Towards Multimodal Simultaneous Neural Machine Translation

Aizhan Imankulova* **Masahiro Kaneko*** **Tosho Hirasawa*** **Mamoru Komachi**
Tokyo Metropolitan University
6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
{imankulova-aizhan, kaneko-masahiro, hirasawa-tosho}@ed.tmu.ac.jp
komachi@tmu.ac.jp

Abstract

Simultaneous translation involves translating a sentence before the speaker's utterance is completed in order to realize real-time understanding in multiple languages. This task is significantly more challenging than the general full sentence translation because of the shortage of input information during decoding. To alleviate this shortage, we propose multimodal simultaneous neural machine translation (MSNMT), which leverages visual information as an additional modality. Our experiments with the Multi30k dataset showed that MSNMT significantly outperforms its text-only counterpart in more timely translation situations with low latency. Furthermore, we verified the importance of visual information during decoding by performing an adversarial evaluation of MSNMT, where we studied how models behaved with incongruent input modality and analyzed the effect of different word order between source and target languages.

1 Introduction

Simultaneous translation is a natural language processing (NLP) task in which translation begins before receiving the whole source sentence. It is widely used in international summits and conferences where real-time comprehension is one of the essential aspects. Simultaneous translation is already a difficult task for human interpreters because the message must be understood and translated while the input sentence is still incomplete, especially for language pairs with different word orders (e.g. SVO-SOV) (Seeber, 2015). Consequently, simultaneous translation is more challenging for machines. Previous works attempt to solve this task by predicting the sentence-final verb (Grissom II et al., 2014), or predicting unseen syntactic constituents (Oda et al., 2015). Given the difficulty of predicting future inputs based on existing limited inputs, Ma et al. (2019) proposed a simple simultaneous neural machine translation (SNMT) approach `wait-k` which generates the target sentence concurrently with the source sentence, but always `k` tokens behind, satisfying low latency requirements.

However, previous approaches solve the given task by solely using the text modality, which may be insufficient to produce a reliable translation. Simultaneous interpreters often consider various additional information sources such as visual clues or acoustic data while translating (Seeber, 2015). Therefore, we hypothesize that using supplementary information, such as visual clues, can also be beneficial for simultaneous machine translation.

To this end, we propose Multimodal Simultaneous Neural Machine Translation (**MSNMT**) that supplements the incomplete textual modality with visual information, in the form of an image. It will predict still missing information to improve translation quality during the decoding process. Our approach can be applied in various situations where visual information is related to the content of speech such as presentations with slides (e.g. TED Talks[1]) and news video broadcasts[2]. Our experiments show that the proposed MSNMT method achieves higher translation accuracy than the SNMT model that does not use images by leveraging image information. To the best of our knowledge, we are the first to propose the incorporation of visual information to solve the problem of incomplete text information in SNMT.

The main contributions of our research are as follows. We propose to combine multimodal and simultaneous NMT, therefore, discovering cases where such multimodal signals are beneficial for

*These authors contributed equally to this paper

[1]https://interactio.io/

[2]https://www.a.nhk-g.co.jp/bilingual-english/broadcast/nhk/index.html

Proceedings of the 5th Conference on Machine Translation (WMT), pages 594–603
Online, November 19–20, 2020.

the end-task. Our MSNMT approach brings significant improvement in simultaneous translation quality by enriching incomplete text input information using visual clues. As a result of a thorough analysis, we conclude that the proposed method is able to predict tokens that have not appeared yet for source-target language pairs with different word order (e.g. English→Japanese). By providing an adversarial evaluation, we showed that the models indeed utilize visual information.

2 Related Work

For simultaneous translation, it is crucial to predict the words that have not appeared yet. For example, it is important to distinguish nouns in SVO-SOV translation and verbs in SOV-SVO translation (Ma et al., 2019). SNMT can be realized with two types of policy: fixed and adaptive policies (Zheng et al., 2019b). Adaptive policy decides whether to wait for another source word or emit a target word in one model. Previous models with adaptive policies include explicit prediction of the sentence-final verb (Grissom II et al., 2014; Matsubara et al., 2000) and unseen syntactic constituents (Oda et al., 2015). Most dynamic models with adaptive policies (Gu et al., 2017; Dalvi et al., 2018; Arivazhagan et al., 2019; Zheng et al., 2019a,c, 2020) have the advantage of exploiting input text information as effectively as possible due to the lack of such information in the first place. Meanwhile, Ma et al. (2019) proposed a simple `wait-k` method with fixed policy, which generates the target sentence only from the source sentence that is delayed by `k` tokens. However, their model for simultaneous translation relies only on the source sentence. In this research, we concentrate on the `wait-k` approach with fixed policy, so that the amount of input textual context can be controlled to analyze better whether multimodality is effective in SNMT.

