AFRL-AFOSR-UK-TR-2023-0034

Referential Grounding in Multimodal Machine Translation

Specia, Lucia
IMPERIAL COLLEGE OF SCIENCE TECHNOLOGY & MEDICINE
EXHIBITION RD
LONDON, ,
GB

12/22/2022
Final Technical Report

DISTRIBUTION A: Distribution approved for public release.

Air Force Research Laboratory
Air Force Office of Scientific Research
European Office of Aerospace Research and Development
Unit 4515 Box 14, APO AE 09421

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20221222	Final	START DATE 20191115	END DATE 20220514

4. TITLE AND SUBTITLE
Referential Grounding in Multimodal Machine Translation

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA8655-20-1-7006	61102F
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Lucia Specia

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
IMPERIAL COLLEGE OF SCIENCE TECHNOLOGY & MEDICINE
EXHIBITION RD
LONDON
GB

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
EOARD UNIT 4515 APO AE 09421-4515	AFRL/AFOSR IOE	AFRL-AFOSR-UK-TR-2023-0034

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project aimed to advance the state of the art in multimodal machine translation (MMT). MMT is an area where a text in the source language is supplemented by visual information (images or video) to be used as additional context to better understand and translate the text into a target language. The core of the advances proposed are on referential grounding, i.e., on guiding the alignment between image regions and source (and/or target) words such that the visual context can be more useful for translation.
Work done during the project in covered the following directions:
1. Improving supervised attention mechanisms to map source or target words to image regions, addressing both attention at encoding time (i.e. learning alignments between source words and objects in the image) and at decoding time (i.e. learning alignments between target words and objects in the image), as well as improving the underlying multimodal neural machine translation architectures and fusion strategies to use such information and exploring more recent and better types of visual features.
2. Leveraging information from multiple vision-and-language tasks and datasets to improve multilingual grounding.
3. Creating resources to facilitate work on referential grounding.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	SAR	2

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
NANDINI IYER	314-235-6161

Referential grounding in multimodal machine translation
FA8655-20-1-7006

Prof Lucia Specia, Imperial College London

Project start 15th Nov 2019
Current project ending 14 May 2022
Reporting period: Jan 2022 – May 2022

Progress report

This project aimed to advance the state of the art in multimodal machine translation (MMT). MMT is an area where a text in the source language is supplemented by visual information (images or video) to be used as additional context to better understand and translate the text into a target language. The core of the advances proposed are on **referential grounding**, i.e., on guiding the alignment between image regions and source (and/or target) words such that the visual context can be more useful for translation.

Work done during the project in covered the following directions:

1. Improving supervised attention mechanisms to map source or target words to image regions, addressing both attention at encoding time (i.e. learning alignments between source words and objects in the image) and at decoding time (i.e. learning alignments between target words and objects in the image), as well as improving the underlying multimodal neural machine translation architectures and fusion strategies to use such information and exploring more recent and better types of visual features.
2. Leveraging information from multiple vision-and-language tasks and datasets to improve multilingual grounding.
3. Creating resources to facilitate work on referential grounding.

This report focuses in the last 4 months of the project, and covers further work towards **direction 1**, namely we proposed the first approach to **Simultaneous Video Translation**, i.e. real-time translation or interpreting where a translation needs to be generated for incomplete source sentences and where a video is available as additional context. An example of application is the translation of the audio stream in live broadcasting such as news. Different from our previous work, where a single image was available as static visual information for each text segment to translate, in our more recent work a video containing multiple pieces of visual information is available for each text segment. This brings many challenges to MMT, including decisions around how to process the video (frame sampling approach, video encoding approach) as well as how to combine multiple pieces of visual (frames or even image regions in frames) and textual (source and/or target words) information. The latter can be seen as a form of **referential grounding** between frames and text sub-segments.

Using videos as visual information for MMT is appealing as if offers richer visual context, especially for longer text segments. It also opens new avenues for research on referential grounding: for a correctly grounded translation, the model needs to identify correspondences between specific video frames or parts of frames that are relevant for the words seen so far in the incomplete textual input that is incrementally made available. A draft summary of the work done is attached to this report (paper to be submitted). In this paper, we use a dataset of videos where people describe their apartments for rental to train and evaluate our simultaneous video translation models.

DISTRIBUTION A: Distribution approved for public release. Distribution unlimited

Publications relevant to reporting period

Wang, Y., Miao, Y., Specia, Lucia. (2022). Translate in the Scenes: Multi-modal Simultaneous Machine Translation (draft of paper to be submitted, attached).

Other cited publications (previous report)

Haralampieva, V., Caglayan, O., Specia, L. (2022). Supervised Visual Attention for Simultaneous Multimodal Machine Translation. Journal of Artificial Intelligence Research (JAIR), 74:1059-1089.

Wang, J., P.S., Figueiredo, J.M., Specia, L. (2022). MultiSubs: A Large-scale Multimodal and Multilingual Dataset. Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), p. 6776–6785

Mitzalis, F., Caglayan, O., Madhyastha, P.S., Specia, L. (2021). BERTGEN: Multi-task Generation through BERT. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021), p. 6440–6455.

Caglayan, O., Ive, J., Haralampieva, V., Madhyastha, P.S., Barrault, L., & Specia, L. (2020). Simultaneous Machine Translation with Visual Context. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), p. 2350-2361.

