Predicting Actions to Help Predict Translations

Zixiu Wu [1] Julia Ive [2] Josiah Wang [1] Pranava Madhyastha [1] Lucia Specia [1]

Abstract

We address the task of text translation on the How2 dataset using a state of the art transformer-based multimodal approach. The question we ask ourselves is whether visual features can support the translation process, in particular, given that this is a dataset extracted from videos, we focus on the translation of actions, which we believe are poorly captured in current static image-text datasets currently used for multimodal translation. For that purpose, we extract different types of action features from the videos and carefully investigate how helpful this visual information is by testing whether it can increase translation quality when used in conjunction with (i) the original text and (ii) the original text where action-related words (or all verbs) are masked out. The latter is a simulation that helps us assess the utility of the image in cases where the text does not provide enough context about the action, or in the presence of noise in the input text.

1. Introduction

Multimodal machine translation (MMT) (Specia et al., 2016) is one of the main applications motivating the creation of the How2 dataset (Sanabria et al., 2018). The goal was to move away from existing datasets – namely Multi30K (Elliott et al., 2016) – with static images and their corresponding simple and short descriptive captions. In the Multi30K dataset, existing work has shown that images can be beneficial, especially in the presence of noisy or incomplete input (Caglayan et al., 2019; Ive et al., 2019).

The language in the How2 dataset is not necessarily descriptive and sentences are longer, less repetitive and structurally more complex. While intuitively this should make the translation task harder and under such conditions one could expect that other modalities could be helpful, the general translation quality obtained by text-only neural machine translation models trained on this dataset is relatively high, as reported in (Sanabria et al., 2018). Additionally, there is not a very close equivalence between the visual and textual modality. For example, many videos are focused on the speaker. Therefore, making use of the additional modality becomes a much harder challenge. As a consequence, previous experiments on MMT on this data thus far have not been able to benefit from images (Sanabria et al., 2018). In this paper we further examine the question of whether visual information can be helpful by (i) using a more advanced model architecture for multimodality, (ii) testing different types of visual features and and different ways of representating these features; and (iii) concentrating on the translation of words which we believe the temporal nature of videos could help with. More specifically, in a similar way to Caglayan et al. (2019), we probe the contribution of images by masking source words to simulate the case of noisy or highly ambiguous input. We focus on actions, which are generally represented by certain verbs, as we believe this is the main additional information one can explore in videos, as compared to static images. We report experiments with a more advanced, transformer-based architecture for MMT than that exploited in Sanabria et al. (2018). Our results show that the visual features, especially those from a CNN fine-tuned for classifying videos into verb-related actions, led to considerable boost in translation quality, with varying levels of improvements for different masking settings. Human evaluation of a subset of the data confirms the automatic evaluation results.

2. Dataset and Masking Strategies

We use the How2 Sanabria et al. (2018) dataset for the experiments, keeping the standard splits:[1] 184,949 training sentences, 2,022 validation sentences and 2,305 test sentences. Our text-only baseline uses the dataset as distributed. For the masking experiments, two strategies to replace words in the source language are defined:

- **Mask action verbs (`ACT`):** All verbs which correspond to an action as defined in the action categori-

[1]Department of Computing, Imperial College London, United Kingdom [2]DCS, Sheffield University, United Kingdom. Correspondence to: Lucia Specia <l.specia@imperial.ac.uk>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

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

sations of the Moments in Time dataset (Monfort et al., 2019) are replaced by a placeholder. The masked words (tokens) make up 2.75%, 2.83%, and 2.84% of the training, validation, and test texts respectively.

- **Mask all verbs (ALL):** All verbs in the sentence are replaced by a placeholder. The masked words (tokens) make up 20.6%, 21.0%, and 20.4% of the training, validation, and test texts respectively.

The masking is performed in all sentences containing (action) verbs in the source language. For that, the data is first POS-tagged and lemmatised using spaCy 2.0.[2] In the case of action verbs, the resulting lemmatised tokens are matched against the 339 lemmatised action verbs from Monfort et al. (2019).[3] The target language remains the same for the purposes of both training and testing. We call the original unmasked sentences ORG.