Multimodal NMT (MNMT) for full-sentence machine translation has been developed to enrich text modality by using visual information (Hitschler et al., 2016; Specia et al., 2016; Elliott and Kádár, 2017). While the improvement brought by visual features is moderate, their usefulness is proven by Caglayan et al. (2019). They showed that MNMT models are able to capture visual clues under limited textual context, where source sentences are synthetically degraded by color deprivation, entity masking, and progressive masking. However, they use an artificial setting where they deliberately deprive the models of source-side textual context by masking. However, our research has discovered an actual end-task and has shown the effectiveness of using multimodal data for it. Compared with the entity masking experiments (Caglayan et al., 2019), where they use a model exposed to only `k` words, our model starts by waiting for the first `k` source words and then generates each target word after receiving every new source token, eventually seeing all input text.

In MNMT, visual features are incorporated into standard machine translation in many ways. Doubly-attentive models are used to capture the textual and visual context vectors independently and then combine these context vectors in a concatenation manner (Calixto et al., 2017) or hierarchical manner (Libovický and Helcl, 2017). Some studies use visual features in a multitask learning scenario (Elliott and Kádár, 2017; Zhou et al., 2018). Also, recent work on MNMT has partly addressed lexical ambiguity by using visual information (Elliott et al., 2017; Lala and Specia, 2018; Gella et al., 2019) showing that using textual context with visual features outperform unimodal models.

In our study, visual features are extracted using image processing techniques and then integrated into an SNMT model as additional information, which is supposed to be useful to predict missing words in a simultaneous translation scenario. To the best of our knowledge, this is the first work that incorporates external knowledge into an SNMT model.

3 Multimodal Simultaneous Neural Machine Translation Architecture

Our main goal is to investigate if image information would bring improvement on SNMT. As a result, two tasks could benefit from each other by combining them.

In this section, we describe our MSNMT model, which is composed by combining an SNMT framework `wait-k` (Ma et al., 2019) and a multimodal model (Libovický and Helcl, 2017). We base our model on the RNN architecture, which is widely used in MNMT research (Libovický and Helcl, 2017; Caglayan et al., 2017a; Elliott and Kádár, 2017; Zhou et al., 2018; Hirasawa et al., 2019). The model takes a sentence and its corresponding image as inputs. The decoder of the MSNMT model outputs the target language sentence in a simultaneous and multimodal manner by attaching

attention not only to the source sentence but also to the image related to the source sentence.[3]

3.1 Simultaneous Translation

We first briefly review standard NMT to set up the notations. The encoder of standard NMT model always takes the whole input sequence $\mathbf{X} = (x_1, ..., x_n)$ of length n where each x_i is a word embedding and produces source hidden states $\mathbf{H} = (h_1, ..., h_n)$. The decoder predicts the next output token y_t using $\mathbf{H}$ and previously generated tokens, denoted $\mathbf{Y}_{<t} = (y_1, ..., y_{t-1})$. The final output is calculated using the following equation:

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{t=1}^{|\mathbf{Y}|} p(y_t|\mathbf{X}, y_{<t}) \quad (1)$$

Different from standard neural translation, in which each y_i is predicted using the entire source sentence $\mathbf{X}$, the simultaneous translation requires the model to translate concurrently with the growing source sentence. We incorporate the `wait-k` approach (Ma et al., 2019) for our simultaneous translation model. Instead of waiting for the whole sentence before translating, this model waits for only the first `k` tokens and starts to generate each target tokens after taking every new source token one by one. It stops taking new input tokens once the whole input sentence is on board. For example, if $\mathtt{k} = 3$, the first target token is predicted using the first 3 source tokens, and the second target token using the first 4 source tokens. The `wait-k` decoding probability $p_{\texttt{wait-k}}$ is:

$$p_{\texttt{wait-k}}(\mathbf{Y}|\mathbf{X}) = \prod_{t=1}^{|\mathbf{Y}|} p(y_t|\mathbf{X}_{\leq g(t)}, y_{<t}) \quad (2)$$

where $g(t)$ is the `wait-k` policy function which decides how much input text to read and translate, $\mathbf{X}_{\leq g(t)} = (x_1, ..., x_{g(t)})$ and $g(t)$ is $0 \leq t \leq n$. $g(t)$ is defined as follows:

$$g(t) = \min\{k + t - 1, n\} \quad (3)$$

When $k + t - 1$ is over source length n, $g(t)$ is fixed to n, which means the remaining target tokens (including current step) are generated using the full source sentence. For full sentence translation, $g(t)$ is constant $g(t) = n$.

