Exploiting Multimodal Reinforcement Learning for Simultaneous Machine Translation

Julia Ive[1], **Andy Mingren Li**[1], **Yishu Miao**[1], **Ozan Caglayan**[1],
Pranava Madhyastha[1], **Lucia Specia**[1,2,3]
Imperial College London[1], University of Sheffield[2], ADAPT - Dublin City University[3]
j.ive@ic.ac.uk, andy.li16@imperial.ac.uk, y.miao20@imperial.ac.uk, o.caglayan@ic.ac.uk
pranava@ic.ac.uk, l.specia@ic.ac.uk

Abstract

This paper addresses the problem of simultaneous machine translation (SiMT) by exploring two main concepts: (a) adaptive policies to learn a good trade-off between high translation quality and low latency; and (b) visual information to support this process by providing additional (visual) contextual information which may be available before the textual input is produced. For that, we propose a multimodal approach to simultaneous machine translation using reinforcement learning, with strategies to integrate visual and textual information in both the agent and the environment. We provide an exploration on how different types of visual information and integration strategies affect the quality and latency of simultaneous translation models, and demonstrate that visual cues lead to higher quality while keeping the latency low.

1 Introduction

Research into automating real-time interpretation has explored deterministic and adaptive approaches to build policies that address the issue of translation delay (Ryu et al., 2006; Cho and Esipova, 2016; Gu et al., 2017). In another recent development, the availability of multimodal data (such as visual information) has driven the community towards multimodal approaches for machine translation (MMT) (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). Although deterministic policies have been recently explored for simultaneous MMT (Caglayan et al., 2020; Imankulova et al., 2020), there are no studies regarding how multimodal information can be exploited to build flexible and adaptive policies for simultaneous machine translation (SiMT).

Applications of reinforcement learning (RL) for unimodal SiMT have highlighted the challenges for the agent to maintain good translation quality while learning an optimal translation path (i.e. a sequence of `READ`/`WRITE` decisions at every time step) (Grissom II et al., 2016; Gu et al., 2017; Alinejad et al., 2018).

Incomplete source information will have detrimental effect especially in the cases where significant restructuring is needed while translating from one language to another.

In addition, the lack of information generally leads to high variance during the training in the RL setup. We posit that **multimodality** in adaptive SiMT could help the agent by providing extra signals, which would in turn improve training stability and thus the quality of the estimator and translation decoder.

In this paper, we present the first exploration on multimodal RL approaches for the task of SiMT.

As visual signals, we explore both image classification features as well as visual concepts, which provide global image information and explicit object representations, respectively. For RL, we employ the Policy Gradient method with a pre-trained neural machine translation model acting as the environment.

As the SiMT model is optimised for both translation quality and latency, we apply a combined reward function that consists of a decomposed smoothed BLEU score and a latency score. To integrate visual and textual information, we propose different strategies that operate both on the agent (as prior information or at each step) and the environment side.

In experiments on standard datasets for MMT, our models achieve the highest BLEU scores on most settings without significant loss on average latency, as compared to strong SiMT baselines. A qualitative analysis shows that the agent benefits from the multimodal information by grounding language signals on the images.

Our **main contributions** are as follows: (1) we

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3222–3233
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

propose the first multimodal approach to simultaneous machine translation based on adaptive policies with RL, introducing different strategies to integrate visual and textual information (Sections 3 and 4); (2) we show how different types of visual information and integration strategies affect the quality and latency of the models (Section 5); (3) we demonstrate that providing visual cues to both agent and environment is beneficial: models achieve high quality while keeping the latency low (Section 5).

2 Related Work

In this section, we first present background and related work on SiMT, and then discuss recent work in MMT and multimodal RL.

2.1 Simultaneous Machine Translation

In the context of neural machine translation (NMT), Cho and Esipova (2016) introduce a greedy decoding framework where simple heuristic waiting criteria are used to decide whether the model should read more source words or instead write a target word. Gu et al. (2017) utilise a pre-trained NMT model in conjunction with an RL agent whose goal is to learn a `READ/WRITE` policy by maximising quality and minimising latency. Alinejad et al. (2018) further extend the latter approach by adding a `PREDICT` action with an aim to capture the anticipation of the next source word. Ma et al. (2019) propose an end-to-end, fixed-latency framework called 'wait-k' which allows *prefix-to-prefix* training using a deterministic policy: the agent starts by reading a specified number of source tokens (k), followed by alternating `WRITE` and `READ` actions.

