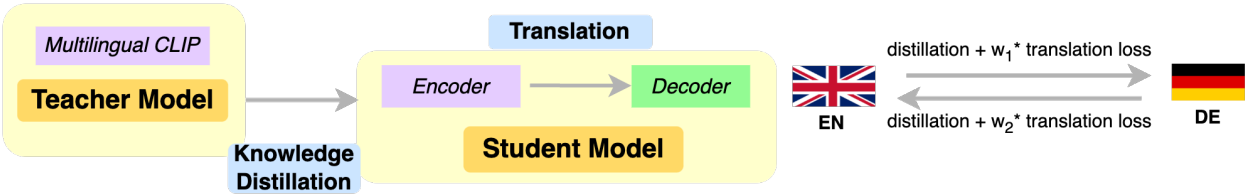

Figure 1: An overview of the architecture of VisDNMT. The student NMT model is jointly optimized with two training objectives: (1) Knowledge distilling from a frozen teacher model M-CLIP. (2) Translating between source and target languages. We also practice bi-directional translation where we translate both source to target language and vice-versa.

3.4 Bi-Direction Translation

To enrich language representation of both source text x_i and the target text y_i, for each step we optimize the student model to translate from x_i to y_i and y_i back to x_i. So the final loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{trans}}(\mathrm{x_i}, \mathrm{y_i}) + w * \mathcal{L}_{\text{KD}}(\mathrm{x_i}) + \mathcal{L}_{\text{trans}}(\mathrm{y_i}, \mathrm{x_i}) + w * \mathcal{L}_{\text{KD}}(\mathrm{y_i})$$

4 Experimental Setting

4.1 Datasets

Multi30k: a multi-lingual image captioning dataset (Elliott et al., 2016a) to study how visual knowledge would benefit the caption translation which has a good amount of content visually grounded on the paired images. We only use the EN-DE (Elliott et al., 2016b) and EN-FR (Elliott et al., 2017) language pairs without images for our analysis. We report results on both test2016 and test2017 splits.

WMT'16: A news translation benchmark[1] to understand the effect of visual knowledge to benefit translation of the text where the visually grounding context is sparse. Following Tang et al. (2022), we sample 100k bilingual pairs from 4.5 million from WMT'16 EN-DE training split to focus on studying the effect of visual information when the training data scale is limited. We use newstest2016 as the test and newstest2014 as the validation set.

4.2 Baselines

We compare VisDNMT to image-retrieval based MMT systems that aim to generalize MMT system for translation setting when the image pair is not available, including PL-UVR (Fang and Feng, 2022), UVR-NMT (Zhang et al., 2020), and IR-NMT (Tang et al., 2022). The details of these methods is covered in section 2. For each of these baselines, we also show the performance of their text-only backbone (Trans) to understand the benefits of retrieved visual content on translation. To understand the effectiveness of VisDNMT, we show results of our text-only backbone gradually adding knowledge distillation with MCLIP and bi-directional translation. As the text-encoder of MCLIP is initialized from MBERT, we also add one more ablation experiment where knowledge distillation is conducted between the student model and MBERT to further understand how visual knowledge from MCLIP can help translation.

[1] https://www.statmt.org/wmt16/

4.3 Implementation Details

We use the MCLIP tokenizer for both the student model and MCLIP to ensure token alignment. We use the M-BERT-Base-ViT-B checkpoint for MCLIP. The embedding size of the student transformer model is 512. The weights of MCLIP are frozen during training. Based on validation, we use 6 layers for base transformer and KD weight 2.0 for Multi30k. For WMT'16, we use 3 layers for base transformer and KD weight 1.25. We use early stopping on BLEU score for validation set with patience 5. VisDNMT is trained on a single NVIDIA-TITAN RTX GPU using a learning rate of 0.0001 with Adam optimizer (Kingma and Ba, 2014) and 0.1 label smoothing.

5 Results & Analysis

We compare the BLEU scores of each method with their own text-only backbone in table 1.

Results on Multi30K On Multi30K, PLUVR dominates on both English to German and English to French over all other methods. The advantage of PLUVR is largely credited to a strong text-only backbone, which outperforms other backbones by a large margin. Comparing every method to its text-only backbone, VisDNMT achieves the most significant gain with a minimum boost of 1.8 BLEU on EN-DE and 2.6 BLEU on EN-FR. We think the statistically significant gain achieved by