Specia, L., Wang, J., Lee, S.J., Ostapenko, A., Madhyastha, P. (2021). Read, Spot and Translate. Machine Translation Journal, 35, 145–165, Springer.

DISTRIBUTION A: Distribution approved for public release. Distribution unlimited

Translate in the Scenes: Multi-modal Simultaneous Machine Translation

Anonymous ACL submission

Abstract

We propose the task of simultaneous video translation, which aims to explore multimodal information to help produce machine translations for videos in real-time. We posit that visual information can help the language channel achieve a good trade-off between translation quality and latency in this very challenging task. To enable research in this area, we created the *ApartmenTour Simultaneous Video Translation Dataset*, which contains 33495 parallel subtitles for 82 hours of videos describing apartments for rental. We then build on previous approaches that address simultaneous translation with multimodal information from static images using RNN and Transformer architectures and show that these lead to considerable improvements compared to text-only translation models. Finally, we propose a frame selection approach to efficiently deal with the problem of processing multiple frames for one sentence.

1 Introduction

Given the recent breakthroughs in machine translation (MT) with neural approaches (Bahdanau et al., 2014; Vaswani et al., 2017) and the impressive improvements in translation quality observed in the last 6-7 years, the community has turned to a number of practical applications, including *simultaneous* MT. This task simulates the job of a human interpreter, who translates simultaneously as source words are produced, in real-time. This has a number of applications such as international summits, conferences and lectures. Compared to the traditional *consecutive* MT, where the entire sentence is available as the input, simultaneous MT (SiMT) is much more challenging, given that the incomplete source language content usually reduces the quality of the translation significantly. In addition, the model is required to consider not only the quality of the translation but also the latency of the reader, i.e. the time the reader has to wait to receive the translation before it is produced. A good trade-off must be obtained between the quality and the latency.

Cho and Esipova (2016) first explored the possibility of SiMT with a heuristic decoding algorithm. After that, several methods were proposed that focus on finding the best timing to perform the translation action through supervised or unsupervised neural networks (Gu et al., 2016; Zheng et al., 2019), or fixing the latency to optimise for translation quality (Ma et al., 2018).

Hoping to utilise additional modalities to compensate for the incomplete semantics during the simultaneous translation, previous work has also proposed simultaneous image translation (Caglayan et al., 2020; Imankulova et al., 2020). Previous work has shown that the visual information from an image can help the model anticipate the missing source context as well as disambiguate source terms, overall helping predict the next source word in the pipeline of simultaneous translation. However, existing work has been limited to static visual input (one image per source sentence) and to datasets where the text is a description of the image (i.e. the Multi30K dataset (Elliott et al., 2016)), where the visual information is clearly well aligned to the source sentence, and can be seen as an alternative representation of the information given by the complete source sentence. When made available to the model at the first timestep of encoding/decoding, the complete visual representation information is therefore bound to provide helpful information to complement the incomplete source context. No previous work has been done in the more realistic, more practical but also more challenging setting addressed in this paper: that of SiMT from videos.

In the simultaneous video translation task, we aim to simultaneously generate the translation using as input the incomplete source text and the

DISTRIBUTION A: Distribution approved for public release. Distribution unlimited

corresponding *partial* visual information, i.e. only the frames visible thus far. The long sequences of the videos bring several complexities to the task. First, there are multiple frames for each sentence and the frames will appear gradually as a stream, i.e. both video and text modalities are simultaneous, requiring different architectures. Second, the text in the subtitles of the videos are not always aligned to the visual content, so there is a weak correspondence between the text and the video which can introduce noise in the process and reduce the translation quality.

To address these challenges, we start by collecting a new dataset – *ApartmenTour Simultaneous Video Translation Dataset*, which includes 534 apartment tour videos, accompanied by English subtitles with timestamps and parallel Chinese translations. We then build the Multi-modal Simultaneous Machine Translation model which can process multiple frames. We exploit different visual features to represent the videos, as well as different fusion methods to fuse two modalities. To tackle the weak semantic equivalence between the content in the text and the video, we propose two masking strategies to mask frames which we expect to contain information that is not related to the text. The experimental results show that our approaches to SiMT from videos perform better than their monomodal counterparts without the video. The masking strategies have proved to filter out some irrelevant frames and improve translation quality.

In Section 2, we provide the background of the monomodal and multimodal simultaneous machine translation. Section 3 outlines the processing of our ApartmenTour Simultaneous Video Translation Dataset. Our proposed Multi-modal Simultaneous Machine Translation model is described in Section 4. Section 5 provides the results of our models and the qualitative analysis. Section 6 is for the conclusion.

2 Related Work

2.1 Simultaneous machine translation

Simultaneous machine translation is challenging as machines have to translate the sentences simultaneously as the words appear. Partial sentences will inevitably impair the quality of the translation compared to typical machine translation task which translates the entire sentences. In addition, both the quality and the latency of the translation have to be considered.

In the earlier years, segmentation-based methods have been proposed. These methods (Fügen et al., 2007; Bangalore et al., 2012; Sridhar et al., 2013; Oda et al., 2014) aim to segment a whole source sentence, translate each part independently, and concatenate all parts to form a whole target sentence to simulate the simultaneous translation scenario.

Cho and Esipova (2016) start to introduce the neural network into simultaneous machine translation to exploit the dependency between different parts of the translation. They use an rnn-based machine translation model trained on full sentences and connected it with a greedy decoding part to do the simultaneous machine translation. They defined two criteria, wait-if-worse and wait-if-diff to heuristically decide the timing to write a target word. However, this method is based on predefined rules, which cannot greatly exploit the information of hidden states of the recurrent neural networks.