Other approaches to SiMT include re-translation of previous outputs depending on new outputs (Arivazhagan et al., 2020; Niehues et al., 2018) or learning adaptive policies guided by a heuristic or alignment-based approaches (Zheng et al., 2019; Arthur et al., 2020). A general theme in these approaches is their reliance on *consecutive* NMT models pre-trained on full-sentences. However, Dalvi et al. (2018) discuss potential mismatches between the training and decoding regimens of these approaches and propose to perform fine-tuning of the models using chunked data or prefix pairs.

2.2 Multimodal Machine Translation

MMT aims at improving the quality of automatic translation using additional sources of information (Sulubacak et al., 2020). Different methods for fusing textual and visual information have been proposed. These include initialising the textual encoder or decoder with the visual information (Elliott and Kádár, 2017; Caglayan et al., 2017), combining the visual information through spatial feature maps using soft attention (Caglayan et al., 2016; Libovický and Helcl, 2017; Huang et al., 2016; Calixto et al., 2017), and projecting a summary of the visual representations to a common context space via a trained projection matrix (Calixto and Liu, 2017; Caglayan et al., 2017; Elliott and Kádár, 2017; Grönroos et al., 2018). Further, recent work has also focused on exploring Multimodal Pivots (Hitschler et al., 2016) and latent variable models (Calixto et al., 2019) in the context of multimodal machine translation. In this paper, we explore all these strategies, and also the use of *visual concepts*, similar to the approach by Ive et al. (2019).

2.3 Multimodal Reinforcement Learning

Previous work has explored RL with language inputs (Andreas et al., 2017; Bahdanau et al., 2018; Goyal et al., 2019) by making use of language to improve the policy or reward function: for example, the task of navigating in the world grid environment using language instructions (Andreas et al., 2016).

Alternatively, RL with language output can be shaped as sequential decision making for language generation, while conditioning on other modalities. This includes image captioning (Ren et al., 2017), video captioning (Wang et al., 2018), question answering (Das et al., 2018), and text-based games (Côté et al., 2018). Our study sits somewhere in between these different types of work. We have both the source language and respective images as input and the target language as output. Our agent is focused only on learning the `READ` and `WRITE` actions while the translation model is fixed for simplicity.

The central aim of the agent is learning to capture the relevant structures and relations of the modalities that can lead to a better SiMT system.

3 Methods

We first present the architectures for consecutive and baseline fixed policy simultaneous MT (Section 3.1). Then we introduce our RL approaches, both the baseline, the proposed multimodal extension (Section 3.2), and the visual features used by

all multimodal approaches (Section 3.3).

3.1 Baselines

Unimodal MT. We implement a standard encoder-decoder baseline with attention (Bahdanau et al., 2014) which incorporates a two-layer encoder and a two-layer decoder with GRU (Cho et al., 2014) units. Given a source sequence of embeddings $X=\{x_1, \ldots, x_S\}$ and a target sequence of embeddings $Y=\{y_1, \ldots, y_T\}$, the encoder first computes the sequence of hidden states $H=\{h_1, \ldots, h_S\}$ unidirectionally.

The attention layer receives H as *key-values* whereas the hidden states of the first decoder GRU provide the *queries*. The context vector $c_t^{\mathbf{T}}$ produced by the attention layer is given as input to the second GRU. Finally, the output token (y_t) probabilities are obtained by applying a softmax layer on top of the concatenation of the previous word embedding, context vector and the second GRU's hidden state.

For **consecutive NMT**, all source tokens are observed before the decoder begins the process of generation.

Multimodal MT. We extend unimodal MT with multimodal attention (Calixto et al., 2016; Caglayan et al., 2016) in the decoder, in order to incorporate visual information into the baseline NMT. Let us denote the visual counterpart of textual hidden states H by V. Multimodal attention simply applies another attention layer on top of V, which yields a visual context vector $c_t^{\mathbf{V}}$ at each decoding timestep t. The final multimodal context vector that would be given as input to the second GRU is simply the sum of both context vectors.