Figure 1 shows some examples of segments from How2 with verbs masked using the two different strategies.

Byte Pair Encoding (BPE) (Sennrich et al., 2015) with 20,000 merge operations is applied on the target training text and each of the differently-masked source training texts separately, leading to 4 distinct vocabularies for ORG, ALL, ACT, and the target language respectively.

3. Visual features

We experiment with three types of visual features:

- **videosum:** the output of the last fully-connected layer of ResNeXt-101 (Xie et al., 2017) with 3D convolutional kernels trained to recognise 400 different actions (Hara et al., 2018);
- **conv4:** the final convolutional layer of a 3D ResNet-50 CNN trained to classify the 339 action verbs from Monfort et al. (2019);
- **emb**: a word embedding matrix for the 339 action verbs, with the embedding of each verb weighted by the final softmax layer of the same CNN for **conv4**.

videosum is provided officially by the How2 Challenge.[4] Each How2 video segment is divided into 16-frame chunks as separate inputs to the network, according to Sanabria et al. (2018), and the average of the 2048-D feature maps for all the chunks is computed as the single-vector feature of the video segment.

We extract **conv4** and **emb** features using a 3D ResNet-50 CNN trained by Monfort et al. (2019), which inflates a 2D ResNet-50 CNN pre-trained on ImageNet and fine-tuned on the Moments in Time dataset. We sample 16 equi-distant frames for each video, feed them to the network, and extract the **conv4** and **softmax** vectors from the CNN as the visual features of the video.

For **emb**, we encode each of the 339 category labels as a vector, more specifically a 300-dimensional CBOW word2vec embedding (Mikolov et al., 2013). In the case of multiword phrases, we average the embeddings for each word in the phrase. For each video and for each category label, we scale the category embedding elementwise by its corresponding CNN softmax posterior prediction.

For each video segment in our experiments, **conv4** is represented as a $7 \times 7 \times 2048$ matrix and **emb** as a 339×300 matrix. The former can be interpreted as 49 video region summaries, where each region is a cell of a 7×7 grid that divides the video spatially. The latter can be seen as a description of the video segment based only on the 339 action categories.

4. MMT model

We base our model on the **transformer architecture** (Vaswani et al., 2017) for neural machine translation. Our architecture is a multi-layer encoder-decoder using the tensor2tensor[5] (Vaswani et al., 2018) library. The encoder and decoder blocks are as follows:

Encoder Block ($\mathcal{E}$): The encoder block comprises 6 layers, with each containing two sublayers of multi-head self-attention mechanism followed by a fully connected feed forward neural network. We follow the standard implementation and employ residual connections between each layer, as well as layer normalisation. The output of the encoder forms the encoder memory which consists of contextualised representations for each of the source tokens ($M_{\mathcal{E}}$).

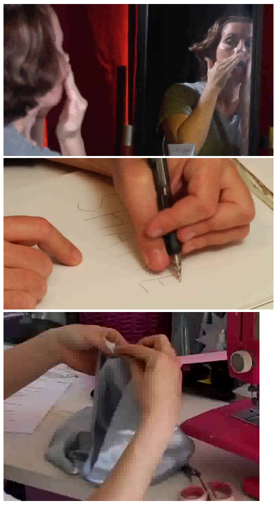
Decoder Block ($\mathcal{D}$): The decoder block also comprises 6 layers. It contains an additional sublayer which performs multi-head attention over the outputs of the encoder block. Specifically, decoding layer d_{l_i} is the result of a) multi-head attention over the outputs of the encoder which in turn is a function of the encoder memory and the outputs from the previous layer: $A_{\mathcal{D}\rightarrow\mathcal{E}} = f(M_{\mathcal{E}}, d_{l_{i-1}})$ where, the keys and values are the encoder outputs and the queries correspond to the decoder input, and b) the multi-head self attention which is a function of the generated outputs from the previous layer: $A_{\mathcal{D}} = f(d_{l_{i-1}})$.

