Transformer-based Cascaded Multimodal Speech Translation

Zixiu Wu[1], Ozan Caglayan[1], Julia Ive[2], Josiah Wang[1], Lucia Specia[1]

[1]Department of Computing, Imperial College London, UK
[2]Department of Computer Science, University of Sheffield, UK

{zixiu.wu18, o.caglayan, josiah.wang, l.specia}@imperial.ac.uk
j.ive@sheffield.ac.uk

Abstract

This paper describes the cascaded multimodal speech translation systems developed by Imperial College London for the IWSLT 2019 evaluation campaign. The architecture consists of an automatic speech recognition (ASR) system followed by a Transformer-based multimodal machine translation (MMT) system. While the ASR component is identical across the experiments, the MMT model varies in terms of the way of integrating the visual context (simple conditioning vs. attention), the type of visual features exploited (pooled, convolutional, action categories) and the underlying architecture. For the latter, we explore both the canonical transformer [1] and its deliberation version [2] with additive and cascade variants which differ in how they integrate the textual attention. Upon conducting extensive experiments, we found that (i) the explored visual integration schemes often harm the translation performance for the transformer and additive deliberation, but considerably improve the cascade deliberation; (ii) the transformer and cascade deliberation integrate the visual modality better than the additive deliberation, as shown by the incongruence analysis.

1. Introduction

The recently introduced How2 dataset [3] has stimulated research around multimodal language understanding through the availability of *300h* instructional videos, English subtitles and their Portuguese translations. For example, [4] successfully demonstrates that semantically rich action-based visual features are helpful in the context of machine translation (MT), especially in the presence of input noise that manifests itself as missing source words. Therefore, we hypothesize that a speech-to-text translation (STT) system may also benefit from the visual context, especially in the traditional *cascaded* framework [5, 6] where noisy automatic transcripts are obtained from an automatic speech recognition system (ASR) and further translated into the target language using a machine translation (MT) component. The dataset enables the design of such multimodal STT systems, since we have access to a bilingual corpora as well as the corresponding audio-visual stream. Hence, in this paper, we propose a cascaded multimodal STT with two components: (i) an English ASR system trained on the How2 dataset and (ii) a transformer-based [1] visually grounded MMT system.

MMT is a relatively new research topic which is interested in leveraging auxiliary modalities such as audio or vision in order to improve translation performance [7]. MMT has proved effective in scenarios such as for disambiguation [8] or when the source sentences are corrupted [9]. So far, MMT has mostly focused on integrating visual features into neural MT (NMT) systems using visual attention through convolutional feature maps [10, 11] or visual conditioning of encoder/decoder blocks through fully-connected features [12, 13, 14, 15].

Inspired by previous research in MMT, we explore several multimodal integration schemes using action-level video features. Specifically, we experiment with visually conditioning the encoder output and adding visual attention to the decoder. We further extend the proposed schemes to the deliberation variant [2] of the canonical transformer in two ways: additive and cascade multimodal deliberation, which are distinct in their textual attention regimes. Overall, the results show that multimodality in general leads to performance degradation for the canonical transformer and the additive deliberation variant, but can result in substantial improvements for the cascade deliberation. Our incongruence analysis [16] reveals that the transformer and cascade deliberation are more sensitive to and therefore more reliant on visual features for translation, whereas the additive deliberation is much less impacted. We also observe that incongruence sensitivity and translation performance are not necessarily correlated.

2. Methods

In this section, we briefly describe the proposed multimodal speech translation system and its components.