Unimodal wait-k NMT. We explore deterministic wait-k (Ma et al., 2019) approach as a unimodal baseline[1] for simultaneous NMT. The wait-k model starts by reading k source tokens and writes the first target token. The model then reads and writes one token at a time to complete the translation process. This implies that the attention layer will now attend to a partial textual representation corresponding to k-words. We use the decoding-only variant which does not require re-training an NMT model i.e. it re-uses the already trained consecutive NMT baselines.

[1]These baselines are equivalent to the deterministic approaches used in Caglayan et al. (2020).

3.2 Policy Learning Framework

RL baseline. We closely follow Gu et al. (2017) and cast SiMT as a task of producing a sequence of `READ` or `WRITE` actions. We then devise an RL model that connects the MT system and these actions. The model is based on a reward function that takes into account both quality and latency. Following standard RL, the framework is composed of an environment and an agent. The agent takes the decision of either reading one more input token or writing a token into the output – hence two actions are possible: `READ` and `WRITE`. The environment is a pre-trained NMT system which is *frozen* during RL training.

The agent is a GRU that parameterises a stochastic policy which decides on the action a_t by receiving as input the observation o_t.[2] In our setup, o_t is defined as $[c_t^{\mathbf{T}}; y_t; a_{t-1}]$, i.e. the concatenation of vectors coming from the environment, as well as the previously produced action sequence. At each time step, the agent receives a reward $r_t = r_t^Q + r_t^D$ where r_t^Q is the quality reward (the difference of smoothed BLEU scores for partial hypotheses produced from one step to another) and r_t^D is the latency reward formulated as:

$$r_t^D = \alpha\,[\text{sgn}(C_t - C^*) + 1] + \beta \lfloor D_t - D^* \rfloor_+$$

where C_t denotes the consecutive wait (CW) metric which is added to avoid long consecutive waits (Gu et al., 2017). CW measures how many source tokens are consecutively read between committing two translations. D_t refers to average proportion (AVP) (Cho and Esipova, 2016), which defines the average proportion of wait tokens when translating the words. D^* and C^* are hyper-parameters that determine the expected/target values. The optimal quality-latency trade-off is achieved by balancing the two reward terms. In our reward implementation we again closely follow Gu et al. (2017).

Multimodal extension. Here we focus on integrating the visual information with the agent (see Figure 1). The basic premise is that the addition of multimodal information, especially in the context of MMT, can result in the agent learning better and more flexible policies. We explore several ways to integrate visual information into this framework:

[2]We note that the use of GRU cells is not critical for the multimodal components. They were chosen as they led to the best performance in our implementation.

- **Multimodal initialisation** (`RL-init`) - the agent network is initialised with the image vector V as d_0. We expect this vector to give the agent some context w.r.t. the source sentence so it can potentially read fewer words before producing outputs.

- **Multimodal attention** (`RL-att`, Figure 1) applies another attention layer on top of V, which yields a visual context vector $c_t^{\mathbf{V}}$ at each agent time step t. This visual context vector is a dot product attention $c_t^{\mathbf{V}} = \text{Attention}(V, \text{query} \leftarrow y_t)$ that computes the similarity between V and the embedding of the target word produced by the decoder at the time step t. In this setting, we expect the agent to pay attention to the information in V that will help in defining whether y_t is good enough to be written to the output (potentially with closer relationship to some part of the image information) or we need to read more source words to produce a better y_t. We concatenate $c_t^{\mathbf{V}}$ to o_t, which now becomes $[c_t^{\mathbf{T}}; y_t; a_{t-1}; c_t^{\mathbf{V}}]$;

- As a **control**, we also study **multimodal environment** (`RL-env`, Figure 1) where we use the MMT baseline as environment. Here, we expect the initial translation quality of SiMT RL models be closer to the quality of the respective consecutive multimodal baseline as the image information is expected to compensate for partial source information. When combined with `RL-init` and `RL-att` settings, we expect the agent to exploit different kinds of image information than the environment.

Learning. To learn the multimodal agent, we introduce an additional neural network with the same structure as that of the agent GRU network to provide for control variates (baselines) that improve the Monte-Carlo policy gradient (REINFORCE (Williams, 1992)). Note that here we depart from the previous work where Gu et al. (2017) use a simple multilayer perceptron as the baseline.