[2] http://spacy.io/ model en_core_web_lg

[3] We retain only the verb component for specialised actions such as *playing+music* and *adult+male+singing*

[4] https://srvk.github.io/how2-challenge/

[5] https://github.com/tensorflow/tensor2tensor



■ simply apply the cleanser or cream to your hands and apply it to the face and begin rubbing.
⧫ simply apply the cleanser or cream to your hands and apply it to the face and begin V .
▲ simply V the cleanser or cream to your hands and V it to the face and V V .

■ you can draw it really lightly , go back and erase it later .
⧫ you can V it really lightly , go back and erase it later .
▲ you V V it really lightly , V back and V it later .

■ what we are going to be doing is folding the top over and making a little casing the ribbon will slip through .
⧫ what we are going to be doing is V the top over and making a little casing the ribbon will V through .
▲ what we V V to V V V V the top over and V a little V the ribbon V V through .

Figure 1. Three example segments from the How2 training dataset with verbs masked. In each example, the first line (■) shows the full text segment, the second line (⧫) shows the segment with verbs from Monfort et al. (2019) masked with V , the third line (▲) shows the segment with all verbs masked with V .

Our multimodal transformer models follow one of the two formulations below for conditioning translations on image information:

- **Additive image conditioning (AIC)** The 2048-D **videosum** feature vector is projected and then added to each of the outputs of the encoder. The projection matrix is jointly learned with the model.
- **Attention over image features (AIF)** The model attends over image features, as in Helcl et al. (2018), where the decoder block now contains an additional cross-attention sub-layer $A_{\mathcal{D}\rightarrow\mathcal{V}}$ which attends to the visual information. The keys and values correspond to the visual information. For **conv4** and **emb**, the attention is distributed across the 49 video regions and the 339 action categories, respectively. For **videosum**, the 2048-D feature vector is reshaped in row-major order into a 32 × 64 matrix, so that the attention is over the 32 rows.

Training We keep the hyperparameter settings as in Ive et al. (2019), i.e. we use the `transformer_big` parameter set with 16 heads, a hidden state size of 1024, a base learning rate of 0.05, and a dropout rate of 0.1 for layer pre- and post-processing at training time. We optimise our models with cross entropy loss and Adam as optimiser (Kingma & Ba, 2014). Training is performed until convergence.[6] We optimise the number of warmup steps during the multi-GPU training according to Popel & Bojar (2018). We apply beam search of size 10 and alpha of 1.0 for inference.

Table 1. Results for the test set. We report BLEU scores. The How2 baselines and the SOTA are from the How2 Challenge official challenge website. Bold highlights our best results.

SETUP	ORG	ACT	ALL
HOW2 UNIMODAL BASELINE	54.4	-	-
HOW2 MULTIMODAL BASELINE	54.4	-	-
HOW2 MULTIMODAL SOTA	55.5	-	-
text-only	51.7	49.6	39.0
AIC-VIDEOSUM	52.0	49.8	40.0
AIF-VIDEOSUM	52.7	50.4	40.7
AIF-CONV4	**54.9**	52.0	40.7
AIF-EMB	54.8	**52.3**	**42.4**

5. Results

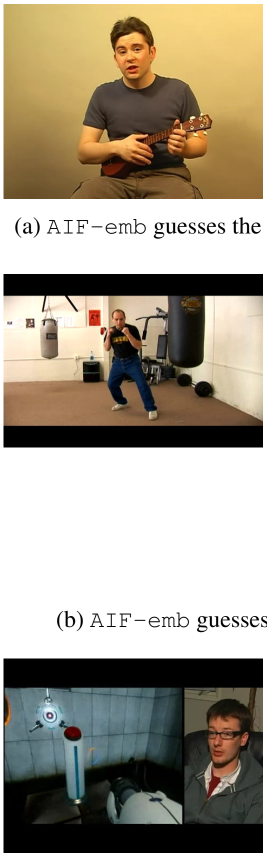
Table 1 reports the results of our experiments using BLEU (Papineni et al., 2002) as metric.[7]

Our unmasked `text-only` baseline achieves a BLEU score of 51.7. As expected, the scores for masked models are lower: for ACT, we observe a BLEU of 49.6; for ALL a BLEU of 39.0.