System	Architecture	Multi30k				WMT'16
		EN-DE		EN-FR		EN-DE
		test2016	test2017	test2016	test2017	100k
UVR-NMT	Trans	35.59	26.31	57.88	48.55	-
	+VR	$35.72^{+0.13}$	$26.87^{+0.56}$	$58.32^{+0.44}$	$48.69^{+0.14}$	-
PLUVR	Trans	39.87	31.78	60.51	52.44	-
	+VR	$\mathbf{40.30}^{+0.43}$	$\mathbf{33.45}^{+1.67}$**	$\mathbf{61.31}^{+0.8}$*	$\mathbf{53.15}^{+0.71}$*	-
IR-NMT	LSTM	37.77	-	-	-	7.99
	+VR	$38.43^{+0.66}$	-	-	-	$8.41^{+0.42}$
Ours	Trans	36.00	27.0	53.99	45.58	7.36
	+MBERT KD	$37.1^{+1.1}$**	$28.16^{+1.16}$**	$54.44^{+0.45}$	$48.56^{+2.98}$*	$8.17^{+0.81}$*
	+KD	$37.51^{+1.51}$**	$28.77^{+1.77}$**	$56.58^{+2.59}$*	$48.72^{+3.14}$*	$8.32^{+0.96}$*
	+KD +BT	$38.45^{+2.45}$**	$28.83^{+1.83}$**	$56.63^{+2.64}$*	$48.83^{+3.25}$*	$\mathbf{8.48}^{+1.12}$*

Table 1: Results of VisDNMT in comparison to baselines. * means that the BLEU scores between the transformer baseline without KD and with KD are statiscally significant with p<0.05; ** is for p<0.01.

Source text	Target text	Text-only backbone	VisDNMT
A man fixing a little girl's bicycle.	Ein Mann repariert das Fahrrad eines kleinen Mädchens (A man repairs a little girl's bicycle)	Ein Mann repariert ein kleines Mädchen (A man repairs a little girl)	Ein Mann repariert das Fahrrad eines kleinen Mädchens (A man repairs a little girl's bicycle)

Figure 2: Qualitative comparison between the translation from VisDNMT and the Text-only Backbone. Text-only backbone highlighted in red misses the context of the bicycle while VisDNMT correct captures it (highlighted green).

VisDNMT over the text-only backbone is mainly due to the following reasons. First, the visually grounded language representations distilled from MCLIP have more accurate cross-modal grounding than retrieved images and image regions. Second, grounding both source and target language on shared visual concepts aligns the two languages using vision as the pivot. We observe that translation performance improves slightly using MBERT as the teacher model, which shows that the pre-trained multilingual language encoder of MCLIP benefits the translation task to some extent. However, when the teacher model is replaced with MCLIP, the student model achieves consistent improvement over all languages in Multi30K, demonstrating the effectiveness of the visual knowledge to benefit translation. Adding bi-directional translation with KD further improves the translation quality.

Results on WMT16 We observe similar trends on news translation benchmark as on Multi30K. Even though our text-only baseline is worse than that of IR-NMT, VisDNMT outperforms IR-NMT. This demonstrates the effectiveness of VisDNMT of benefitting general translation setting with visual knowledge even on sparse visually grounded text.

Qualitative Analysis To further analyze how visual knowledge impacts translation quality, we conduct qualitative analysis by comparing translations from VisDNMT with its text-only backbone. We observe some visual cues like noun phrase-verb alignment are accurately captured by the model with KD. For eg in Figure 2, model with KD correctly generates "Ein Mann repariert das Fahrrad eines kleinen Mädchens" while the model without KD misses the bicycle and generates "Ein Mann repariert ein kleines Mädchen" (A man repairs a little girl). More examples and human evaluation to provide statistical evidence of better translation from VisDNMT are given in Appendix A.

6 Conclusion

In this work, we propose VisDNMT that distills visual knowledge from a pre-trained multi-lingual vision language model to benefit general translation task without paired images. On both Multi30K and WMT'16, we observe that our approach outperforms the backbone NMT with a consistently larger margin than previous image-retrieval methods. Ablation study also verifies the effectiveness of VisDNMT to learn visually enriched language representations and bi-directional translation to connect source and target languages via the shared visual ground. In the future, we plan to further investigate the effectiveness of knowledge distillation on more general translation benchmarks whose texts are weakly grounded on visual concepts.

Limitations

Training NMT with knowledge distillation and bidirectional translation will significantly increase the training time and the occupied GPU memory, which makes it challenging evaluate them on large translation benchmarks. Besides, the pre-trained MCLIP uses machine translated sentences to learn visually grounded language representation, whose noise can also hurt the performance of the translation model. We also only evaluate our method on two small scale translation benchmarks and small amount of languages. The current conclusion may not generalize to the evaluation on larger benchmark or other languages. We leave this exploration to the future work.

Ethics Statement

By distilling information from pre-trained MCLIP, our model will capture any biases (however limited) that the teacher model has learnt from its training corpus. Therefore, we do not recommend use our model for any real word translation task but only for research purposes. To ensure the reproducibility of our experiment results, we provide details of the hyperparameter setting in our paper and will also publish our code later.

References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

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. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439, Copenhagen, Denmark. Association for Computational Linguistics.

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1317–1324, Online. Association for Computational Linguistics.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

Yevgen Chebotar and Austin Waters. 2016. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023. mCLIP: Multilingual CLIP via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016a. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

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

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5687–5698, Dublin, Ireland. Association for Computational Linguistics.

