On Leveraging the Visual Modality for Neural Machine Translation: A Case Study on the How2 Dataset

Vikas Raunak [*1] **Sang Keun Choe** [*1] **Quanyang Lu** [*1] **Yi Xu** [*1] **Florian Metze** [1]

Abstract

Leveraging visual modality effectively for Neural Machine Translation (NMT) remains an open problem in computational linguistics. We posit that effectively leveraging visual information requires reconciliation of the high-level visual features (e.g., derived from action recognition, scene descriptions etc.) and low-level text features (at word or subword level). In this work, we address this semantic gap between text and the visual modality by carefully selecting the places of fusion of text and visual features for the Machine Translation task. We propose and evaluate 3 novel techniques, each along a key component in the Sequence-to-Sequence transduction pipeline, namely step-wise decoder fusion, multimodal attention modulation and visual-semantic supervision to effectively leverage the higher-level visual modality for token prediction. However, we get only modest incremental gains by the application of each technique when compared against a strong Sequence-to-Sequence baseline model, leading us to the conclusion the features provided in the How2 dataset do not lend themselves to increasing the discriminativeness between the vocabulary elements at token level prediction. We further validate this claim by comparing the visual features against the Multi30K dataset through Principal Component Analysis, wherein we find that the How2 visual feature space is even less discriminative in terms of the visual context provided.

1. Introduction

A number of works have explored integrating the visual modality for Neural Machine Translation (NMT) models (Sanabria et al., 2018b), though, there has been relatively modest gains or no gains at all by incorporating the visual modality in the translation pipeline (Caglayan et al., 2019). In particular, (Elliott & Kádár, 2017) leverage multitask learning, (Sanabria et al., 2018b) use visual adaptive training, while (Caglayan et al., 2016; Libovickỳ & Helcl, 2017; Huang et al., 2016) use a number of modality fusion techniques to incorporate features obtained from the visual modality.

Regarding the seemingly low utility of visual modality in machine translation, (Lazaridou et al., 2014) hypothesize that highly relevant visual properties are often not represented by linguistic models because they are too obvious to be explicitly mentioned in text (e.g., birds have wings, violins are brown). Similarly, (Louwerse, 2011) argue that perceptual information is already sufficiently encoded in textual cues. However, recently (Caglayan et al., 2019) have demonstrated that neural models are capable of leveraging the visual modality for translations, albeit under limited source side context. We draw upon their work and posit that since Neural models are capable of leveraging visual modality under limited source side context, they can be effectively exploited to *enhance* the discriminativeness between the vocabulary elements at token level predictions even under the presence of full linguistic context. To this end, we hypothesize that to effectively use the visual context to improve token level predictions, we must reconcile the intrinsic feature abstraction discrepancy between natural language and visual modalities and to address this problem we propose three novel techniques to integrate visual features in Neural Machine Translation.

In the upcoming sections, we first describe the How2 Multimodal Machine translation dataset and the baseline used throughout the experiments. We then outline our proposed approaches in detail, followed by experimental results and analysis of the proposed mechanisms. Finally, we list the key differences between the proposed techniques and the related literature and conclude by pointing out a few directions for further investigation.

1.1. The How2 Dataset

Throughout our experiments, we use the 300 hours subset of How2 [1](Sanabria et al., 2018a) dataset, which contains 300 hours of videos, sentence-level time alignments to the ground-truth English subtitles, and Portuguese translations of English subtitles. Detailed statistics of the dataset are presented in Table 1. The How2 dataset has 2048 dimensional

[1]Dataset Links https://github.com/srvk/how2-dataset

pre-trained ResNet embeddings (action features) (Xie et al., 2017) available for each of the video clips aligned to the sentences.

		Videos	Hours	Clips/Sentences
300h	train	13,168	298.2	184,949
	val	150	3.2	2,022
	test	175	3.7	2,305
	held	169	3.0	2,021
2000h	train	73,993	1766.6	-
	val	2,965	71.3	-
	test	2,156	51.7	-

Table 1. Statistics of How2 dataset

1.2. Baseline Sequence-to-Sequence Model

Our baseline model is the canonical Sequence-to-Sequence (Seq2Seq) model (Sutskever et al., 2014) consisting of bidirectional LSTM as encoder and decoder, general Bahadanau attention (Bahdanau et al., 2014) and Length normalization (Wu et al., 2016). In all cases we use an embedding size of 300 and hidden size of 512.