2.1. Automatic Speech Recognition

The baseline ASR system that we use to obtain English transcripts is an attentive sequence-to-sequence architecture with a stacked **encoder** of 6 bidirectional LSTM layers [17]. Each LSTM layer is followed by a *tanh* projection layer. The mid-

dle two LSTM layers apply a temporal subsampling [18] by skipping every other input, reducing the length of the sequence X from T to $T/4$. All LSTM and projection layers have 320 hidden units. The forward-pass of the encoder produces the source encodings on top of which attention will be applied within the decoder. The hidden and cell states of all LSTM layers are initialized with 0. The **decoder** is a 2-layer stacked GRU [19], where the first GRU receives the previous hidden state of the second GRU in a transitional way. GRU layers, attention layer and embeddings have 320 hidden units. We share the input and output embeddings to reduce the number of parameters [20]. At timestep $t=0$, the hidden state of the first GRU is initialized with the average-pooled source encoding states.

2.2. Deliberation-based NMT

A human translator typically produces a translation draft first, and then refines it towards the final translation. The idea behind the deliberation networks [21] simulates this process by extending the conventional attentive encoder-decoder architecture [22] with a second pass *refinement* decoder. Specifically, the encoder first encodes a source sentence of length N into a sequence of hidden states $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$ on top of which the first pass decoder 1P applies the attention. The pre-softmax hidden states $\{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_M\}$ produced by the decoder leads to a first pass translation $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M\}$. The second pass decoder 2P intervenes at this point and generates a second translation by attending separately to both $\mathcal{H}$ and the concatenated state vectors $\{[\hat{s}_1; \hat{y}_1], [\hat{s}_2; \hat{y}_2], \ldots, [\hat{s}_M; \hat{y}_M]\}$. Two context vectors are produced as a result, and they are joint inputs with s_{t-1} (previous hidden state of 2P) and y_{t-1} (previous output of 2P) to 2P to yield s_t and then y_t.

A transformer-based deliberation architecture is proposed by [2]. It follows the same two-pass refinement process, with every second-pass decoder block attending to both the encoder output $\mathcal{H}$ and the first-pass pre-softmax hidden states $\hat{\mathcal{S}}$. However, it differs from [21] in that the actual first-pass translation $\hat{Y}$ is not used for the second-pass attention.

2.3. Multimodality

2.3.1. Visual Features

We experiment with three types of video features, namely average-pooled vector representations (AvgPool), convolutional layer outputs (Conv), and Ten-Hot action category embeddings (Emb). The AvgPool features are provided by the How2 dataset using the following approach: a video is segmented into smaller parts of 16 frames each, and the segments are fed to a 3D ResNeXt-101 CNN [23], trained to recognise 400 action classes [24]. The 2048-D fully-connected features are then averaged across the segments to obtain a single AvgPool feature vector for the overall video.

In order to obtain the Conv features, 16 equi-distant frames are sampled from a video, and they are then used as input to an inflated 3D ResNet-50 CNN [25] fine-tuned on the *Moments in Time* action video dataset. The CNN hence takes in a video and classifies it into one of 339 categories. The Conv features, taken at the CONV_4 layer of the network, has a $7 \times 7 \times 2048$ dimensionality.

Higher-level semantic information can be more helpful than convolutional features. We apply the same CNN to a video as we do for Conv features, but this time the focus is on the softmax layer output: we process the embedding matrix to keep the 10 most probable category embeddings intact while zeroing out the remaining ones. We call this representation ten-hot action category embeddings (Emb).

2.3.2. Integration Approaches

Encoder with Additive Visual Conditioning (Enc-Cond)

In this approach, inspired by [8], we add a projection of the visual features to each output of the vanilla transformer encoder (Enc-Van). This projection is strictly linear from the 2048-D AvgPool features to the 1024-D space in which the self attention hidden states reside, and the projection matrix is learned jointly with the translation model.

Decoder with Visual Attention (Dec-Attn)

In order to accommodate attention to visual features at the decoder side and inspired by [26], we insert one layer of visual cross attention at a decoder block immediately before the fully-connected layer. We name the transformer decoder with such an extra layer as Trans-Dec-Attn, where this layer is immediately after the textual attention to the encoder output. Specifically, we experiment with attention to Conv, Emb and AvgPool features separately. The visual attention is distributed across the 49 video regions in Conv, the 339 action category word embeddings in Emb, or the 32 rows in AvgPool where we reshape the 2048-D AvgPool vector into a 32×64 matrix.