Therefore, with the reward r_t at each time step, we obtain the estimation of the gradients by subtracting the baselines $b(o_t)$:

$$\nabla_\theta J(\theta) = \mathbb{E}[\sum_{t=0}^{T-1} \nabla_\theta \log \pi(a_t|o_t)(r_t - b(o_t))]$$

To further reduce the variance of the gradient estimator, we also introduce a temperature τ for controlling the interpolation between discrete action samples and continuous categorical densities, which yields to a Gumbel-Softmax reparameterisation (Jang et al., 2017) that smooths the learning. To be more precise, we use the Gumbel-Softmax distribution instead of argmax while sampling. So the probability of the `WRITE` action is given to the agent network instead of the index of the action.

3.3 Visual Features

In order to represent the visual information, we explore two settings that differ in the organisation of the spatial structure. Regardless of the setting, the image features are linearly projected into the hidden space of the decoder to yield the tensor V.

Image classification features (OC) are *global* image information represented by convolutional feature maps, which are believed to capture spatial cues. These features are extracted from the final convolution layer of a ResNet-50 convolutional neural network (CNN) (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) for object classification. The size of the final feature tensor being 8x8x2048, the visual attention is applied on a grid of 64 equally-sized regions.

Visual Concepts (VC) are explicit object representations where *local* regions are detected as objects and subsequently encoded with 100-dimensional word representations. For a given image, the detector provides 36 object and 36 attribute region proposals which are abstract concepts associated with the image. We represent each of the detected region with its corresponding GloVe (Pennington et al., 2014) word vectors. An image is thus represented by a feature tensor of size 72x100 and the visual attention is now applied on these visual concepts, rather than the uniform grid of the first approach above. We hypothesise that this type of information can result in better referential grounding by using conceptually meaningful units rather than global features. The detector used here is a Faster R-CNN/ResNet-101 object detector (with 1600 object labels) (Anderson et al., 2018)[3] pre-trained on the Visual Genome dataset (Krishna et al., 2017).

[3] https://hub.docker.com/r/airsplay/bottom-up-attention

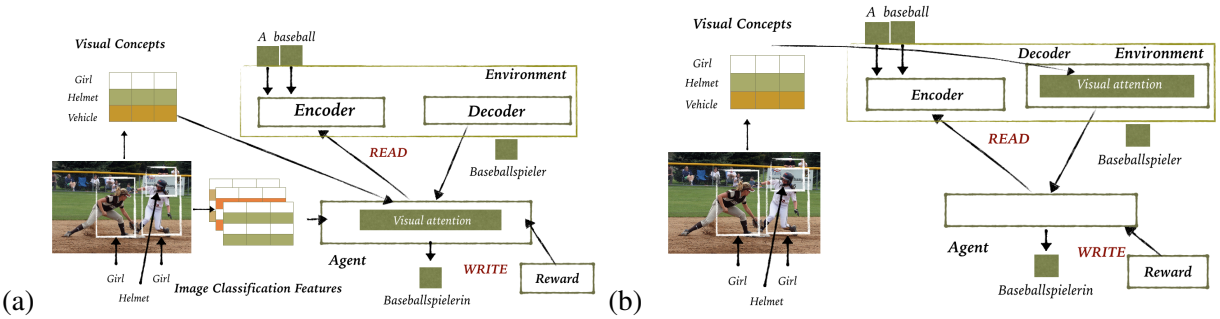


Figure 1: Our multimodal RL SiMT models: the agent interacts with the environment to receive new translation and at each time step produces the `READ/WRITE` action. For each action it receives a reward. The image information can be integrated into the agent by means of an attention mechanism (a, `RL-att`), or into the environment decoder (b, `RL-env`) producing the next translation.

4 Experimental Setup

4.1 Dataset

We perform experiments on the Multi30k dataset (Elliott et al., 2016)[4] which extends the Flickr30k image captioning dataset (Young et al., 2014) with caption translations in German and French (Elliott et al., 2017). Multi30k is a standard MMT dataset that contains parallel sentences in two languages that describe the images. The training set for each language direction comprises 29,000 image-source-target triplets whereas the development and the test sets have around 1,000 samples. We use the corresponding test sets from 2016, 2017 and 2018 for evaluation.