Further, whenever the visual modality is used, we encode each of the visual features through a video encoder which is also trained end-to-end with the Seq2Seq model . Figure 1 outlines all of the proposed techniques, which we describe later. As illustrated in Figure 1, the Video Frame Encoder consists of a linear layer, followed by ReLU non-linearity and a batch norm layer.

2. Proposed Methods

In this section, we describe the three proposed methods.

2.1. Step-Wise Decoder Fusion

Our first proposed technique is the step-wise decoder fusion of visual modality during every prediction step. The motivation behind this technique comes from the fact the visual encoding is provided at the sentence level, while the decoder has to make prediction for the next token at the token level. Therefore, instead of passing a single visual context at the beginning of the decoding process as in (Huang et al., 2016), we concatenate the visual encoding obtained from the video frame encoder as context at each step of the decoding process. Our hypothesis is that by proving the decoder with the sentence-level visual encoding at each step, the decoder would be able to leverage the visual encoding for any specific token prediction, and ignore it when it is not necessary. This partially solves the problem of the abstraction-level discrepancy between the features, since the visual encoding now acts as a feature during each prediction step. The technique is illustrated in Figure 1.

2.2. Multimodal Attention Modulation

Another aspect of the Sequence-to-Sequence model where the visual modality could be fused to induce more discriminativeness in the prediction is attention computation.

2.2.1. Motivation

For the proposed technique of multimodal attention modulation, we derive the motivation from a number of neuroimaging studies, which indicate that processing a word activates areas in the brain that correspond to the associated sensory modality of its semantic category, e.g. action-related words like "kick" trigger activity in the motor cortex and object-related words like cup activate visual areas. Further, it has been widely accepted that conceptual and sensorimotor representations interact with each other. (Garagnani & Pulvermüller, 2016) In the proposed technique, we model this interaction through the attention mechanism, which acts a mediator between modalities. i.e. it modulates the information flow in one modality (natural language) by input from another modality similar to human perception (De Vries et al., 2017).

2.2.2. Multimodal Attention

We consider a simplified version of the Bahdanau attention proposed in (Bahdanau et al., 2014). It is referred to as general attention in (Luong et al., 2015). We first reiterate the attention mechanism. In general attention, we first consider all the hidden states of the encoder when deriving the context vector c_t. Then, a variable-length alignment vector a_t, whose size equals the number of time steps on the source side, is derived by comparing the current target hidden state h_t with each source hidden state h_t. And the score function is a content based scoring mechanism as described below:

$$\boldsymbol{a}_t(s) = \text{align}\left(\boldsymbol{h}_t, \overline{\boldsymbol{h}}_s\right)$$

$$\boldsymbol{a}_t(s) = \frac{\exp\left(\text{score}\left(\boldsymbol{h}_t, \overline{\boldsymbol{h}}_s\right)\right)}{\sum_{s'} \exp\left(\text{score}\left(\boldsymbol{h}_t, \overline{\boldsymbol{h}}_{s'}\right)\right)}$$

$$\text{score}\left(\boldsymbol{h}_t, \overline{\boldsymbol{h}}_s\right) = \boldsymbol{h}_t^\top \boldsymbol{W}_{\boldsymbol{a}} \overline{\boldsymbol{h}}_s$$

For a multimodal extension of this attention, we propose to use the encoding obtained from the Video Frame Encoder to calculate an attention distribution over the source encodings. We use the same form for attention computation as above except that the visual encoding from the Video frame encoder is used to compute the scores:

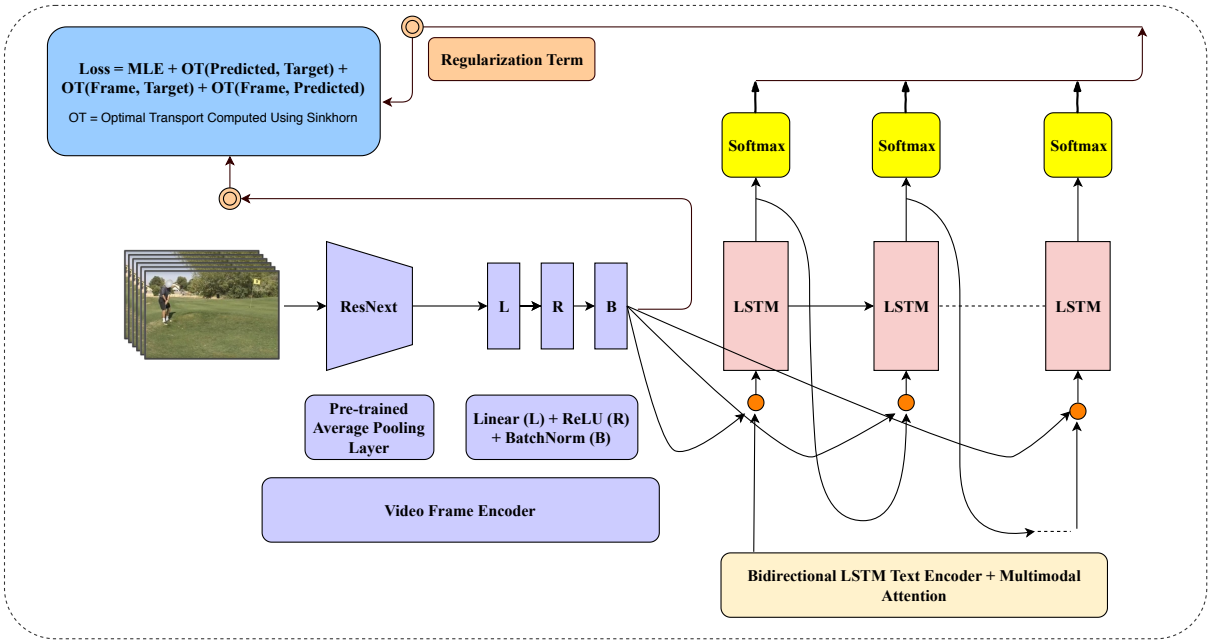

Figure 1. Decoder of the Visually Grounded Machine Translation Model with various Proposed Components

$$\boldsymbol{a}_{tv}(s) = \text{align}\left(\boldsymbol{v}_t, \overline{\boldsymbol{h}}_s\right)$$

$$\boldsymbol{a}_{tv}(s) = \frac{\exp\left(\text{score}\left(\boldsymbol{v}_t, \overline{\boldsymbol{h}}_s\right)\right)}{\sum_{s'} \exp\left(\text{score}\left(\boldsymbol{v}_t, \overline{\boldsymbol{h}}_{s'}\right)\right)}$$

$$\text{score}\left(\boldsymbol{v}_t, \overline{\boldsymbol{h}}_s\right) = \boldsymbol{v}_t^\top \boldsymbol{W}_{\boldsymbol{v}} \overline{\boldsymbol{h}}_s$$

Finally, the true attention distribution is computed as an interpolation between the visual and text based attention scores.

$$\boldsymbol{a}_t(s) = (1-\gamma) \cdot \boldsymbol{a}_t(s) + \gamma \cdot \boldsymbol{a}_{tv}(s)$$

2.3. Visual-Semantic Supervision

The well-known Stroop effect (Scarpina & Tagini, 2017) demonstrates the visual and language-processing modalities need to be close in the some "semantic space" to allow for fast language processing. However, there is no explicit mechanism in multi-source Sequence-to-sequence models to allow for this high-level alignment to emerge. In this section, we explicitly model this condition by using the distance between visual encoding and the predicted and target sentence embeddings as a regularizer during the training stage.