2.3.3. Multimodal Transformers

The vanilla text-only transformer (Trans-Baseline) is used as a baseline, and we design two variants: with additive visual conditioning (Trans-Cond) and with attention to visual features (Trans-Attn). A Trans-Cond features a Enc-Cond and a vanilla transformer decoder (Dec-Van), therefore utilising visual information only at the encoder side. In contrast, a Trans-Attn is configured with a Enc-Van and a Trans-Dec-Attn, exploiting visual cues only at the decoder. Figure 1 summarises the two approaches.
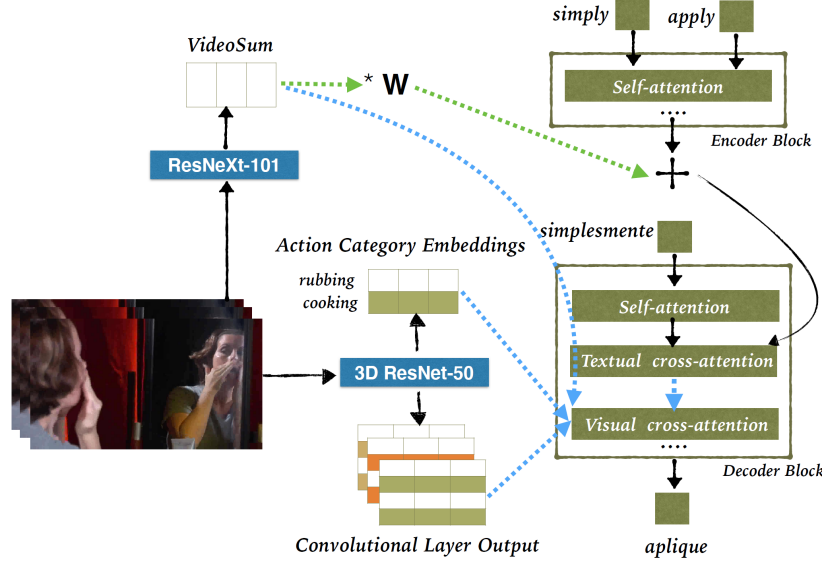
Figure 1: Unimodal and multimodal transformers: `Trans-Cond` and the `Trans-Attn` extend the text-only `Trans-Baseline` with dashed green- and blue-arrow routes, respectively. Each multimodal model activates either the dashed green-arrow route for `Trans-Cond` or one of the three dashed blue-arrow routes (i.e. VideoSum, Action Category Embeddings or Convolutional Layer Output, as shown) for `Trans-Attn`.

2.3.4. Multimodal Deliberation

Our multimodal deliberation models differ from each other in two ways: whether to use additive (`A`) [8] or cascade (`C`) **textual** deliberation to integrate the **textual** attention to the original input and to the first pass, and whether to employ **visual** attention (`Delib-Attn`) or additive **visual** conditioning (`Delib-Cond`) to integrate the **visual** features into the textual MT model. Figures 2 and 3 show the configurations of our additive and cascade deliberation models, respectively, each also showing the connections necessary for `Delib-Cond` and `Delib-Attn`.

Additive (`A`) & Cascade (`C`) Textual Deliberation

In an **additive-deliberation** second-pass decoder (`A-Delib-2P`) block, the first layer is still self-attention, whereas the second layer is the addition of two separate attention sub-layers. The first sub-layer attends to the encoder output in the same way `Dec-Van` does, while the attention of the second sub-layer is distributed across the concatenated first pass outputs and hidden states. The input to both sub-layers is the output of the self-attention layer, and the outputs of the sub-layers are summed as the final output and then (with a residual connection) fed to the visual attention layer if the decoder is multimodal or to the fully connected layer otherwise.

For the **cascade** version, the only difference is that, instead of two sub-layers, we have two separate, **successive** layers with the same functionalities.