Gu et al. (2016) propose a reinforcement learning-based method to train an agent to generate the READ/WRITE action sequences. The agent is trained using the reward function which considers both the quality and the latency.

As reinforcement learning-based methods are not stable and difficult to train, Ma et al. (2018) propose the wait-k policy, which forces the translation (the first WRITE action) to start after reading k source words, and then performs the READ/WRITE actions alternatively. The policy enables the model to have the ability to anticipate the future words.

However, as the latency for the wait-k method is predefined, it cannot learn the best timing for READ/WRITE actions for different data. In addition, the anticipation of the future words can often be incorrect. Zheng et al. (2019) propose a method that combines the adaptivity of the reinforcement learning-based method and the easy training properties of the predefined latency methods. They design a supervised action sequence generation network, which takes in the ground truth action sequences generated using a heuristic algorithm.

2.2 Multimodal Simultaneous Translation

Multimodal simultaneous translation is a newly proposed task that seeks to introduce another modality to compensate for the miss semantic content.

DISTRIBUTION A: Distribution approved for public release. Distribution unlimited

Imankulova et al. (2020) propose the RNN-based Multimodal Simultaneous Neural Machine Translation model (MSNMT) by combining the wait-k simultaneous translation model (Ma et al., 2018) and a multimodal model (Libovický and Helcl, 2017). A hierarchical attention mechanism (Libovický and Helcl, 2017) is adopted to fuse the text modality and image modality in the decoder side. Based on the two-GRU-layer architecture described in (Caglayan et al., 2016), Caglayan et al. (2020) introduce object classification features extracted from ResNet50 (He et al., 2016) and object detection features extracted from the "buttom-up-top-down(BUTD)" (Anderson et al., 2018) extractor. Also, they exploit encoder attention and decoder attention mechanisms which can only attend to the available hidden states to simulate the simultaneous translation setting. Ive et al. (2021) propose the first RL-based multimodal simultaneous translation model to introduce the visual cues to the agent and environment. Both the global image classification features (He et al., 2016) and the local visual concept features are explored. The research on multimodal simultaneous translation is mostly based on static images but not videos.

3 ApartmenTour Dataset

We pick up 534 apartment tour videos from the ApartmenTour dataset (Zhong et al., 2020) and extracted WAV files using FFmpeg[1]. Next, the English transcription (splitted by paragraphs) with punctuation and the corresponding timestamps for each video are extracted using Microsoft Azure Speech Translation service[2] by uploading the WAV files (as shown in Figure 1). The timestamp for each paragraph includes the offset and the duration in the unit of 100 nanoseconds (1 nanosecond = 1×10^{-9} second). We translated the English transcription to Chinese using the Youdao Translation API[3][4]. As most paragraphs of transcription generated by the Microsoft Azure Speech Translation service are very long and contain many sentences, we split each paragraph by the period mark to get several sentences. In addition, since the subtitles are converted from audio, they contain many hesitations and other symbols that are not very meaningful and sometimes correspond to a whole sentence, such as "So." and "Yeah.". To focus on more meaningful and hard to translate sentences, we remove those sentences whose word lengths are less than eight words. As a result, our final dataset includes 20,974 sentences for training, 6,521 sentences for validation, and 6,000 sentences for testing. The average length of the sentences is around 20 words. Some statistics of the dataset are shown in Table 1.

	Train	Val	Test
Sentence / Video clip	20974	6521	6000
English words	426521	129649	120872
English vocabulary	16691 in total		
Chinese words	387089	117270	109500
Chinese vocabulary	19928 in total		

Table 1: Statistics about the ApartmenTour Simultaneous Video Translation Dataset

Since using all frames from the video would make the process very inefficient, we resort to frame sampling, a common practice in video processing (Lei et al., 2021). For convenience, we sample one frame every five words (as shown in Figure 2). The frame at timestep 0 is always sampled as the first frame for the corresponding sentence. As we only get sentence-level timestamps using Microsoft Azure Speech Translation service, we roughly compute timestamp for each frame using offset $+ (\frac{i \times 5}{\text{word length}}) \times$ duration for the i-th frame (starting from 0).

4 Models

As shown in Figure 3, our model consists of three parts: language channel, video channel and the translation networks.

4.1 Language channel

The architecture is the same as that in Vaswani et al. (2017), but using a pre-norm variant. In the simultaneous translation task, we can only use the unidirectional information, so we will generate a lookahead mask to make each word in the sentence not attend to the words behind itself in the self-attention layers. For simultaneous translation, we adopt the wait-k policy proposed by Ma et al. (2018). Though the authors report better results using training time wait-k, that is, training and testing on partial sentences, we find that testing time wait-k works better on our dataset (this was also re-

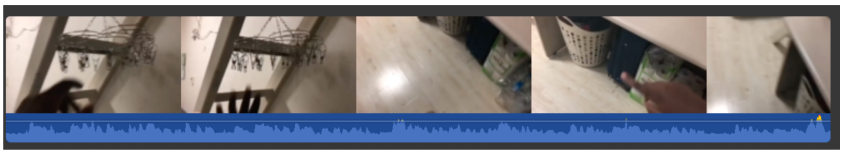
[1] http://ffmpeg.org/

[2] https://azure.microsoft.com/en-us/services/cognitive-services/speech-translation/

[3] https://ai.youdao.com/product-fanyi-text.s

[4] We also created German translation using Azure, which we did not use in our experiments but will release with the data.