[3]Our code is publicly available at: https://github.com/toshohirasawa/mst. We fixed our code based on the comments of Ozan Caglayan.

3.2 Multimodal Translation

We use a hierarchical attention combination technique (Libovický and Helcl, 2017) to incorporate visual and textual features into an MNMT model. This model calculates the independent context vectors from the textual features $\boldsymbol{h}^{\text{txt}} = (h_1^{\text{txt}}, ..., h_n^{\text{txt}})$ and the visual features $\boldsymbol{h}^{\text{img}} = (h_1^{\text{img}}, ..., h_m^{\text{img}})$, which are extracted by the textual encoder and the image processing model, respectively. It then combines the resulting two vectors using a second attention mechanism, which helps to perform simultaneous translation taking into account visual information.

Specifically, we compute the context vectors c_i^{f} for each image ($\text{f} = \text{img}$) and text ($\text{f} = \text{txt}$) modality independently using the following equations:

$$e_{i,j}^{\text{f}} = \Omega^{\text{f}}(s_i, h_j^{\text{f}}) \quad (4)$$

$$\alpha_{i,j}^{\text{f}} = \frac{\exp(e_{i,j}^{\text{f}})}{\sum_{l=1}^{|\mathbf{h}^{\text{f}}|} \exp(e_{i,l}^{\text{f}})} \quad (5)$$

$$c_i^{\text{f}} = \sum_{j=1}^{|\mathbf{h}^{\text{f}}|} \alpha_{i,j}^{\text{f}} h_j^{\text{f}} \quad (6)$$

where Ω^{f} is a feedforward network for each modality f; s_i is i-th decoder hidden state.

We project these image and text context vectors into a common space and compute another distribution over the projected context vectors and their corresponding weighted average using the second attention:

$$\tilde{e}_i^{\text{f}} = \Psi(s_i, c_i^{\text{f}}) \quad (7)$$

$$\beta_i^{\text{f}} = \frac{\exp(\tilde{e}_i^{\text{f}})}{\sum_{\text{r}\in\{\text{img,txt}\}} \exp(\tilde{e}_i^{\text{r}})} \quad (8)$$

$$\tilde{c}_i = \sum_{\text{r}\in\{\text{img,txt}\}} \beta_i^{\text{r}} W^{\text{r}} c_i^{\text{r}} \quad (9)$$

where Ψ is a feedforward network. Equation 8 calculates the second attention to combine the image and text vectors. W^{r} is a weight matrix used to compute the context vector $\tilde{c}_i$ calculated from image and text features. The final hypothesis $\mathbf{Y}$ has the probability:

$$p_{\text{mnmt}}(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) = \prod_{t=1}^{|\mathbf{Y}|} p(y_t|\mathbf{X}, \mathbf{Z}, y_{<t}) \quad (10)$$

where $\mathbf{Z}$ represents input image features.

3.3 Multimodal Simultaneous Neural Machine Translation

In this subsection, we describe the structure of the MSNMT model, which is a combination of the models described in Sections 3.1 and 3.2. The method for calculating the image context vector is the same as for MNMT; however, the text context vector (Equation 6) for the t-th step is calculated as follows:

$$\hat{c}_i^{\text{txt}} = \sum_{j=1}^{g(t)} \alpha_{i,j}^{\text{txt}} h_j^{\text{txt}} \tag{11}$$

Thus $\hat{c}_i^{\text{txt}}$ is calculated from the input text prefix determined by `wait-k` policy function $g(t)$. Then we apply the second attention to $\hat{c}_i^{\text{txt}}$ and c_i^{img} in order to calculate $\tilde{c}_i$ (Equation 9).