Pre-processing. We use Moses scripts (Koehn et al., 2007) to lowercase, normalise and tokenise the sentences. We then create word vocabularies on the *training* subset of the dataset. We did not use subword segmentation to avoid its potential side effects on fixed policy SiMT and to be able to better analyse the grounding capability of the models. The resulting English, French and German vocabularies contain 9.8K, 11K and 18K tokens, respectively.

4.2 Evaluation

We use **BLEU** (Papineni et al., 2002) for quality, and perform significance testing via bootstrap resampling using the `Multeval` tool (Clark et al., 2011). For latency, we measure **Average proportion** (AVP) (Cho and Esipova, 2016). AVP is the average number of source tokens required to commit a translation. This metric is sensitive to the difference in lengths between source and target. Hence, as our main latency metric we measure **Average Lagging** (AVL) (Ma et al., 2019) which estimates the number of tokens the "writer" is lagging behind the "reader", as a function of the number of input tokens read.

4.3 Training

Hyperparameters. We set the embeddings dimensionality and GRU hidden states to 200 and 320, respectively. We use the ADAM (Kingma and Ba, 2014) optimiser with the learning rate 0.0004 and the batch size of 64. We use `pysimt` (Caglayan et al., 2020) with PyTorch (Paszke et al., 2019) v1.4 for our experiments.[5] We early stop w.r.t. the validation BLEU with the patience of 10 epochs. On a single NVIDIA RTX2080-Ti GPU, the training takes around 35 minutes for the unimodal model and around 1 hour for the multimodal model. The number of learnable parameters is between 6.9M and 9.3M depending on the language pair and the type of multimodality.

For the **RL systems**, we follow (Gu et al., 2017).[6] The agent is implemented by a 320-dimensional GRU followed by a softmax layer and the baseline network is similar to the agent except with a scalar output layer.[7] We use ADAM as the optimiser and set the learning rate and mini-batch size to 0.0004 and 6, respectively. For each sentence pair in a batch, 5 trajectories are sampled. Following best practises in RL, the baseline network is trained to reduce the MSE loss between the predictions and the rewards using a second op-

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

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

[6]https://github.com/nyu-dl/dl4mt-simul-trans

[7]Note that that Gu et al. (2017) use a 2-hidden layer feed-forward network as the baseline network. In our implementation GRUs have demonstrated better performance.

timiser.

For inference, greedy sampling is used to pick action sequences. We set the hyperparameters C^*=2, D^*=0.3, α=0.025 and β= $-$ 1. To encourage exploration, the negative entropy policy term is weighed empirically with 0.001. Following (Gu et al., 2017), we choose the model that maximises the quality-to-latency ratio (BLEU/AVP) on the validation set with a patience of 5 epochs.[8] On a single NVIDIA RTX2080-Ti GPU, the training takes around 2 hours. The number of learnable parameters is around 6M.

Model configurations. We experiment with seven different configurations (below). We consider visual concepts (VC) as the main source of multimodal information. Visual concepts are more abstract forms of multimodal information. Unlike spatial image representation or region of interest-based object representations, where the representation for the same concept can vary significantly across images, visual concepts remain constant. For example, the visual concept "dog" is the same regardless of the breed, colour, size, position, etc. of the concept in different images. Image classification (OC) features are used as a contrastive setting.

- Unimodal RL baseline (`RL-base`): This baseline follows (Gu et al., 2017) where the environment is a text-only NMT model.
- Multimodal agent with VC initialisation (`RL-init` VC): We initialise the agent GRU using a projection of the flattened 72x100 matrix of visual concepts.
- Multimodal agent with attention over VC (`RL-att` VC): The agent attends over the set of visual concepts at each step.
- Multimodal agent with attention over OC (`RL-att` OC): The agent attends over the set of image classification-based spatial feature maps at each step.
- Visually initialised multimodal agent with attention over VC (`RL-init-att` VC): Similar to `RL-att` VC but the agent is also initialised with VC.
- Multimodal environment with unimodal RL agent (`RL-env` VC): The environment is an MMT model, however the agent is a standard RL agent akin to the baseline.
- Multimodal agent with multimodal environment (`RL-env-init-att` VC): This merges all the variants in that both the multimodal environment and the multimodal agent attend to visual concepts, the latter is also initialised with visual information.