DISTRIBUTION A: Distribution approved for public release. Distribution unlimited



Timestamp	English	Chinese
[2526800000, 226800000] offset duration	Only a washer, so I hang up my clothes right here to dry on this little clothespin thing. And then I also kind of lay things out on here as well. And then I also use it for some storage like I keep my water and my paper towels and toilet paper. I've got like a carry on bag. I've got my dirty clothes hamper. I've got a bunch of stuff guys.	只有洗衣机，所以我把衣服挂在这里，用这个小晾衣夹晾干。然后我也会把东西放在这里。我也用它来储存水、纸巾和厕纸。我有个随身包。我有脏衣服篮。我有一堆东西，伙计们。

Figure 1: An illustration of the generation of the source and translation text. The WAV file of the video was extracted and uploaded to the Microsoft Azure speech translation service to generate the English subtitles, Chinese translation and timestamps. The timestamp for each paragraph includes the offset and the duration in the unit of 100 nanoseconds (1 nanosecond = 1×10^{-9} second)

Figure 2: An illustration of the frame sampling. One frame is sampled for every five words. The frame at timestep 0 is also sampled.

ported to be the case in (Caglayan et al., 2020)). For the language channel, we train using the full sentences and test using partial sentences. To be more specific, only the source text feature of $t+k$ source words at test time can be seen when generating the t-th target word. We also apply the GRU-based simultaneous translation method (Caglayan et al., 2020) in the experiment section.

4.2 Video channel

In what follows we describe what variants of visual features we extract from the sampled frames, as well as the strategy we propose to mask out irrelevant frames.

Visual features We adopt the Vision-and-Language Transformer (ViLT) (Kim et al., 2021) to create a masked language model for each frame sampled from the video and get the text feature for 8 words (nouns and adjectives describing the image) from one frame, and the shape is [8, 768] for each frame. The features for the same sentence but different frames are stacked together and then padded to get a feature matrix of shape [176, 768] for each sentence (as the maximum number of frames for one sentence is 22, the maximum number of text features is 22 * 8 = 176).

Masking strategies As with any video dataset, our data will contain many frames that are irrelevant to what the Youtubers are talking about, and thus would represent noise rather than helpful information to improve SiMT. Here we propose two strategies to mask the irrelevant frames and reduce the impact of noise:

In the apartment tour videos, there are quite a lot of clips where the Youtubers face the camera and talk about some background knowledge of filming the videos. The frames sampled from these clips are always irrelevant to the corresponding caption and contain very little information. By checking the dataset, we find that there are a large number of frames in which the Youtuber's face takes up most of the place, making up more than 5% of the frame. We here roughly define these frames as "irrelevant frames". As shown in Figure 4, the left frame is our irrelevant frame and the middle frame is considered as a relevant frame as there is no face appearing in the frame. The Youtuber's face also appears in the right frame, but as it does not take up more than 5% of the frame, we still consider it a relevant frame. Haar Cascades in OpenCV[5] is used to do face detection in every frame and mask the frame

[5] https://docs.opencv.org/3.4/d2/d99/tutorial_js_face_detection.html

DISTRIBUTION A: Distribution approved for public release. Distribution unlimited

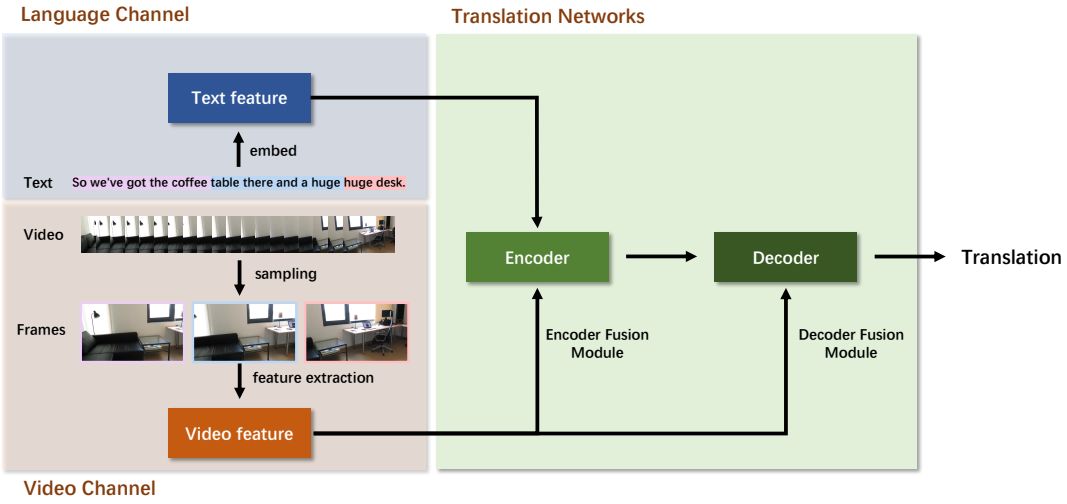


Figure 3: Multi-modal Simultaneous Machine Translation model which consists of language channel, video channel and the translation networks. Language channel is for text processing and video channel is for visual feature extraction. In the translation networks, two modalities are fused in the decoder side or the encoder side.