The decoding probability becomes as follows:

$$p_{\text{msnmt}}(\mathbf{Y}|\mathbf{X},\mathbf{Z}) = \prod_{t=1}^{|\mathbf{Y}|} p(y_t|\mathbf{X}_{\leq g(t)},\mathbf{Z},y_{<t}) \tag{12}$$

4 Experimental Setup

4.1 Dataset

We experiment with our model in four translation directions consisting of 5 languages: English (En), German (De), French (Fr), Czech (Cs), and Japanese (Ja). All language pairs include En on the source side.

We used the train, development, and test sets from the Multi30k (Elliott et al., 2016) dataset published in the WMT16 Shared Task, which is a benchmark dataset generally used in MNMT research (Libovický and Helcl, 2017; Caglayan et al., 2019; Elliott and Kádár, 2017; Zhou et al., 2018; Hirasawa et al., 2019) for En→De, En→Fr and En→Cs.

Nakayama et al. (2020) released F30kEnt-JP dataset[4] which contains Japanese translations of first two original English captions for each image of the Flickr30k Entities dataset (Plummer et al., 2017). They follow the same annotation rules as the Flickr30k Entities dataset using exactly the same tags with entity types and IDs. We preprocessed this data as follows: 1) The parallel En→Ja data was created by taking alignment using corresponding IDs assigned to each Japanese translation entity with the IDs of Flickr30k entities.[5] 2) The created parallel data was aligned with its corresponding images using text files named $(image\_id).txt$ corresponding to each image in Flickr30k. 3) Finally, the created multimodal data was split to train, dev, and test following data splits of Multi30k using the same Multi30k image IDs. Note that the English side of En→Ja parallel data extracted from F30kEnt-JP and English side of Multi30k data are thought to be somewhat comparable but not strictly the same while their corresponding images are the same.

Data split for all language pairs were as follows: training set, 29,000 sentence pairs, development set, 1,014 sentence pairs, and 1,000 sentence pairs for the test set. This dataset's average sentence length is 12-13 tokens for En, De, Fr, Cs and 20 tokens for Ja.

We limit the vocabulary size of the source and the target languages after concatenating them to 10,000 sub-words (Sennrich et al., 2016). All sentences are preprocessed with lower-casing, tokenizing, and normalizing the punctuation using the Moses script[6]. To tokenize Japanese sentences, we used MeCab[7] with the IPA dictionary.

Visual features are extracted using pre-trained ResNet (He et al., 2016). Technically, we encode all images in Multi30k with ResNet-50 and pick out the hidden state in the pool5 layer as a 2,048-dimension visual feature.

4.2 Systems

We compare the following models: **1. SNMT:** We use only text modality for training data as a baseline for each `wait-k` model. **2. MSNMT:** We use image modality along with text modality for a training data for each `wait-k` model.

To train the above models, we utilize attention NMT (Bahdanau et al., 2015) with a 2-layer unidirectional GRU encoder and a 2-layer conditional GRU decoder. We use the open-source implementation of the `nmtpytorch` toolkit v3.0.0 (Caglayan et al., 2017b). We first pre-train the MSNMT model for each `k` until convergence using only text data and use zeros for visual features. Then we continue training MSNMT on multimodal data for

[4] https://github.com/nlab-mpg/Flickr30kEnt-JP

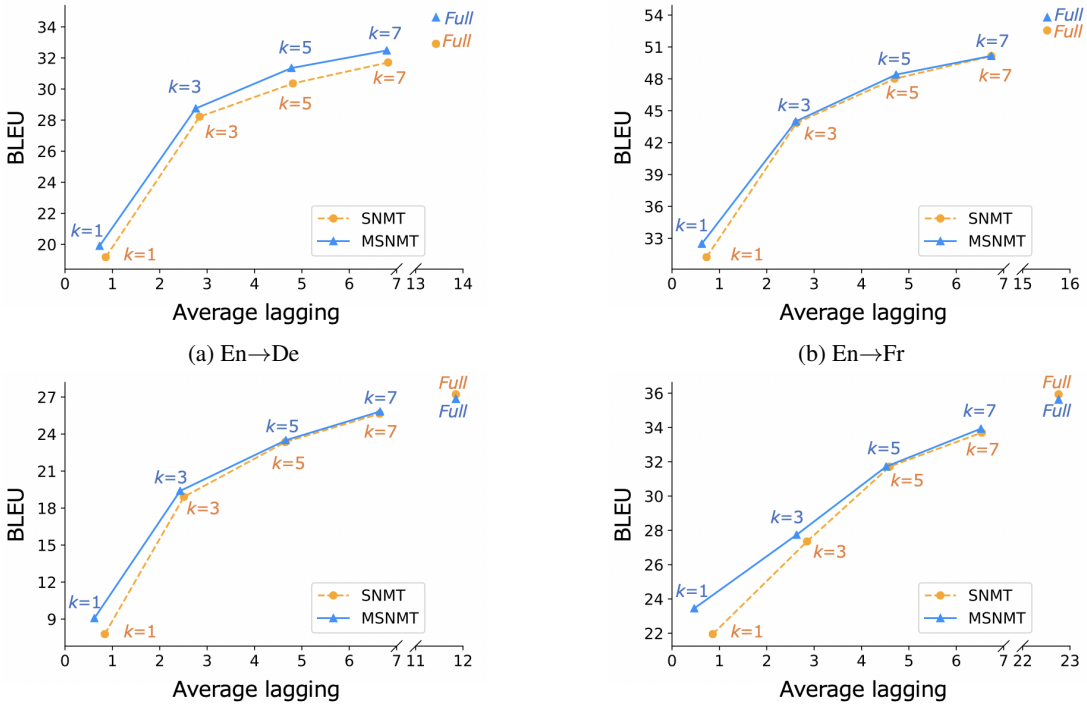
[5] We used the second translations due to some empty translations of the first captions.