Overall, `AIC-videosum` brings slight improvement for ORG and ACT. For ALL, this improvement is more substantial (1.0 BLEU). `AIF-videosum` leads to more noticeable score increases: 1.0 for ORG, 0.8 for ACT, and 1.7 for ALL. This reveals the potential of a dense feature vector being utilised as segmented sub-vectors for attention as opposed to being processed as a whole.

[6] We use early stopping with patience of 10 epochs based on the validation BLEU score.

[7] We measure the performance with Multeval (Clark et al., 2011). We use tokenised and lowercased reference and hypotheses both with punctuation removed, as in the official challenge.



EN	hi, we're learning to play ukulele today.
text-only	oi, estamos aprendendo a fazer o ululele hoje.
AIF-conv4	oi, estamos aprendendo a falar do ukuele hoje.
AIF-emb	oi, estamos aprendendo a tocar o ukulele hoje.
PT	oi, estamos aprendendo a tocar cavaquinho hoje.

(a) AIF-emb guesses the masked word play correctly as *tocar*, while the other models translate as *fazer* (make) and *falar* (speak)

EN	we're going to slide the front foot and drag the back foot whenever you're moving forward.
text-only	vamos deslizar o pé da frente e deslizar o pé de trás sempre que você estiver avançando.
AIF-conv4	vamos deslizar o pé da frente e deslizar o pé de trás sempre que você seguir em frente.
AIF-emb	vamos deslizar o pé da frente e arrastar o pé de trás sempre que estiver se movendo para a frente.
PT	vamos deslizar o pé da frente e arrastar o pé de trás , sempre que você estiver se movendo para frente.

(b) AIF-emb guesses slide and drag correctly as *deslizar* and *arrastar*; the other models translate both words as *deslizar*

EN	we press this button, which opens the door.
text-only	nós viramos este botão, que bate na porta.
AIF-conv4	nós abrimos este botão, que abre a porta.
AIF-emb	nós viramos este botão, que bloqueia a porta.
PT	nós pressionamos este botão, que abre a porta.

(c) AIF-conv4 guesses open correctly as *abre*, while others translate it as *bate* (knock) or *bloqueia* (block)

Figure 2. Examples of improvements of AIF-conv4 and AIF-emb over the text-only baseline. Underlined text denotes masked words and their translations.

AIF-conv4 and AIF-emb features contribute to even greater improvements. AIF-conv4 achieves deltas of 3.2, 2.4 and 1.7 points over the text-only model respectively for ORG, ACT and ALL. The AIF-emb models achieve increase of 3.1, 2.7 and 3.4 BLEU points on the same settings.

While AIF-conv4 and AIF-emb both comfortably outperform with comparable scores the **videosum**-based models for ORG and ACT, AIF-conv4 is on par with AIF-videosum for ALL. In contrast, AIF-emb is able to achieve greater improvement for ALL than AIF-conv4 (1.7 BLEU more). For ORG, our AIF-conv4 and AIF-emb setups give around 0.5 BLEU improvement over the official task baseline.

To sum up, an important finding is the general superior performance of the multimodal models with respect to the text-only one, even in the case of unmasked words (up to 3.2 BLEU points) which is different from what was reported in previous work on this dataset (Sanabria et al., 2018). Surprisingly, this delta in BLEU between text-only and multimodal models for the masked datasets is not larger than for the original dataset. It is smaller for ACT (2.7 BLEU points), and similar for ALL (3.4 BLEU points).

Additionally, we note that in none of the model settings are the video features able to help bridge the gap between ORG and ACT performances, not even in AIF-conv4 and AIF-emb where the visual features are from a CNN finetuned for classifying videos into classes whose labels are closely related to the verbs masked in ACT. However, the gap between ORG and ALL is slightly smaller for some multimodal models than for the text-only model. Finally, our two best multimodal models with masked action verbs in ACT perform better than the text-only baseline without masked words (by 0.7 BLEU points).

To probe more into why the performance gap persists, adversarial evaluation (Elliott, 2018) could be conducted, where the visual features are shuffled and paired with mismatched texts to see if the performance of the model will be correspondingly affected. We leave this to future work.