Figure 4: An example of "irrelevant" and "relevant" frames.

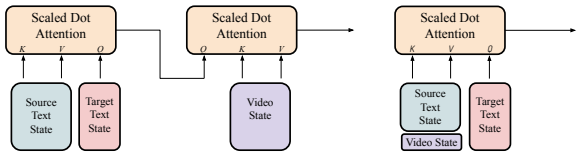


Figure 5: An illustration of the serial attention layer (left) and the concatenation attention layer (right)

if the face region takes up more than 5%, so the model will not utilise the feature of this frame.

We also use the Universal Sentence Encoder (Cer et al., 2018) to embed the subtitle (the source text) and the caption of the corresponding frame, then get the semantic textual similarity by calculating the inner product of the encodings. If the similarity between the subtitle and the frame's caption is lower than a threshold, this frame will be masked as it is not so related to the source text.

4.3 Fusion in the translation networks

In what follows, we describe our fusion strategies to build a multimodal model from the language and video channel representations.

4.3.1 Decoder Fusion

Two modalities will be fused in the cross attention layers in the decoder. We try two methods for the cross-attention layer: serial attention and concatenation attention (shown in Figure 5).

The serial attention layer includes two scaled dot attention layers. The first scaled dot attention layer will attend the query (the target state) to the source text state (the key and value are both the text state). Then the second scaled dot attention layer will attend the output from the first scaled dot attention layer to the visual state (the query is the output from the last attention layer, the key and value are both the visual state). The output from the second attention layer will be returned. As for the concatenation attention layer, the source text state and the visual state are concatenated together (set as key and value). The attention layer attends the target state (set as the query) to the concatenated state to get the context vector and return it.

Also, we do some modifications to do the simultaneous machine translation for the Transformer model. The number of the frames that are available at timestep t is computed by $N = \text{math.ceil}((t + \delta)/5 + 0.0001)$, and the index for the latest frame is computed by $i = N - 1$ ($\delta \leq k$ to guarantee that the model will not access to the future image information). The frames$[i-1]$ and frames$[i]$ are the visible frames when translating the t-th word (when i is 0, only the frames$[i]$ is visible). During the training, all the source words together with the

DISTRIBUTION A: Distribution approved for public release. Distribution unlimited

		BLEU	METEOR	AL
w1	Baseline	20.982	25.534	1.544
	MSMT	21.541	25.997	1.552
w2	Baseline	22.849	26.924	2.224
	MSMT	23.123	27.047	2.180
w3	Baseline	24.708	27.853	2.904
	MSMT	25.069	28.142	2.896

Table 2: The results of the baseline Monomodal Transformer model and our proposed Multi-modal Simultaneous Machine Translation model (MSMT) from wait-1 to wait-3.

frames$[i-1]$ and frames$[i]$ will be visible. Only the $t+k$ words together with the frames$[i-1]$ and frames$[i]$ will be visible when generating the t-th word at test time.

4.3.2 Encoder Fusion

As it is an encoder side fusion method, the visual feature will be introduced earlier. Merged attention (Chen et al., 2020; Hendricks et al., 2021) is adopted to concatenate the source text feature and the visual feature together and pass the merged feature through the stacked self-attention layers and feed-forward layers. The scaled dot attention layer is used for the decoder to attend the target state to the hidden state got from the encoder (set the target state as query, the hidden state as key and value).

5 Experiments

5.1 Multimodal Simultaneous Machine Translation

In table 2, we compare our Multi-modal Simultaneous Machine Translation model (decoder side fusion + serial attention layer + ViLT feature + encoder mask) to the baseline. When doing simultaneous translation from wait-1 to wait-3, introducing the ViLT feature with encoder mask can improve the translation quality without sacrificing the simultaneity of the translation (the Average Lagging of the multimodal model is even a little bit smaller than the one of the monomodal model for wait-2 and wait-3 tasks).

5.2 Ablation study

5.2.1 Architecture

As visual information can play a more important role when k is smaller (Caglayan et al., 2020), here we do an early feature and architecture search on wait-1 to decide the best practice. The feature is fixed as the ResNet50 feature to search for the best architecture. The results are shown in Table 3.

		BLEU	
Baseline	RNN	16.073	
	Transformer	20.982	
MSMT		No mask	OpenCV mask
	RNN Concat (decoder fusion)	18.294	-
	Transformer Merged (encoder fusion)	20.378	21.086
	Transformer Concat (decoder fusion)	20.939	20.678
	Transformer Serial (decoder fusion)	**21.276**	**21.239**

Table 3: Architecture search. We fix the feature and experiment across different architectures on wait-1 to find the best architecture.

The Transformer baseline model is a lot better than the RNN baseline, improving the BLEU score by around 5 points. The transformer can greatly learn the contextual information using self-attention layers.

From Table 3, when using decoder side fusion with serial attention layer, the Transformer-based model can get the best BLEU scores (the bold numbers in Table 3). It shows that an extra scaled dot layer attending the text hidden state to the visual hidden state is more effective than directly concatenating these two states together. Though better improvement in the image-guided translation has been shown using the encoder side fusion, it is not the case for our task here. There is only one image for one sentence for the image-guided translation task, and the image is manually picked to guarantee that it is very relevant to the text. But in our task, there are multiple images for one sentence, and the images are extracted without the manually further process. So encoder side fusion will introduce a lot of noise when generating the hidden state. When the masking strategy is applied to mask the irrelevant frames, there is an improvement in the encoder side fusion model compared to the model using the raw features, which can somehow confirm that irrelevant frames will deteriorate the translation quality.