It is worth mentioning that we introduce the attention to the first pass only at the initial three decoder blocks out of the total six of the second pass decoder (`Delib-2P`), following [8].

Additive Visual Conditioning (`Delib-Cond`) & Visual Attention (`Delib-Attn`)

`Delib-Cond` and `Delib-Attn` are simply applying `Enc-Cond` and `Dec-Attn` respectively to a deliberation model, therefore more details have been introduced in Section 2.3.2.

For `Delib-Cond`, similar to in `Trans-Cond`, we add a projection of the visual features to the output of `Enc-Van`, and use `Dec-Van` as the first pass decoder and either additive or cascade deliberation as the `Delib-2P`.

For `Delib-Attn`, in a similar vein as `Trans-Attn`, the encoder in this setting is simply `Enc-Van` and the first pass decoder is just `Dec-Van`, but this time `Delib-2P` is responsible for attending to the first pass output as well as the visual features. For both additive and cascade deliberation, a visual attention layer is inserted immediately before the fully-connected layer, so that the penultimate layer of a decoder block now attends to visual information.

3. Experiments

3.1. Dataset

We stick to the default training/validation/test splits and the pre-extracted speech features for the How2 dataset, as provided by the organizers. As for the **pre-processing**, we low-
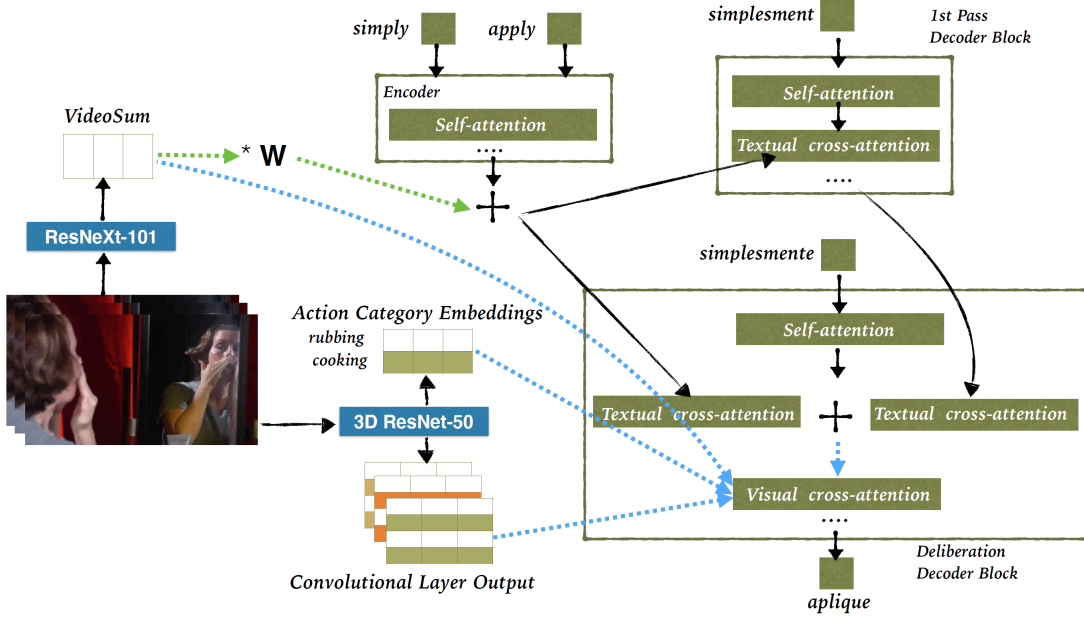
Figure 2: Unimodal and multimodal **additive** deliberation: `Delib-Cond` and `Delib-Attn` extend the text-only `Delib-Baseline` with dashed green- and blue-arrow routes, respectively. Each multimodal model activates either the dashed green-arrow route for `Delib-Cond` or one of the three dashed blue-arrow routes (i.e. VideoSum, Action Category Embeddings or Convolutional Layer Output, as shown) for `Delib-Attn`.