[6] We applied preprocessing using task1-tokenize.sh from https://github.com/multi30k/dataset.

[7] http://taku910.github.io/mecab, version 0.996.

wait-k	En→De		En→Fr		En→Cs		En→Ja	
	S	M	S	M	S	M	S	M
1	19.18	†**19.90**	31.23	†**32.49**	7.78	†**9.07**	21.95	†**23.45**
3	28.22	†**28.75**	43.85	**43.99**	18.91	†**19.39**	27.35	†**27.74**
5	30.38	†**31.48**	48.01	†**48.40**	23.35	**23.50**	31.71	**31.72**
7	31.72	**32.14**	50.14	**50.16**	25.65	**25.83**	33.70	**33.93**
Full	34.64	**34.84**	53.55	**53.78**	**27.22**	26.85	**35.93**	35.62

Table 1: BLEU scores of SNMT (S) and MSNMT (M) models for four translation directions on test set. Results are the average of four runs. **Bold** indicates the best BLEU score for each `wait-k` for each translation direction. "†" indicates statistical significance of the improvement over SNMT.



(a) En→De

(b) En→Fr

(c) En→Cs

(d) En→Ja

Figure 1: Average Lagging scores. Results are the average of four runs.

each `k`. We employ early-stopping: the training was stopped when the BLEU score did not increase on the development set for 10 epochs for MSNMT pre-training, 5 epochs for MSNMT fine-tuning, and 15 epochs for SNMT training.

In order to keep our experiments as pure as possible, we will not use additional data or other types of models. It will allow us to control the amount of input textual context, so we can easily analyze the relationship between the amount of textual and visual information.

4.3 Hyperparameters

We use the same hyperparameters for SNMT and MSNMT for a fair comparison as follows. All models have word embeddings of 200 and recurrent layers of dimensionality 400 units with 2way sharing of embeddings in the network. We used Adam (Kingma and Ba, 2015) with a learning rate of 0.0004. Decoders were initialized with zeros. We used a minibatch size of 64 for training and 32 for fine-tuning. Rates of dropout applied on source embeddings, source encoder states and pre-softmax activations were 0.4, 0.5, and 0.5, respectively. We set the max length of the input to 100. `wait-k` experiments were conducted for 1, 3, 5, 7, and Full settings. For MSNMT only hyperparameters, the sampler type was set to approximate, and channels were set to 2048. The fusion type was set to hierarchical mode.

4.4 Evaluation

We report BLEU scores calculated using Moses' multi-bleu.perl, which is a widely used evalu-

wait-k	En→De		En→Fr		En→Cs		En→Ja	
	C	I	C	I	C	I	C	I
1	†**19.90**	8.19	†**32.49**	18.00	†**9.07**	6.83	†**23.45**	17.57
3	†**28.75**	26.78	†**43.99**	42.31	†**19.39**	18.78	†**27.74**	24.51
5	**31.48**	31.08	**48.40**	48.19	†**23.50**	22.81	†**31.72**	28.57
7	†**32.14**	32.04	**50.16**	50.15	†**25.83**	25.09	†**33.93**	31.03
Full	**34.84**	34.40	†**53.78**	53.10	**26.85**	26.84	**35.62**	35.59