[8] We also attempted to choose the model that maximises BLEU or BLEU/AVL but those stopping criteria resulted in instability of convergence.

5 Results

In this section, we first provide the results from our experiments (Section 5.1) and then analyse the behaviour of the (multimodal) agents (Section 5.2).

5.1 Quantitative Results

SiMT vs. Consecutive. We present the main results in Table 1. The top block for each language pair shows the textual Consecutive model and its multimodal counterpart (Consecutive+VC). These are our upperbounds since they have access to the entire source before translating. As expected, they have better BLEU but much larger AVL.

RL SiMT vs. Deterministic policy. The second block in Table 1 shows the deterministic policy Wait-2 and Wait-3 approaches. `RL-base` performs on par with the Wait-2 (English-French) and Wait-3 (English-German). We however emphasise the flexibility of the stochastic policies with RL models. These are particularly beneficial in the multimodal scenario and allow for exploitation of the image information more efficiently especially towards reducing the average lag. We further expand on this later in Section 5.2.

Unimodal RL vs. Multimodal RL. The third block in Table 1 compares all multimodal RL variants against the text-only SiMT RL (`RL-base`). In general, the multimodal RL models produce translations that are significantly better than `RL-base`.

Across Multimodal RL Setups. With regard to different configurations, we observe (1) an increase in quality for the `RL-att` models when compared to `RL-base` which is consistent in both types of visual inputs OC and VC, and (2) a decrease in the lag for the `RL-init` models at a small decrease in quality (for VC `RL-init` in comparison to `RL-base`).

This observation suggests that the RL model with the agent explicitly attending over image information leads to an increase in quality, as the

		test 2016			test 2017			test 2018		
		BLEU↑	AVL↓	AVP↓	BLEU↑	AVL↓	AVP↓	BLEU↑	AVL↓	AVP↓
English → French	Consecutive	58.0	13.1	1.0	50.6	11.1	1.0	36.0	13.8	1.0
	+VC	59.1	13.1	1.0	51.0	11.1	1.0	36.5	13.8	1.0
	Wait-2	48.1	2.6	0.7	42.9	2.6	0.7	32.1	2.7	0.7
	Wait-3	54.0	3.5	0.7	48.6	3.5	0.7	35.5	3.5	0.7
	RL	50.8	3.3	**0.7**	44.3	3.0	**0.7**	32.1	3.5	**0.7**
	+att-OC	53.0*	4.1	0.8	46.4*	3.9	0.8	33.3*	4.4	0.8
	+att-VC	53.0*	4.0	**0.7**	46.5*	3.7	0.8	33.3*	4.2	**0.7**
	+init-VC	49.6	**2.8**	**0.7**	43.3	**2.6**	**0.7**	31.5	**2.9**	**0.7**
	+init-att-VC	52.6*	3.8	**0.7**	46.3*	3.6	**0.7**	33.3*	4.1	**0.7**
	+env-VC	54.0*	3.3	**0.7**	47.2*	3.1	**0.7**	33.7*	3.4	**0.7**
	+env-init-att-VC	**54.0***	3.9	**0.7**	**47.7***	3.8	0.8	**34.4***	4.2	**0.7**
English → German	Consecutive	35.5	13.1	1.0	27.7	11.1	1.0	25.8	13.8	1.0
	+VC	35.9	13.1	1.0	27.0	11.1	1.0	25.4	13.8	1.0
	Wait-2	28.3	2.2	0.6	22.5	2.2	0.7	20.1	2.2	0.6
	Wait-3	32.6	3.0	0.7	25.4	3.0	0.7	24.1	3.0	0.7
	RL	31.0	2.7	0.7	23.0	2.6	0.7	22.0	2.7	0.7
	+att-OC	33.9*	3.7	0.7	**25.8***	3.4	0.7	**24.5***	3.8	0.7
	+att-VC	33.3*	3.3	0.7	24.7*	3.0	0.7	23.0*	3.2	0.7
	+init-VC	29.7	2.8	0.7	21.3	2.4	0.7	20.5	2.5	0.6
	+init-att-VC	**34.1***	3.3	0.7	25.3*	3.1	0.7	24.1*	3.4	0.7
	+env-VC	30.0	**2.5**	**0.6**	21.7	**2.2**	**0.6**	19.7	**2.2**	**0.6**
	+env-init-att-VC	31.4	3.0	0.7	24.0*	2.9	0.7	22.4	3.0	0.7