Figure 3: Unimodal and multimodal **cascade** deliberation: `Delib-Cond` and `Delib-Attn` extend the text-only `Delib-Baseline` with dashed green- and blue-arrow routes, respectively. Each multimodal model activates either the dashed green-arrow route for `Delib-Cond` or one of the three dashed blue-arrow routes (i.e. VideoSum, Action Category Embeddings or Convolutional Layer Output, as shown) for `Delib-Attn`.
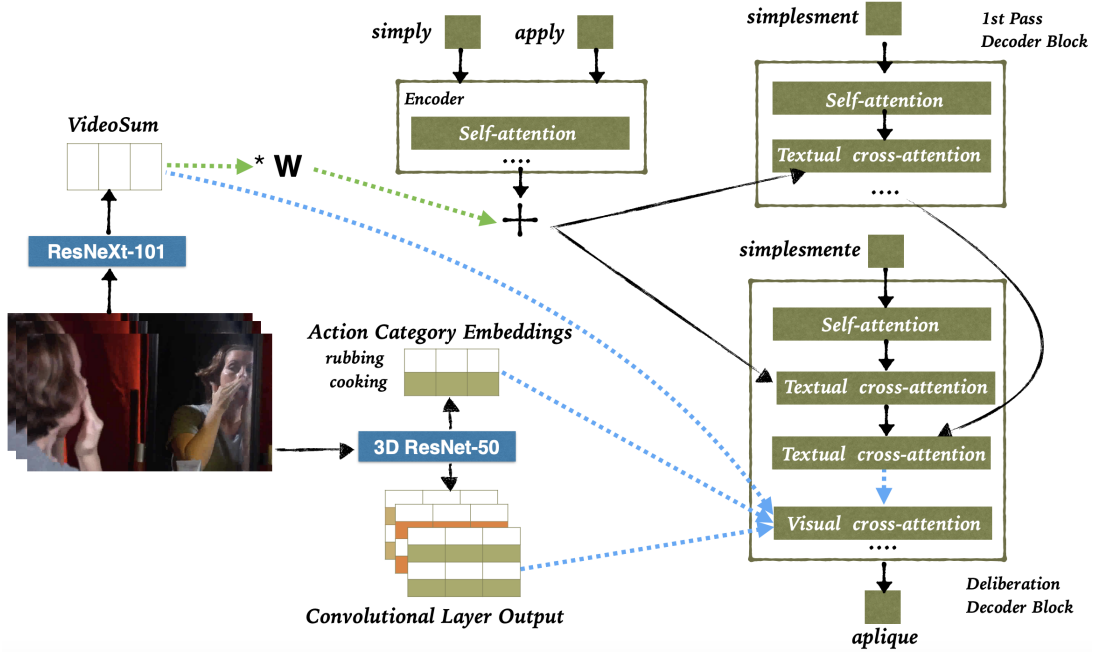
ercase the sentences and then tokenise them using Moses [27]. We then apply subword segmentation [28] by learning separate English and Portuguese models with 20,000 merge operations each. The English corpus used when training the subword model consists of both the ground-truth video subtitles and the noisy transcripts produced by the underlying

ASR system. We do not share vocabularies between the source and target domains. Finally for the **post-processing** step, we merge the subword tokens, apply recasing and detokenisation. The recasing model is a standard Moses baseline trained again on the parallel How2 corpus.

The baseline ASR system is trained on the How2 dataset as well. This system is then used to obtain noisy transcripts for the whole dataset, using beam-search with beam size of 10. The pre-processing pipeline for the ASR is different from the MT pipeline in the sense that the punctuations are removed and the subword segmentation is performed using SentencePiece [29] with a vocabulary size of 5,000. The test-set performance of this ASR is around 19% WER.