5.2.2 Feature & Masking strategy

Table 4 shows that, for wait-1 and wait-2 simultaneous translation tasks, introducing different kinds of no mask image features can all bring improvement to the translation quality. However, using OpenCV mask will lead to worse results for the ResNet feature and the image captioning feature than those using raw image features. Also, the encoder mask does not improve the translation qual-

DISTRIBUTION A: Distribution approved for public release. Distribution unlimited

		BLEU		
	Baseline	20.964		
		No mask	OpenCV mask	Encoder mask
w1	MSMT w/ ResNet	21.276	21.239	21.302
	MSMT w/ Image cap.	21.404	21.035	21.097
	MSMT w/ ViLT	21.408	21.306	21.541
	Baseline	22.849		
		No mask	OpenCV mask	Encoder mask
w2	MSMT w/ ResNet	23.122	22.992	23.134
	MSMT w/ Image cap.	23.285	22.760	23.010
	MSMT w/ ViLT	23.085	23.197	23.123
	Baseline	24.708		
		No mask	OpenCV mask	Encoder mask
w3	MSMT w/ ResNet	24.439	24.479	24.520
	MSMT w/ Image cap.	24.376	24.465	24.866
	MSMT w/ ViLT	24.698	24.960	25.069

Table 4: The results of the baseline Transformer model and the multimodal Transformer model with the serial attention layer and different features from wait-1 to wait-3.

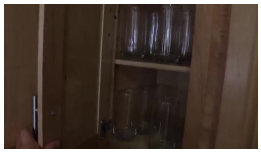
ity too much for wait-1 and wait-2. But when it comes to wait-3, the masking strategy is important. From Table 4, using features without masking on the wait-3 task makes the translation quality even worse than the baseline model's translation quality. For example, we get 24.376 using the image captioning feature, which is worse than the baseline 24.708. With masking strategy (especially the encoder masking strategy), we can get the translations of higher quality. That is because, as k increases, the language channel can get more source words which can directly improve the translation quality while most kinds of image features are comparatively noisy. When the monomodal baseline can get quite a great result when k is large (the language channel is more powerful), the introduction of the noisy video modality will affect more and impair the translation quality. However, when k is smaller, the importance of exposing to more video information to help anticipate overweighs the importance of reducing noise.

The feature and the masking strategy selected will both affect the translation quality. Encoder masking strategy is better than the OpenCV mask strategy as it can rule out more frames irrelevant to the text. Taking all three tasks (wait-1 to wait-3) into consideration, using ViLT feature with the encoder mask is the best practice.

5.3 Qualitative Analysis

5.3.1 Disambiguation

As we all know, polysemy is a difficulty in machine translation tasks. It will be challenging for the translation model to determine the correct meaning in different sentences. Using the image or the video can help disambiguate the polysemous words by providing additional cues.



Source: I just have **glasses** that I literally never use .
Target: 我 (I) 只是 (just) 有 (have) 我 (I) 从不 (literally never) 使用 (use) 的 **杯子 (drinking glasses)** 。
Baseline: 我 (I) 只 (just) 有 (have) 一个 **眼镜 (eyeglasses)** ，我 (I) 从来 不 (literally never) 用 (use) 。
MSMT: 我 (I) 只 (just) 有 (have) **杯子 (drinking glasses)** ，我 (I) 从来 不 (literally never) 用 (use) 。

Figure 6: Multimodality can help disambiguate the polysemous word. In this example, the multimodal model can utilise the visual information to correctly translate "glasses" to "drinking glasses" in Chinese while the monomodal one translates it to "eyeglasses".

The word "glasses" has two common definitions(GLA): (1) a device for correcting vision, which is also known as "eyeglasses". (2) the plural form of the container for drinks made of glass. From Figure 6, when using the monomodal translation model, the word "glasses" in the sentence "I just have glasses that I literally never use." was incorrectly translated as "眼镜" (the first meaning we specified), but when using the MSMT, the corresponding frame provides the information about the "glasses" which is a container, and the model successfully disambiguate the word and generate correct translation "杯子" (the second meaning we specified).

5.3.2 Extra visual information to anticipate

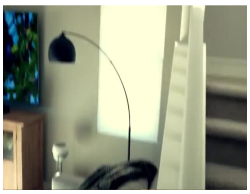
This aspect is more related to the simultaneous translation task. As we adopt the wait-k policy to control the latency, some WRITE actions will be executed before getting enough context from the text, leading to the inaccurate translation. However, an extra modality can provide the context required to translate some specific words.

Compound word As this task is a simultaneous task, every step is simultaneous, including the speech recognition. The words in the source language are generated one after another, so there is always the case that the translation model will only get one part of a compound word and translate this part directly. As shown in Figure 7, when using the wait-1 policy, the monomodal model fails to translate the compound word as a whole but only generate the translation for "gas" ("气体"), while MSMT successfully "anticipate" the following word using the extra information from the frame to generate

DISTRIBUTION A: Distribution approved for public release. Distribution unlimited

Source: I used to have like a gas stove (煤气灶) .
Baseline (w1): 我 以前 以前 已经 有 一个 气体 (gas) 。
MSMT (w1): 我 以前 以前 有 过 一个 煤气灶 (gas stove) 。
Target: (consecutive): 我 以前 有 个 煤气炉 (gas stove) 。

Figure 7: Multimodality can help predict the following word during the simultaneous translation. In this example, MSMT can correctly predict that the word “gas" would be followed by the word “stove" according to the visual information, and generate the correct translation for the compound word “gas stove", while the monomodal model directly translates the “gas".