Table 2: Image Awareness results on the test set. BLEU scores of MSNMT Congruent (C) and Incongruent (I) settings for four translation directions. Results are the average of four runs. **Bold** indicates the best BLEU score for each wait-k for each translation direction. "†" indicates the statistical significance of the improvement over Incongruent settings.

ation metric in MT, on our test sets for each wait-k model.[8] Statistical significance ($p < 0.05$) on the difference of BLEU scores was tested by Moses' *bootstrap-hypothesis-difference-significance.pl*. "Full" means that the whole input sentence is used as an input for the model to start translating. All reported results are the average of four runs using four different random seeds.

Additionaly, we use open-sourced Average Lagging (AL) latency metric proposed by Ma et al. (2019) to evaluate the latency for SNMT and MSNMT systems.[9] It calculates the degree of out of sync time with the input, in terms of the number of source tokens as follows:

$$\mathrm{AL}_g(\mathbf{X}, \mathbf{Y}) = \frac{1}{\tau_g(|\mathbf{X}|)} \sum_{t=1}^{\tau_g(|\mathbf{X}|)} g(t) - \frac{t-1}{r} \quad (13)$$

where $r = |\mathbf{Y}|/|\mathbf{X}|$ is the target-to-source length ratio and τ_g is the decoding step when source sentence finishes:

$$\tau_g(|\mathbf{X}|) = \min\{t | g(t) = |\mathbf{X}|\} \quad (14)$$

5 Results

Table 1 illustrates the BLEU scores of MSNMT and SNMT models on the test set. MSNMT systems show significant improvements over SNMT systems for all language pairs when input textual information is limited. Note that the difference of BLEU scores between MSNMT and SNMT becomes larger as the k gets smaller, especially when the target language is distant from English in terms of word order (e.g. Cs and Ja). On the other hand, the availability of more tokens during the decoding process ($k \geq 5$) leads to the text information becoming sufficient in some cases.

[8]Due to space constraints, we show results only for test sets.

[9]https://github.com/SimulTrans-demo/STACL

Figure 1 shows translation quality against AL for four language directions. In all these figures, we observe that, as k increases, the gap between BLEU scores for MSNMT and SNMT decreases. We also observe that AL scores are better for MSNMT as k decreases. From these results, it can be seen that in terms of latency, the smaller k is, the more beneficial the visual clues become.

6 Analysis

In this section, we provide a thorough analysis to further investigate the effect of visual data to produce a simultaneous translation by (a) providing adversarial evaluation; and (b) analyzing the impact of different word order for En→Ja language pair.

6.1 Adversarial Evaluation

In order to determine whether MSNMT systems are aware of the visual context (Elliott, 2018), we perform the adversarial evaluation on the test set. We present our system with correct visual data with its source sentence (Congruent) as opposed to random visual data as an input (Incongruent) (Elliott, 2018).Therefore, we reversed the order of 1,000 images of the test set, so there will be no overlapping congruent visual data. Then we reconstruct image features for those images to use as an input.

Results of image awareness experiments are shown in Table 2. We can see the large difference in BLEU scores between MSNMT congruent (C columns) and incongruent (I columns) settings when k are small. This implies that our proposed model utilizes images for translation by learning to extract needed information from visual clues. The interesting part is for a full translation, where scores for the incongruent setting are very close to those of the congruent setting. The reason is that when textual information is enough, visual information becomes not that relevant in some cases.

6.2 How Source-Target Word Order Affects Translation

In `wait-k` translations, for the En→Ja language pair with different word orders (SVO vs. SOV), some source tokens should be translated before they are presented to the decoder for grammaticality and fluency purposes. Hence, the model also needs to handle such cases well apart from the "usual" order. We hypothesized that MSNMT models, given additional visual information, are able to translate such cases better than SNMT models. Therefore, we investigated how many tokens were correctly translated that are not given as input yet.