3.2. Training

We train our transformer and deliberation models until convergence largely with `transformer_big` hyperparameters: 16 attention heads, 1024-D hidden states and a dropout of 0.1. During inference, we apply beam-search with beam size of 10. For deliberation, we first train the underlying transformer model until convergence, and use its weights to initialise the encoder and the first pass decoder. After freezing those weights, we train `Delib-2P` until convergence. The reason for the partial freezing is that our preliminary experiments showed that it enabled better performance compared to updating the whole model. Following [21], we obtain 10-best samples from the first pass with beam-search for source augmentation during the training of `Delib-2P`.

We train all the models on an Nvidia RTX 2080Ti with a batch size of 1024, a base learning rate of 0.02 with 8,000 warm-up steps for the Adam [30] optimiser, and a patience of 10 epochs for early stopping based on `approx-BLEU` (`tensor2tensor`) for the transformers and 3 epochs for the deliberation models. After the training finishes, we evaluate all the checkpoints on the validation set and compute the real BLEU [31] scores, based on which we select the best model for inference on the test set. The transformer and the deliberation models are based upon the `tensor2tensor`[1] library [32] (v1.3.0 RC1) as well as the vanilla transformer-based deliberation[2] [21] and their multimodal variants[3] [8].

4. Results & Analysis

4.1. Quantitative Results

We report tokenised results obtained using the *multeval* toolkit [33]. We focus on single system performance and thus, do not perform any ensembling or checkpoint averaging.

---

[1]https://github.com/tensorflow/tensor2tensor
[2]https://github.com/ustctf/delibnet
[3]https://github.com/ImperialNLP/MMT-Delib

Table 1: BLEU scores for the test set: bold highlights our best results. † indicates a system is significantly different from its text-only counterpart (p-value ≤ 0.05).

SETUP	TRANS	A-DELIB	C-DELIB
BASELINE	**39.8**	**37.6**	36.4
COND-AVGPOOL	39.7	36.0 †	36.2
ATTN-AVGPOOL	39.7	**37.6**	37.4 †
ATTN-EMB	39.7	37.0 †	37.3 †
ATTN-CONV	**39.8**	37.2	37.0

The BLEU scores of the models are shown in Table 1. Evident from the table is that the best models overall are `Trans-Baseline` and `Trans-Attn-Conv` with a BLEU score of 39.8, and the other multimodal transformers have slightly worse performance, showing score drops around 0.1. Also, none of the multimodal transformer systems are significantly different from the baseline, which is a sign of the limited extent to which visual features affect the output.

For additive deliberation (`A-Delib`), the performance variation is considerably larger: `Attn-AvgPool` and `Baseline` take the lead with 37.6 BLEU, but the next best system (`Attn-Conv`) plunges to 37.2. The other two (`Cond-AvgPool` & `Attn-Emb`) also have noticeably worse results (36.0 and 37.0). Overall, however, `A-Delib` is still similar to the transformers in that the baseline generally yields higher-quality translations.

Cascade deliberation, on the other hand, is different in that its text-only baseline is outperformed by most of its multimodal counterparts. Multimodality enables boosts as large as around 1 BLEU point in the cases of `Attn-AvgPool` and `Attn-Emb`, both of which achieve about 37.4 BLEU and are significantly different from the baseline.

Another observation is that the deliberation models as a whole lead to worse performance than the canonical transformers, with BLEU deterioration ranging from 2.3 (across `Attn-AvgPool` variants) to 3.5 (across `Cond-AvgPool` systems), which defies the findings of [8]. We leave this to future investigations.

4.2. Incongruence Analysis

To further probe the effect of multimodality, we follow the incongruent decoding approach [16], where our multimodal models are fed with mismatched visual features. The general assumption is that a model will have learned to exploit visual information to help with its translation, if it shows substantial performance degradation when given wrong visual features. The results are reported in Table 2.

Overall, there are considerable parallels between the trans-

Table 2: Incongruent decoding results for the test set: BLEU changes are w.r.t the congruent counterparts from Table 1. † marks incongruent decoding results that are significantly different (p-value ≤ 0.05) from congruent counterparts.