Table 1: Results for the test sets 2016, 2017 and 2018 (averaged over 3 runs): * marks statistically significant increases in BLEU w.r.t. RL-base (p-value ≤ 0.05). Bold highlights best scores across the RL approaches.

multimodal agent model is more selective towards the word choice. The RL-init configuration with prior image context on the other hand reduces the lag and seems to use WRITE actions more often than READ actions. It is interesting that OC and VC features result in similar quality translations, however we see that on average the average lag is lower with VC. We hypothesise that this could be due to the fact that the representations remain constant across images (see Section 4.3).

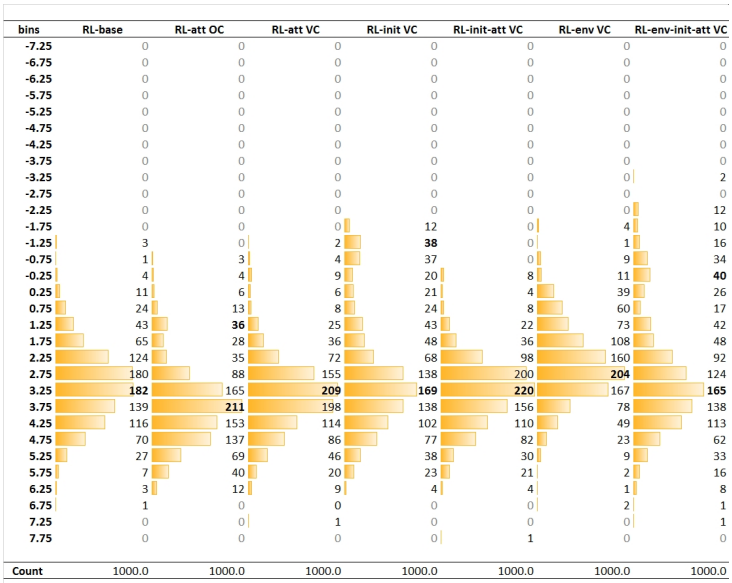
The RL-init-att configuration represents a middle ground and we see similar quality improvement to RL-att across setups (a gain of 2 BLEU points on average) but with a slightly lower latency. We however observe that RL-env-init-att has a slightly inferior performance with a a pronounced latency when compared to the RL-env model. We investigate this aspect in the next sections.

Investigating Average Lag. To further study the impact of our configurations on the sentence level lag, in Figure 2 we present the binned-histograms of sentence lags over the English→German test 2016 set. Generally, the models which are initialised with image information seem to have more mass towards the smaller delay bins. In terms of RL-init and RL-env-init-att setups, we also observe the presence of two modes around the lag value 3 as well as around two negative values (around -0.25 and -1.25 respectively). These negative lag values are due the difference in length between source and target sentences which is typical for the English→German. This also shows that the agent initialised with the image information tends to prefer WRITE actions with fewer READ actions. Further, on manual inspection of some samples, we observed that in the cases with negative lag the model begins with a WRITE action straight after reading the first token (See Table 2). As the agent is a GRU model, this behavior resembles that of an image captioning model. We also observe similar trends for English→French with RL-init models predominantly having more mass towards smaller delay bins (see Figure 3).

5.2 Agent Attention over Visual Inputs

In Figure 4 we visualize the agent's attention at each time step. On average, the agent actions correlate with the objects it attends to when producing the translation.