First, we quantitatively analyze how well we can translate entities that are not presented from the source yet but should exist in target sentences. To align the source and target entities, we use the entities' annotation attached to both the source and target sentences. Given that annotated entities have the same IDs and tags for both English and Japanese, we can align, calculate, and extract those entities from source and target sentences. If the index of the first token of the aligned target entity is not given as input at timestep `k` yet, we count them for each `k` scenario as # total entities (Table 4). For example, in Table 3 a `wait-3` model should start translating after a token "rappelling" is presented to the model. And if an ID of the entity of "海 (a body of water)" is in the target sentences but not in the inputted part yet, we count it as an entity that should be translated before being inputted to the model. Similarly, an entity of "断崖 (cliff)" is already presented to the model at timestep 5, so we do not count those entities. If the same entity ID appears more than once in one sentence, we exclude those entities due to the impossibility of alignments. Finally, for each model during decoding, if those entities are included in the model's translation results with a perfect match from pre-calculated # total entities, we consider them as correctly translated.[10]

Table 4 demonstrates the results. `k` column is to determine how many tokens a model waits before starting translating. Note that `k`=Full is not included because all entities are given at the time of translation. The reason that the total number of entities that were not inputted yet decreases when `k` increases (# total entities column) is that more entities are already available for the model for translation. `wait-k` columns show how many entities were correctly translated by `wait-k` SNMT and MSNMT models from # total entities for each `k` scenario. Columns `Full` show upper-bounds of how many entities can be correctly translated if the models were trained with full sentences for entities from each `k`. Comparing `Full` results to `wait-k` for both SNMT and MSNMT shows that it is hard to correctly translate entities when `k` is small. Furthermore, comparing `wait-k` results of SNMT to MSNMT, it can be seen that the smaller value of `k`, the better MSNMT can handle different source-target word order than SNMT.

(a) A person rappelling a cliff.

(b) Eight men on motorcycles.

Figure 2: Images presented in translation examples (Table 5).

As an example, we sampled sentences and their images from the En→Ja test set (Figure 2) to compare the outputs of our systems. Table 5 lists their translations generated by SNMT (S) and MSNMT (M) models. In the first example, an SNMT model with `wait-3` could not predict "海 (sea, a body of water)" which appears at the end of the source sentence and generated an erroneous "岩 (rock)" which is not present neither in source text nor in a corresponding image. Contrarily, the MSNMT model with `wait-3` was able to correctly predict "海 (body of water)" even before it was inputted by capturing visual information. When a full sentence is given as an input, MSNMT translated it correctly using more information, unlike SNMT, which translated only from the given text and generated incorrect "登って (climbing)" instead of "降りて (rappelling)". Interestingly, in the second example, the MSNMT model with `wait-3` predicted "自転車 (bicycles)" instead of "オートバイ (motorcycles)" at the beginning of the sentence, while the SNMT model with `wait-3` was not able to generate any vehicle entities. Also, both MSNMT models with `wait-3` and `Full` correctly captured that there were eight men, whilst both SNMT models incorrectly predicted about one and two men. From these results, we can conclude that visual clues pos-

[10] We can not create # total entities from decoded tokens directly due to unavailability of entity annotations.

t	1	2	3	4	5	6	7	8	9	10	11										
Source	a	person	rappelling	a	cliff	above	a	body	of	water	.										
Target, k=3				海	の	上	に	ある	断崖	を	降り	て	いる	一	人	の	男性	。			
Entity count				✓					✗								✗				

Table 3: Example of En→Ja translation to count entities that should be translated before introducing it to a model in case of wait-3 (see Figure 2a). A wait-k model starts translating after k tokens are inputted. Colors represent the same entities. ✓ indicates entities that are not presented to the model at timestep t yet and ✗ indicates entities that are already seen by the model at timestep t. We count only those entities marked with ✓ for # total entities (Table 4).

k	# total entities	# correct entities by S		# correct entities by M	
		wait-k	Full	wait-k	Full
1	1,343	251	**716**	**270**	707
3	852	229	**433**	**242**	432
5	502	147	**247**	**151**	243
7	320	106	**160**	106	159

Table 4: Number of entities that were correctly translated before being presented to the model by SNMT (S) and MSNMT (M) models with their for each k. Results are the average of four runs.