2.3.1. Optimal Transport Loss

Our second proposed technique is the inclusion of visual-semantic supervision to the machine translation model. To this end we propose a multimodal extension to a recently proposed distance based loss function. (Chen et al., 2019) proposed an optimal transport based loss function which computes the distance between the word embeddings of the predicted sentence and the target sentence and uses it as a regularizer. The purpose of this term is to provide the model with sequence level supervision. We propose to leverage this idea by including a distance term between the visual encoding (which is already at the sentence level) and the target/predicted sentence embeddings as well. The purpose of this distance term is to provide sequence level supervision by aligning the visual and text embeddings. Further, to integrate sequence level supervision, even though we closely follow and implement the method used in (Chen et al., 2019), we diverge from the specific parameterization proposed in (Chen et al., 2019). As in (Chen et al., 2019), at each time step t, we use an annealing parameter, such that when the decoder outputs a logit vector v_t, it is passed to the annealed Softmax operator to produce a prediction $\widehat{w}_t^{\text{pred}} = \text{soft-max}(\frac{v_t}{\tau})$, where τ is the the annealing parameter which is fixed to be 0.01 in our experiments. Further, we multiply a pair of source and target sequences w^{pred} and w^{tgt} by the decoder's word embedding $E^{\text{dec}} \in R^{d \times V^{\text{tgt}}}$ to obtain their corresponding vector representations $v^{\text{pred}} = E^{\text{dec}} w^{\text{pred}}$ and $v^{\text{tgt}} = E^{\text{dec}} w^{\text{tgt}}$, where V^{tgt} denotes the target vocabulary size and d is the embedding size. From here, we pass the vectors v^{pred} and v^{tgt} into the Sinkhorn solver (Cuturi, 2013) to obtain the entropy-regularized OT distance

$$L_{ot}^{\text{tgt}} = W(v^{\text{pred}}, v^{\text{tgt}}) = \min_{\pi \in \Pi(\mu^{\text{pred}}, \mu^{\text{tgt}})} \langle \pi, C \rangle - \varepsilon H(\pi),$$

where μ^{pred} and μ^{tgt} are probability distributions induced by the bag-of-words representation of v^{pred} and v^{tgt}, $\Pi(\mu^{\text{pred}}, \mu^{\text{tgt}})$ is the collection of joint probability distribution with marginals μ^{pred} and μ^{tgt}, $H(\pi) = -\sum_{i,j} \pi_{i,j}(\log(\pi_{i,j}) - 1)$ is the discrete entropy, and $C_{i,j} = c(v_i^{\text{pred}}, v_j^{\text{tgt}})$ with $c = c(x, y)$ being the cosine similarity. We utilize the Geomloss library [2], which provides a batched implementation of the Sinkhorn algorithm.

2.3.2. Visual-Semantic (VS) Regularizer

To implement the proposed Visual Semantic regularizer, we apply the same procedure described in the previous paragraph to the vectors w^{pred}, w^{tgt} and w^{img} to obtain the distance $L^{\text{img}} = W(v^{\text{pred}}, v^{\text{img}}) + W(v^{\text{tgt}}, v^{\text{img}})$. In this case, the vector v^{img} is obtained from the video-frame encoder's output, which is implemented to have the same embedding size as the decoder's embedding E^{dec}. Combining these two loss terms, our final loss can be expressed as

$$L = L_{\text{mle}} + L_{\text{ot}}^{\text{tgt}} + L^{\text{img}},$$

In practice, we find that that introducing a hyperparameter in the form below gets the best result:

$$L = (1 - \gamma) \cdot L_{\text{mle}} + \gamma \cdot (L_{\text{ot}}^{\text{tgt}} + L^{\text{img}}),$$

where γ is a hyper-parameter balancing the effect of MLE and OT. Further, this reparametrization is different from the one originally proposed (Chen et al., 2019), not only in the inclusion of new terms with the visual embeddings but also in that it rescales the maximum-likelihood term.

In practice the distance term to incorporate the visual-semantic supervision could use any metric in this formulation, not necessarily an Optimal Transport Metric.

3. Results and Analysis

3.1. Experimental Results

For all the translation experiments, we preprocess the data by lowercasing and removing the punctuations. Whenever BPE (Sennrich et al., 2015) is used, it is used with 10K vocabulary on both the source and target side, while the Sentence piece model (Kudo & Richardson, 2018) is used with 5K vocabulary on both the source and target sides. The learning rate is set to 0.001 with Adam Optimizer and a learning rate decay of 0.5 is used in all the experiments. The performances of the key models are summarized in Table 2, along with the gains in BLEU points.

From Table 2, we can make a few observations:

1. The visual modality leads to modest gains in the BLEU scores. The proposed VS regularizer leads to slightly higher gain when compared to Decoder-Fusion and Attention modulation techniques for the En-Pt language pair.