5.1. Human Analysis

Automatic metrics often fail to capture nuances in translation quality, such as the ones we expect the visual modality to help with, which – according to human perception – lead to better translations (Elliott et al., 2017; Barrault et al., 2018). We thus performed human evaluation of our best outputs involving native speakers of Portuguese (three annotators) who are fluent speakers of English. We focused only on the evaluation for ACT.

The annotators were asked to rank randomly selected test samples according to how well they convey the meaning of the source (50 samples per annotator). For each source segment, the annotator was shown the outputs of three systems: text-only, AIF-conv4 and AIF-emb. They also had access to reference translations. A rank could be assigned from 1 to 3, allowing ties (Bojar et al., 2017). Annotators could assign zero rank to all translations if they were judged incomprehensible. Following the common practice in human evaluation for many machine translation shared tasks (Bojar et al., 2017), each system was then assigned a score which reflects the proportion of times it was judged to be better than or equal to the other two systems.

Table 2 shows the human evaluation results. They are consistent with the automatic evaluation results when it comes to the preference of humans towards the AIF-conv4 and AIF-emb setups. Between these two, AIF-conv4 was ranked better more often. Figure 2 illustrates some cases where AIF-emb or AIF-conv4 outperforms the text-only model.

text-only	AIF-conv4	AIF-emb
0.42	0.56	0.51

Table 2. Human ranking results for ACT: micro-averaged rank over three annotators.

6. Conclusions

We investigated a state of the art multimodal machine translation approach on the How2 dataset. Our focus was on exploring visual features that attempt to represent action information, and on probing their contribution when the input text is corrupted to remove action-related words. The hypothesis was that a well designed multimodal model based on informative visual features should be able to recover from the lack of textual information by leveraging the visual information. Our main results are as follows: (i) all of the multimodal models tested perform substantially better than the text-only baseline, with a gap as large as 3.2 BLEU points on experiments with unmasked text; (ii) a clear gap in performance is observed when action-related words (a subset of verbs or all verbs) are masked in the source sentence for all models, and the multimodal models are better able to recover from such a gap when all verbs are removed, as compared to text-only models. Interestingly, our two best multimodal models with masked action verbs perform better than the text-only baseline with unmasked words. These are promising results for multimodal machine translation and for the use of action-related visual features in this context.

Acknowledgement

This work was supported by the MultiMT (H2020 ERC Starting Grant No. 678017) and MMVC (Newton Fund Institutional Links Grant, ID 352343575) projects. We also thank the annotators for their valuable help.

References

Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., and Frank, S. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 304–323. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/W18-6402.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pp. 169–214. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-4717. URL http://aclweb.org/anthology/W17-4717.

Caglayan, O., Madhyastha, P., Specia, L., and Barrault, L. Probing the need for visual context in multimodal machine translation. *CoRR*, abs/1903.08678, 2019. URL http://arxiv.org/abs/1903.08678.

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 176–181, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P11-2031.

Elliott, D. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2974–2978. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/D18-1329.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. Multi30k: Multilingual english-german image descriptions. In *5th Workshop on Vision and Language*, pp. 70–74, Berlin, Germany, 2016. URL http://aclweb.org/anthology/W16-3210.

Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. 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*, pp. 215–233, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-4718.

Hara, K., Kataoka, H., and Satoh, Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, 2018.

Helcl, J., Libovický, J., and Varis, D. CUNI system for the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 616–623. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/W18-6441.

Ive, J., Madhyastha, Swaroop, P., and Specia, L. Distilling Translations with Visual Awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y. (eds.), *Proceedings of the 1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, AZ, USA, May 2013. URL http://arxiv.org/abs/1301.3781.

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfruend, D., Vondrick, C., et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–8, 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2019.2901464.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. URL http://www.aclweb.org/anthology/P02-1040.

Popel, M. and Bojar, O. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70, 2018.

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, 2018. URL http://arxiv.org/abs/1811.00347.

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

Specia, L., Frank, S., Sima'an, K., and Elliott, D. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation*, pp. 543–553. Association for Computational Linguistics, 2016. doi: 10.18653/v1/W16-2346. URL http://www.aclweb.org/anthology/W16-2346.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., et al. Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416*, 2018.

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.