Source: By the way , check out this crazy lamp .
Baseline (w2): 顺便 说 一下 , 看看 这个 疯狂 的 灯 。
MSMT (w2): 顺便 说 一下 , 看看 这 盏 疯狂 的 灯 。
Baseline (w3): 顺便 说 一下 , 看看 这个 疯狂 的 灯 。
MSMT (w3): 顺便 说 一下 , 看看 这 盏 疯狂 的 灯 。
Target: (consecutive): 顺便 说 一句 , 看看 这 盏 疯狂 的 灯 。

Figure 8: Multimodality can help generate the accurate measure word. In this example, using the extra information from the video frame, MSMT successfully generates the correct measure word/classifier for the word “lamp".

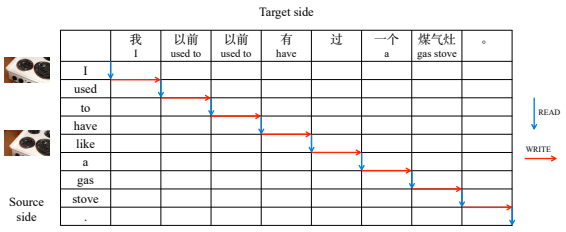


Figure 9: The process of the wait-1 multimodal simultaneous translation.

the correct translation “煤气灶 (gas stove)". From Figure 9, MSMT successfully anticipate and generate the correct translation “煤气灶 (gas stove)" before reading the word “stove".

Measure word This point is a more language-specific one. In Chinese, a measure word, also known as a classifier, must be used together with the numerals or the demonstrative before the nouns (Li and Thompson, 1989). Measure words should be chosen according to the noun. For example, for the noun “pencil", the commonly used measure word is “支" and for the noun “table” it should be “张". In a consecutive translation task, it will be quite easy to figure out the correct measure word using the context. However, in the simultaneous translation task, it will be more difficult due to the incomplete semantics. For example, as shown in 8, when using wait-2, when translating something like “this crazy lamp", the model can only see the English word “this" without knowing the noun after it. Model without visual information might directly translate the English word “this" to “这个", where “个" is not the appropriate measure word for “lamp". However, with the visual information, MSMT can get the accurate measure word “盏". For wait-3 model, we can also see that the visual information help accurately generate the measure word.

6 Conclusion

In this paper, we have explored the newly proposed Simultaneous Video Translation task. We have first created the *ApartmenTour Simultaneous Video Translation Dataset* specifically for this task, which consists of descriptive apartment tour videos and the corresponding multi-language captions with timestamps. Secondly, we have designed the Multi-modal Simultaneous Machine Translation model to process the videos. We have experimented on different fusion timings (encoder side fusion and decode side fusion), cross attention layers (concatenated, serial and merged), features (ResNet50 feature, image captioning feature and ViLT feature). Furthermore, we have proposed the preliminary idea of the masking strategy which can help exclude some noises for the visual space features. Experiment results have shown that our Multi-modal Simultaneous Machine Translation model can gain improvement over the monomodal one and at the same time keep the same latency level. Finally, we have found that when using the Transformer, decoder side fusion with serial cross attention layer and the ViLT feature with encoder mask, we can obtain the greatest translation quality improvement.

DISTRIBUTION A: Distribution approved for public release. Distribution unlimited

References

Glasses | meaning in the cambridge english dictionary. https://dictionary.cambridge.org/dictionary/english/glasses. (Accessed on 08/25/2021).

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

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

Srinivas Bangalore, Vivek Kumar Rangarajan Sridhar, Prakash Kolan, Ladan Golipour, and Aura Jimenez. 2012. Real-time incremental speech-to-speech translation of dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 437–445.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost Van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? *arXiv preprint arXiv:1605.09186*.

Ozan Caglayan, Julia Ive, Veneta Haralampieva, Pranava Madhyastha, Loïc Barrault, and Lucia Specia. 2020. Simultaneous machine translation with visual context. *arXiv preprint arXiv:2009.07310*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *arXiv preprint arXiv:1606.02012*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.

Christian Fügen, Alex Waibel, and Muntsin Kolss. 2007. Simultaneous translation of lectures and speeches. *Machine translation*, 21(4):209–252.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK Li. 2016. Learning to translate in real-time with neural machine translation. *arXiv preprint arXiv:1610.00388*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *arXiv preprint arXiv:2102.00529*.

Aizhan Imankulova, Masahiro Kaneko, Tosho Hirasawa, and Mamoru Komachi. 2020. Towards multimodal simultaneous neural machine translation. *arXiv preprint arXiv:2004.03180*.

Julia Ive, Andy Mingren Li, Yishu Miao, Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2021. Exploiting multimodal reinforcement learning for simultaneous machine translation. In *EACL*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*.

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

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341.

Charles N Li and Sandra A Thompson. 1989. *Mandarin Chinese: A functional reference grammar*, volume 3. Univ of California Press.

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

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 *European conference on computer vision*, pages 740–755. Springer.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, et al. 2018. Stacl: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. *arXiv preprint arXiv:1810.08398*.