SETUP	TRANS	A-DELIB	C-DELIB
COND-AVGPOOL	↓ 0.5 †	↑ 0.1	↓ 0.6 †
ATTN-AVGPOOL	↓ 0.3	0	↓ 0.1
ATTN-EMB	↓ 0.4 †	↑ 0.1	↓ 0.2 †
ATTN-CONV	↓ 0.1	↓ 0.2	↓ 0.2

formers and the cascade deliberation models in terms of the incongruence effect, such as universal performance deterioration (ranging from 0.1 to 0.6 BLEU) and more noticeable score changes (↓ 0.5 BLEU for `Trans-Cond-AvgPool` and ↓ 0.6 BLEU for `C-Delib-Cond-AvgPool`) in the `Cond-AvgPool` setting compared to the other scenarios. Additive deliberation, however, manifests a drastically different pattern, showing almost no incongruence effect for `Attn-AvgPool`, only a 0.2 BLEU decrease for `Attn-Conv`, and even a 0.1 BLEU boost for `Attn-Emb` and `Cond-AvgPool`.

Therefore, the determination can be made that `Trans` and `C-Delib` models are considerably more sensitive to incorrect visual information than `A-Delib`, which means the former better utilise visual clues during translation.

Interestingly, the extent of performance degradation caused by incongruence is not necessarily correlated with the congruent BLEU scores. For example, `Trans-Attn-Conv` is on par with `Trans-Attn-Emb` in congruent decoding (differing by around 0.1 BLEU), but the former suffers only a 0.1-BLEU loss with incongruence whereas the figure for the latter is 0.4, in addition to the fact that the latter becomes significantly different after incongruent decoding. This means that some multimodal models that are sensitive to incongruence likely complement visual attention with textual attention but without getting higher-quality translation as a result.

The differences between the multimodal behaviour of additive and cascade deliberation also warrant more investigation, since the two types of deliberation are identical in their utilisation of visual features and only vary in their handling of the textual attention to the outputs of the encoder and the first pass decoder.

5. Conclusions

We explored a series of transformers and deliberation based models to approach cascaded multimodal speech translation as our participation in the How2-based speech translation task of IWSLT 2019. We submitted the `Trans-Attn-Conv` system, which is a canonical transformer with visual atten-

tion over the convolutional features, as our primary system with the remaining ones marked as contrastive ones. The primary system obtained a BLEU of 39.63 on the public IWSLT19 test set, whereas `Trans-Baseline`, the top contrastive system on the same set, achieved 39.85. Our main conclusions are as follows: (i) the visual modality causes varying levels of translation quality damage to the transformers and additive deliberation, but boosts cascade deliberation; (ii) the multimodal transformers and cascade deliberation show performance degradation due to incongruence, but additive deliberation is not as affected; (iii) there is no strict correlation between incongruence sensitivity and translation performance.

6. Acknowledgements

7. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[2] H. Hassan, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, *et al.*, "Achieving human parity on automatic chinese to english news translation," *arXiv preprint arXiv:1803.05567*, 2018.

[3] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: a large-scale dataset for multimodal language understanding," in *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS, 2018. [Online]. Available: http://arxiv.org/abs/1811.00347

[4] Z. Wu, J. Ive, J. Wang, P. Madhyastha, and L. Specia, "Predicting actions to help predict translations," *arXiv preprint arXiv:1908.01665*, 2019.

[5] F. Casacuberta, M. Federico, H. Ney, and E. Vidal, "Recent efforts in spoken language translation," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 80–88, May 2008.

[6] A. Waibel and C. Fugen, "Spoken language translation," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 70–79, 2008.