2. Table 3 shows the results of training bigger models on the En-Pt dataset using specialized machine translation techniques such as subword vocabulary. We find that using the subword vocabulary and transformer architectures lead to much larger gains, when compared to the gains coming from using visual modality on the baseline [3].

3. Further, the gains due to incorporating the visual modality are less for Multimodal Attention and VS Regularization in the case of the reversed language pair of Pt-En (Table 4), even though the visual modality is common to both the languages This may be due to the dataset creation process wherein first the videos were aligned with English sentences and then the Portuguese translations were created, implying a reduction in correspondence with the visual modality due to errors introduced in the translation process.

Methods	BLEU	Improvement
Baseline (En-Pt)	51.32	
+ Decoder Fusion (En-Pt)	51.79	**+0.47**
+ Multimodal Attention (En-Pt)	51.85	**+0.53**
+ VS Regularization (En-Pt)	52.00	**+0.68**

Table 2. BLEU Score Comparison of the proposed methods

Methods	BLEU
3 Layers LSTM + BPE (En-Pt)	54.86
Unimodal Transformer + SPM (En-Pt)	55.28

Table 3. BLEU Scores for Specialized NMT Models

3.2. Discussion

In this subsection, we inspect the dataset as well as the proposed mechanisms.

Methods	BLEU	Improvement
Baseline (Pt-En)	49.12	
+ Decoder Fusion (Pt-En)	49.68	**+0.56**
+ Multimodal Attention (Pt-En)	49.49	**+0.37**
+ VS Regularization (Pt-En)	49.31	**+0.19**

Table 4. BLEU Score Comparison of the proposed methods

[2] https://github.com/jeanfeydy/geomloss

[3] We have submitted the output of the Transformer model as an entry to the How2 Machine Translation Challenge.

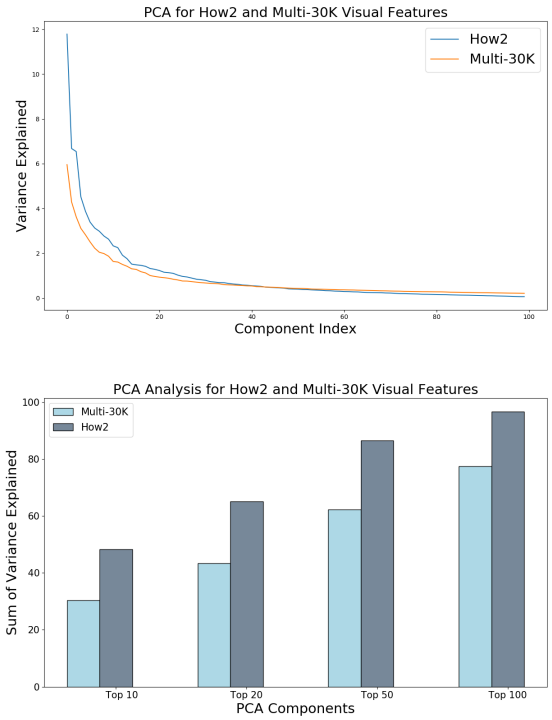

Figure 2. **Top**: Variance Explained by the Top 100 Components. **Bottom**: Cumulative Variance Explained by the Top Components.

3.2.1. PCA Based Analysis of Visual Features

To analyze the reasons for modest gains, despite incorporating multiple techniques to effectively leverage the visual modality for machine translation, we conduct an investigation of feature qualities of the How2 dataset with respect to the widely used Multi-30K dataset [4]. To analyze the discriminativeness of the visual features for both of these datasets, we leverage an analysis mechanism first used in (Mu & Viswanath, 2018) in the context of analyzing embedding discriminativeness.