DISTRIBUTION A: Distribution approved for public release. Distribution unlimited

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Optimizing segmentation strategies for simultaneous speech translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 551–556.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan. 2013. Segmentation strategies for streaming speech translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 230–238.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019. Simpler and faster learning of adaptive policies for simultaneous translation. *arXiv preprint arXiv:1909.01559*.

Yujie Zhong, Linhai Xie, Sen Wang, Lucia Specia, and Yishu Miao. 2020. Watch and learn: Mapping language and noisy real-world videos with self-supervision. *arXiv preprint arXiv:2011.09634*.

A Visual Feature extraction

ResNet50 object classification features We resize each image to 224x224 and input the image to the ResNet50 (He et al., 2016) network pretrained on the ImageNet database (Deng et al., 2009) to get the [1, 1000] top layer feature, where each element is the score for one of 1000 object categories. As there are several images for one sentence, we concatenate the features of multiple images for the same sentence by the first axis and pad to a feature matrix of shape [22, 1000] as the maximum length in the corpus will not exceed 105 (the corresponding to 22 frames being sampled).

Image captioning features We use the image captioning network proposed by Xu et al. (2015) pretrained on Microsoft COCO dataset (Lin et al., 2014) to extract features. The flattened feature from the last layer of the encoder of the image captioning network for each sampled frame is collected. We expect that the image captioning model can anticipate the corresponding sentence for one image and generate feature which is more suitable in the simultaneous video translation task.

B Experimental Setting

We use the pysimt framework[6] for our experiments. As for the evaluation metrics, BLEU score (Papineni et al., 2002) is used for the quality evaluation and Average Lagging (Ma et al., 2018) for latency measuring. For the RNN model, we run each experiment with the max epoch set to 100 and an early stop criterion: when there is no improvement for 10 epochs, the experiment will stop. The initial learning rate is 0.0004, and if the BLEU score on the validation set does not improve for two epochs, the learning rate will decay by multiplying 0.5, and the minimum learning rate is 1e-6. Adam (Kingma and Ba, 2014) optimiser is adopted, and the batch size is set to 64. We use the ResNet50 feature as the visual feature for the RNN model. As we pay more attention to the Transformer architecture, we do not try too many different kinds of features for the RNN model, just to see whether the introduction of the video can improve the quality of the simultaneous translation.

For the Transformer model, we run each experiment with 350 epochs. As specified in (Vaswani et al., 2017), we use the Noam learning rate scheduler, setting the initial learning rate to 0.2, warmup

[6] https://github.com/ImperialNLP/pysimt

DISTRIBUTION A: Distribution approved for public release. Distribution unlimited

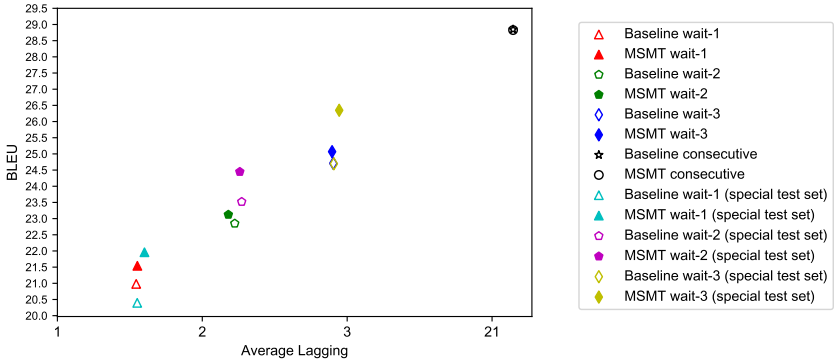


Figure 10: BLEU scores versus Average Lagging for the baseline monomodal model and the multimodal model on two different test sets.

step to 4000 and model dimension to 512. Each experiment is trained on one GeForce GTX 1080 Ti GPU and takes around 7 hours to finish. The number of learnable parameters is around 70M.

We train the models on the training set and record the checkpoints which can get the highest BLEU scores on the validation set, then evaluate these models on the test set. We run experiments for different settings three times and calculate the average.

C Special test set

We create a special test set in which the sentences are all highly related to the frames' content, and all have the pattern 'a/this + noun' (measure words are needed in the Chinese translation) to show that the visual information can better improve the translation quality of these sentences. The results are shown in Figure 10.

As we can see from Figure 10, from wait-1 to wait-3, the improvements of the Multi-modal Simultaneous Machine Translation model (MSMT) on the special test set are more significant than the improvements on the original test set. It shows that when the frames are highly related to the text, the visual cues can better help anticipate to improve the translation quality of the sentences with the pattern 'a/this + noun' (measure words are needed in the Chinese translation). For wait-3, though the language channel will get more source words and be more powerful, the visual information still helps increase the BLEU score by around 1.6 on the special test set.

D Ethical Consideration

One consideration is the collecting and using of the data. The videos in the *ApartmenTour Simultaneous Video Translation Dataset* are all public on Youtube. When uploading a public video to Youtube, the Youtuber grants each other user to use the content. All the videos collected are under fair use. Microsoft Azure Service used for generating transcripts follows the GDPR. Only the transcripts, translations, corresponding timestamp, the packed image features and the public Youtube video id of the videos in the dataset will be released. The pysimt framework used in this paper is under MIT License.

DISTRIBUTION A: Distribution approved for public release. Distribution unlimited