Qingkai Fang and Yang Feng. 2023. Understanding and bridging the modality gap for speech translation. In *Proceedings of the 61st Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 15864–15881, Toronto, Canada. Association for Computational Linguistics.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 417–434. Springer.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. 2019. Bag of tricks for image classification with convolutional neural networks. In *CVPR 2019*.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

Zehao Huang and Naiyan Wang. 2017. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Rachel N Simons, Danna Gurari, and Kenneth R Fleischmann. 2020. " i hope this is helpful" understanding crowdworkers' challenges and motivations for an image description task. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26.

ZhenHao Tang, XiaoBing Zhang, Zi Long, and XiangHua Fu. 2022. Multimodal neural machine translation with search engine based image retrieval. In *Proceedings of the 9th Workshop on Asian Translation*, pages 89–98, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. Vidlankd: Improving language understanding via video-distilled knowledge transfer. *Advances in Neural Information Processing Systems*, 34:24468–24481.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive representation distillation. In *International Conference on Learning Representations*.

Huy Tran Erman Tjiputra Quang D. Tran Tuong Do, Thanh-Toan Do. 2019. Compact trilinear interaction for visual question answering. In *ICCV*.

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

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3035, Online. Association for Computational Linguistics.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

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

A Qualitative Analysis

We also conduct qualitative analysis to compare the generated translation from VisDNMT to the transformer backbone. We observe some common visual cues being accurately translated by the model with KD as compared to without. For example, for the ground truth sentence "Ein Mann repariert das Fahrrad eines kleinen Mädchens"(meaning: A man repairs a little girl's bicycle), model with KD generates the correct sentence "Ein Mann repariert das Fahrrad eines kleinen Mädchens" while the model without KD misses the bicycle altogether and generates "Ein Mann repariert ein kleines Mädchen" (meaning: A man repairs a little girl). Here we can observe that visual cues like "the bicycle" which needs to be repaired and not "the girl" are captured well by the model with KD. Another example is the ground truth sentence "Mehrere Frauen führen vor einem Gebäude einen Tanz auf"(meaning: Several women perform a dance in front of a building), the model without KD don't translate additional visual details like "the building" while the model with KD is able to translate them well. The model without KD generates "Mehrere Frauen führen einen Tanz vor"(meaning: Several women perform a dance) while the model with KD geenrates "Mehrere Frauen führen vor einem Gebäude einen Tanz auf"(meaning: Several women perform a dance in front of a building) which is the correct sentence. Observing various examples, it is evident that the model with visual knowledge distillation is able to trnaslate visual details like "where the task is happening" or "in what manner". The model without KD is not able to understand common visual cues like "the girl" is not the thing to be repaired while the model with KD understands that it's the cycle that the man repairs. You can refer to more examples in Fig 3.

w/ KD	w/o KD	tie	visual cues help
25	4	21	22

Table 2: Human Evaluation on 50 random generated translation from VisDNMT and the text-only model without knowledge distillation from Multi30K EN-DE 2016 test split.

Human Evaluation To provide statistical evidence of the better translation from the proposed VisDNMT than the original NMT, we conducted a human evaluation on 50 random samples of EN-DE Multi30k test predictions from the two methods. We ask a german speaking human evaluator to compare the two generated translations against the source sentence, where they need to identify which one is a better translation or if the quality of both translations is similar. As we can see from Table 2, human evaluators prefer translations with KD 25 times, without KD 4 times, similar quality for 21 times. The result demonstrates that translation from VisDNMT has a significantly better quality than the ones from the base NMT model. To further understand whether the better translation is led by the visual knowledge distilled from MCLIP, we also ask the human evaluator to check if the better translation of VisDNMT is due to words such as nouns, verbs, adjectives that are more grounded onto visual context instead of grammar, fluency, and conciseness. As can be seen from the results in Table 2, the majority of the better translations are influenced by visual cues.

	Source text	Text-only backbone	VisDNMT
	Fans cheer while the band plays a song.	Fans jubeln und spielen ein Lied. (Fans cheer and play a song)	Fans jubeln, während die Band ein Lied singt. (Fans cheer as the band sings a song)
	A guy in a white shirt strolls in hand with a drink.	Ein Mann in einem weißen Hemd läuft mit einem Getränk in der Hand. (A man in a white shirt runs with a drink in his hand)	Ein Mann in einem weißen Hemd geht mit einem Getränk in der Hand spazieren. (A man in a white shirt walks with a drink in his hand)
	Four girls and a woman learn to tinker.	Vier Mädchen und eine Frau lernen wie Kunstwerke. (Four girls and a woman learn like works of art)	Vier Mädchen und eine Frau lernen, wie man etwas Kunstwerke fertig machen. (Four girls and a woman learn how to finish some works of art)

Figure 3: Qualitative comparison between the translation from VisDNMT and the Text-only Backbone. Text-only backbone highlighted in red misses the context the band (top), man "walking" while holding a drink (middle), "how to make works of art" (bottom) while VisDNMT captures them (highlighted in green).