Figure 2 shows the variance explained by the Top 100 principal components, obtained by applying PCA on the How2 and Multi-30K training set visual features. The original feature dimensions are 2048 in both the cases. It is clear from the Figure 2 that most of the energy of the visual feature space resides in a low-dimensional subspace (Mu & Viswanath, 2018). Figure 2 also shows the cumulative variance explained by Top 10, 20, 50 and 100 principal components respectively. It is clear that the visual features in the case of How2 dataset are much more dominated by the "common" dimensions, when compared to the Multi-30K dataset. Further, this analysis is still at the sentence level, i.e. the How2 visual features are much less discriminative among individual sentences, further aggravating the problem at the token level. This leads us to the conclusion that the How2 visual features aren't sufficient enough to expect benefits from the visual modality in Neural Machine Translation task and the problem of constructing a good Multimodal Machine translation dataset, as described in (Caglayan et al., 2019) is still open.

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

3.2.2. Comparison of Visual and Text Based Attention

In this section, we analyze the visual and text based attention mechanisms. We find that the visual attention is very sparse, in that just one source encoding is attended to, thereby limiting the use of modulation. Thus, in practice, we find that a small weight ($\gamma = 0.1$) is necessary to prevent degradation due to this sparse visual attention component.

This is also consistent with our observations that the video encodings are similar for sentences belonging to the same video. Figure 3 shows the comparison of visual and text based attention for the source sentence, *'they do have salaries they have many sponsors they have managers and everything like that getting them jobs all the time'* which is translated to *'eles têm salários eles têm muitos patrocinadores eles têm gerente e tudo assim fazendo trabalhos o tempo todo'*. Further, through inspection we find that the visual attention usually focuses on verbs/adverbs (e.g. in Figure 3 it is the word 'getting' and in Figure 4, it is the word 'down') or pronouns, although we haven't analyzed it quantitatively yet. Figure 4 demonstrates another comparison of the attention mechanisms, showing the same behavior.

Further, we would also like to point out an experiment on the How2 dataset which confirm our conclusion regarding the very limited discriminativeness of the visual features in the How2 dataset. We tried to fuse image captioning and machine translation models by combining the probabilities at token-level prediction. However, after training a Show-Attend-and-Tell (Xu et al., 2015) Image captioning model using the visual features in the How2 dataset, we obtained very poor results for the captions. We observed that the captions were getting repeated for different sentences. We inspected this and found that the video-frame features for these sentences were either almost same or exactly the same, further validating our conclusion.

4. Comparison to Related Work

Multimodal Attention: The proposed multimodal attention technique differs from (Caglayan et al., 2016) in that we use both the natural language as well as the visual modalities to compute attention over the source sentence, rather than having attention over images. Since, attention is computed over the same source embeddings (arising from a sin-

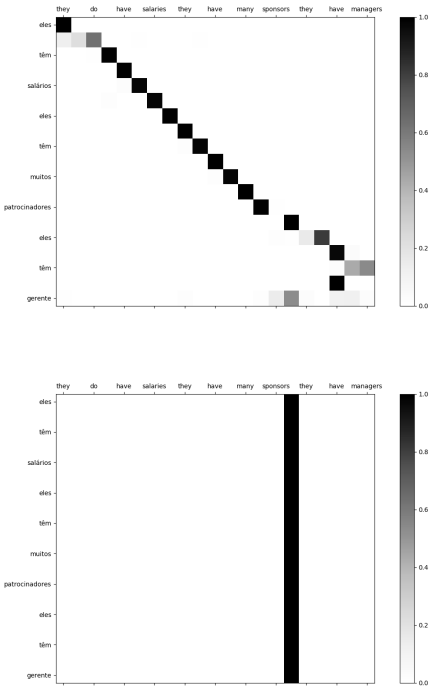

Figure 3. **Top**: Text Based Attention (Horizontal Direction Represents the Source Sentence) **Bottom**: Visual Attention for the same sentence.