Source	a person rappelling a cliff above a body of water .
Target	海の上にある断崖を降りている一人の男性。
S wait-3	誰かが、岩の上で崖を登る。 (someone climbs a cliff on a rock.)
M wait-3	人が海の上で崖を降りている。 (a person is rappelling a cliff above the sea.)
S Full	人が水域の上の崖を登っている。 (a person is climbing a cliff above a body of water.)
M Full	人が水域の上で崖を降りている。 (a person is rappelling a cliff above a body of water.)
Source	eight men on motorcycles dressed in red and black are all lined up on the side of the street .
Target	赤と黒の服を着たオートバイに乗っている８人の男性が通りの脇にずらりと並んでいる。
S wait-3	白い服を着て、黒と黒の服を着た１人の男性が、通りの脇に並んでいる。 (a man in white and black and black is standing beside the street.)
M wait-3	自転車に乗っている赤と黒の服を着た８人の男性が、通りの側面にある。 (eight men in red and black clothes riding a bicycle are on the side of the street.)
S Full	赤と黒の服を着た、オートバイに乗った２人の男性が、通りの脇で並んでいる。 (two men on motorcycles, dressed in red and black, line up by the side of the street.)
M Full	赤と黒の服を着た、オートバイに乗った８人の男性が、通りの側面に並んでいる。 (eight men on motorcycles, dressed in red and black, line the side of the street.)

Table 5: Examples of En→Ja translations from test set using SNMT (S) and MSNMT (M) models (also refer to Figure 2). In () are shown their English meanings. The same colors indicate the same entity types.

itively impact generated translations where there is still a lack of textual information, especially when we deal with language pairs with different word order.

7 Conclusion

In this paper, we proposed a multimodal simultaneous neural machine translation approach, which takes advantage of visual information as an additional modality to compensate for the shortage of input text information in the simultaneous neural machine translation. We showed that in a wait-k setting, our model significantly outperformed its text-only counterpart in situations where only a few input tokens are available to begin translation. We showed the importance of the visual information for simultaneous translation, especially in the low latency setup and for a language pair with word-order differences. We hope that our proposed method can be explored even further for various tasks and datasets.

In this paper, we created a separate model for each value of wait-k. However, in future work, we plan to experiment on having a single model for all k values (Zheng et al., 2019b). Furthermore, we acknowledge the importance of investigating MSNMT effects on more realistic data (e.g. TED), where the utterance does not necessarily match a shown image while speaking and/or where its context can not be guessed from the shown image.

Acknowledgments

We are immensely grateful to Raj Dabre and Rob van der Goot who provided expertise, support, and insightful comments that significantly improved the manuscript. We would also like to show our gratitude to Desmond Elliot for valuable feedback and paper discussions. We want to thank Ozan Caglayan for pointing out critical bugs in our previous implementation.

References

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017a. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439.

Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017b. NMTPY: A flexible toolkit for advanced neural machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 109(1):15–28.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4159–4170.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, and Stephan Vogel. 2018. Incremental decoding and training methods for simultaneous translation in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 493–499.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Desmond Elliott and Àkos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141.

Spandana Gella, Desmond Elliott, and Frank Keller. 2019. Cross-lingual visual verb sense disambiguation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1998–2004.

Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1342–1352.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK Li. 2017. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1, Long Papers)*, pages 1053–1062.

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.

Tosho Hirasawa, Hayahide Yamagishi, Yukio Matsumura, and Mamoru Komachi. 2019. Multimodal machine translation with embedding prediction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 86–91.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2399–2409.

Diederik Kingma and Jimmy Ba. 2015. Adam: a method for stochastic optimization. In *The International Conference on Learning Representations*.

Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3810–3817.

Jindřich Libovický and Jindřich Helcl. 2017. 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)*, pages 196–202.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036.

Shigeki Matsubara, Kiyoshi Iwashima, Nobuo Kawaguchi, Katsuhiko Toyama, and Yoichi Inagaki. 2000. Simultaneous Japanese-English interpretation based on early prediction of English verb. In *Proceedings of The Fourth Symposium on Natural Language Processing*, pages 268–273.

Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4204–4210.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Syntax-based simultaneous translation through prediction of unseen syntactic constituents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 198–207.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123(1):74–93.

Kilian G Seeber. 2015. Simultaneous interpreting. In *The Routledge Handbook of Interpreting*, pages 91–107.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistic*, pages 1715–1725.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: (Volume 2: Shared Task Papers)*, pages 543–553.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019a. Simpler and faster learning of adaptive policies for simultaneous translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1349–1354.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019b. Simultaneous translation with flexible policy via restricted imitation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5816–5822.

Renjie Zheng, Mingbo Ma, Baigong Zheng, and Liang Huang. 2019c. Speculative beam search for simultaneous translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1395–1402.

Renjie Zheng, Mingbo Ma, Baigong Zheng, Kaibo Liu, and Liang Huang. 2020. Opportunistic decoding with timely correction for simultaneous translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 437–442.

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