[7] L. Specia, S. Frank, K. Sima'an, and D. Elliott, "A shared task on multimodal machine translation and crosslingual image description," in *Proceedings of the First Conference on Machine Translation*,

Berlin, Germany, August 2016, pp. 543–553. [Online]. Available: http://www.aclweb.org/anthology/W/W16/W16-2346

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

[9] O. Caglayan, P. Madhyastha, L. Specia, and L. Barrault, "Probing the need for visual context in multimodal machine translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019, pp. 4159–4170. [Online]. Available: https://www.aclweb.org/anthology/N19-1422/

[10] O. Caglayan, W. Aransa, Y. Wang, M. Masana, M. García-Martínez, F. Bougares, L. Barrault, and J. van de Weijer, "Does multimodality help human and machine for translation and image captioning?" in *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 627–633. [Online]. Available: http://www.aclweb.org/anthology/W/W16/W16-2358

[11] J. Libovický and J. 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)*. Association for Computational Linguistics, 2017, pp. 196–202. [Online]. Available: http://aclweb.org/anthology/P17-2031

[12] O. Caglayan, W. Aransa, A. Bardet, M. García-Martínez, F. Bougares, L. Barrault, M. Masana, L. Herranz, and J. van de Weijer, "LIUM-CVC submissions for WMT17 multimodal translation task," in *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 432–439. [Online]. Available: http://www.aclweb.org/anthology/W17-4746.pdf

[13] I. Calixto and Q. Liu, "Incorporating global visual features into attention-based neural machine translation." in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 992–1003. [Online]. Available: https://www.aclweb.org/anthology/D17-1105

[14] P. S. Madhyastha, J. Wang, and L. Specia, "Sheffield multimt: Using object posterior predictions for multimodal machine translation," in *Proceedings of the Second Conference on Machine Translation*, 2017, pp. 470–476.

[15] S.-A. Grnroos, B. Huet, M. Kurimo, J. Laaksonen, B. Merialdo, P. Pham, M. Sjberg, U. Sulubacak, J. Tiedemann, R. Troncy, and R. Vzquez, "The MeMAD submission to the WMT18 multimodal translation task," in *Proceedings of the Third Conference on Machine Translation*. Belgium, Brussels: Association for Computational Linguistics, October 2018, pp. 609–617. [Online]. Available: http://www.aclweb.org/anthology/W18-64066

[16] D. Elliott, "Adversarial evaluation of multimodal machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2974–2978.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 4960–4964.

[19] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014. [Online]. Available: http://arxiv.org/abs/1412.3555

[20] O. Press and L. Wolf, "Using the output embedding to improve language models," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 157–163. [Online]. Available: https://www.aclweb.org/anthology/E17-2025

[21] Y. Xia, F. Tian, L. Wu, J. Lin, T. Qin, N. Yu, and T.-Y. Liu, "Deliberation networks: Sequence generation beyond one-pass decoding," in *Advances in Neural Information Processing Systems*, 2017, pp. 1784–1794.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: http://arxiv.org/abs/1409.0473

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

[24] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet,"

in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[25] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick, *et al.*, "Moments in time dataset: one million videos for event understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–8, 2019.

[26] J. Helcl, J. Libovick, and D. Varis, "CUNI system for the WMT18 multimodal translation task," in *Proceedings of the Third Conference on Machine Translation*. Belgium, Brussels: Association for Computational Linguistics, October 2018, pp. 622–629. [Online]. Available: http://www.aclweb.org/anthology/W18-64068

[27] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ser. ACL '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 177–180. [Online]. Available: http://dl.acm.org/citation.cfm?id=1557769.1557821

[28] R. Sennrich, B. Haddow, and A. 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)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://www.aclweb.org/anthology/P16-1162

[29] T. Kudo, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, October 2018.

[30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[32] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. Gomez, S. Gouws, L. Jones, Ł. Kaiser, N. Kalchbrenner, N. Parmar, R. Sepassi, N. Shazeer, and J. Uszkoreit, "Tensor2Tensor for neural machine translation," in *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*. Boston, MA: Association for Machine Translation in the Americas, Mar. 2018, pp. 193–199. [Online]. Available: https://www.aclweb.org/anthology/W18-1819

[33] J. H. Clark, C. Dyer, A. Lavie, and N. A. Smith, "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*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 176–181. [Online]. Available: https://www.aclweb.org/anthology/P11-2031