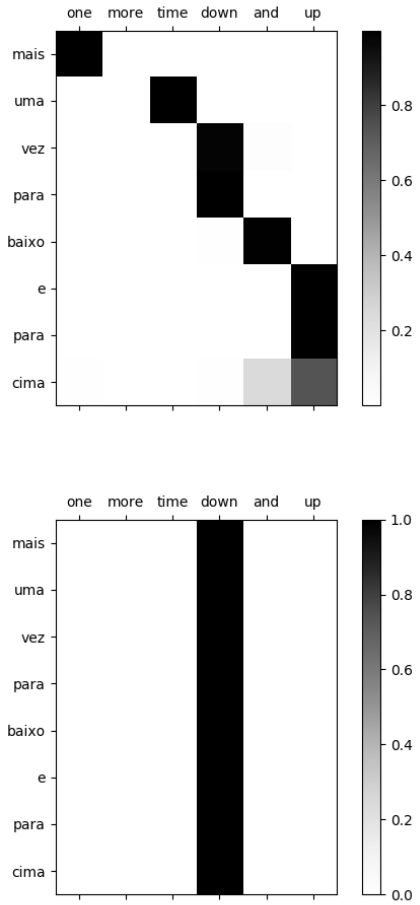

Figure 4. **Top**: Text Based Attention (Horizontal Direction Represents the Source Sentence) **Bottom**: Visual Attention for the same sentence.

gle encoder), using two different modalities, our approach also differs from (Libovickỳ & Helcl, 2017), who focus on combining the attention scores of multiple source encoders.

Step-Wise Decoder Fusion: The proposed step-wise decoder fusion approach differs from the usual practice of passing the visual feature only at the beginning of the decoding process (Huang et al., 2016).

Visual-Semantic Supervision: In terms of leveraging the visual modality for supervision, (Elliott & Kádár, 2017) use multi-task learning to learn grounded representations through image representation prediction. However, in this work, we use the distance between the visual encodings and the embeddings arising from the target and predicted sentences as a supervision signal. To our knowledge, visual-semantic supervision hasn't been much explored for multimodal translation in terms of loss functions.

5. Conclusions and Future Work

To conclude, we tried to draw upon the recent the work of (Caglayan et al., 2019) to construct techniques to better leverage the visual modality. Though our results on How2 dataset confirm the general consensus that the visual modality does not lead to any significant gains, we attribute the relatively modest gains to limited discriminativeness offered by the How2 visual features through fine-grained dataset as well as attention inspection. We hope that our work could help lead to more useful techniques and better visual features for multimodal machine translation. We intend to further extend our work by further experimenting with the proposed approaches on larger models, constructing more discriminative features for the visual modality, categorizing the visual attention distribution in terms of parts of speech tags as well as analyzing the utility of the grounded representations in other tasks.

References

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Caglayan, O., Barrault, L., and Bougares, F. Multimodal

attention for neural machine translation. *arXiv preprint arXiv:1609.03976*, 2016.

Caglayan, O., Madhyastha, P., Specia, L., and Barrault, L. Probing the need for visual context in multimodal machine translation. *arXiv preprint arXiv:1903.08678*, 2019.

Chen, L., Zhang, Y., Zhang, R., Tao, C., Gan, Z., Zhang, H., Li, B., Shen, D., Chen, C., and Carin, L. Improving sequence-to-sequence learning via optimal transport. *arXiv preprint arXiv:1901.06283*, 2019.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 2292–2300. Curran Associates, Inc., 2013.

De Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., and Courville, A. C. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pp. 6594–6604, 2017.

Elliott, D. and Kádár, A. Imagination improves multimodal translation. *arXiv preprint arXiv:1705.04350*, 2017.

Garagnani, M. and Pulvermüller, F. Conceptual grounding of language in action and perception: a neurocomputational model of the emergence of category specificity and semantic hubs. *European Journal of Neuroscience*, 43(6): 721–737, 2016.

Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pp. 639–645, 2016.

Kudo, T. and Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

Lazaridou, A., Bruni, E., and Baroni, M. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 1403–1414, 2014.

Libovickỳ, J. and Helcl, J. Attention strategies for multi-source sequence-to-sequence learning. *arXiv preprint arXiv:1704.06567*, 2017.

Louwerse, M. M. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2): 273–302, 2011.

Luong, M.-T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Mu, J. and Viswanath, P. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkuGJ3kCb.

Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., and Metze, F. How2: A large-scale dataset for multimodal language understanding. *CoRR*, abs/1811.00347, 2018a. URL http://arxiv.org/abs/1811.00347.

Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., and Metze, F. How2: A large-scale dataset for multimodal language understanding. *CoRR*, abs/1811.00347, 2018b. URL http://arxiv.org/abs/1811.00347.

Scarpina, F. and Tagini, S. The stroop color and word test. *Frontiers in psychology*, 8:557, 2017.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057, 2015.