We now examine the general pattern of agent attention over the visual concepts across the four configurations using attention norm: a) RL-att-VC; b) RL-att-OC; c) RL-init-att; and d)



bins	RL-base	RL-att OC	RL-att VC	RL-init VC	RL-init-att VC	RL-env VC	RL-env-init-att VC
-7.25	0	0	0	0	0	0	0
-6.75	0	0	0	0	0	0	0
-6.25	0	0	0	0	0	0	0
-5.75	0	0	0	0	0	0	0
-5.25	0	0	0	0	0	0	0
-4.75	0	0	0	0	0	0	0
-4.25	0	0	0	0	0	0	0
-3.75	0	0	0	0	0	0	0
-3.25	0	0	0	0	0	0	2
-2.75	0	0	0	0	0	0	0
-2.25	0	0	0	0	0	0	12
-1.75	0	0	0	12	0	4	10
-1.25	3	0	2	**38**	0	1	16
-0.75	1	3	4	37	0	9	34
-0.25	4	4	9	20	8	11	**40**
0.25	11	6	6	21	4	39	26
0.75	24	13	8	24	8	60	17
1.25	43	**36**	25	43	22	73	42
1.75	65	28	36	48	36	108	48
2.25	124	35	72	68	98	160	92
2.75	180	88	155	138	200	**204**	124
3.25	**182**	165	**209**	**169**	**220**	167	**165**
3.75	139	**211**	198	138	156	78	138
4.25	116	153	114	102	110	49	113
4.75	70	137	86	77	82	23	62
5.25	27	69	46	38	30	9	33
5.75	7	40	20	23	21	2	16
6.25	3	12	9	4	4	1	8
6.75	1	0	0	0	0	2	1
7.25	0	0	1	0	0	0	1
7.75	0	0	0	0	1	0	0
Count	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0	1000.0

Figure 2: Histogram of per sentence lag values in test 2016 English-German. Y axis shows mean values per bin. Bold highlights modes for each distribution.

`SRC`: the red car is ahead of the two cars in the background .
`REF`: das rote auto fährt vor den beiden autos im hintergrund .
'the red car goes before the both cars in the background'
`RL-init`: die person ist im begriff , die rote mannschaft auf dem roten auto versammelt .
'the person is in concept, that red manhood on the red car gathered'
`Actions`: 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 1 0 0 1
`BLEU`: 3.7
`LAG`: -1.875

Table 2: Example of a German VC `RL-init` setup sentence with a negative lag, where the model tends to write more before reading new words.

`RL-env-init-att`. The attention norm is simply the average ℓ_2 norm between two consecutive attention time-steps. This can help in measuring the average visual attention per time step for a given sentence. We then compare the attention norm distributions over all the sentences in the English→German test 2016 set for the four different agent attention configurations. We present the result in Figure 5. Overall, `RL-init` and `RL-att` models are significantly more peaky than the `RL-env-init-att`. This suggests that `RL-env-init-att` model is generally spread across the 72 visual concepts more uniformly than the other two models. This perhaps is one of the causes for the slightly inferior performance of the model. We hypothesise that further regularisation of the attention distribution can ameliorate this behavior and leave it as future work.

6 Conclusion

In this paper we presented the first thorough exposition of multimodal reinforcement learning strategies for simultaneous machine translation. We demonstrate the efficacy of visual information and show that it leads to adaptive policies which substantially improve over the deterministic and unimodal RL baselines. Our empirical results indicate that both agent-side and environment-side visual information can be exploited to achieve higher quality translations with lower latency.

Throughout the experimental journey, we observed that the optimisation of simultaneous machine translation for dynamic policies is non-trivial, due to the two competing objectives: translation quality versus latency. For unimodal simultaneous machine translation, RL approaches tend to achieve translation quality on par with the quality of the deterministic policies within the same average lag. We believe that the fundamental issue is related

bin	RL-base	RL-att OC	RL-att VC	RL-init VC	RL-init-att VC	RL-env VC	RL-env-init-att VC
-7.25	0	0	0	0	0	0	0
-6.75	1	0	0	0	0	0	0
-6.25	0	0	0	0	0	0	0
-5.75	0	0	0	0	0	0	0
-5.25	0	0	0	0	0	0	0
-4.75	0	0	0	0	0	0	0
-4.25	0	0	0	0	0	0	0
-3.75	0	0	0	1	0	0	0
-3.25	0	0	0	0	0	0	0
-2.75	0	0	0	0	0	0	0
-2.25	0	0	0	0	0	0	0
-1.75	1	2	0	1	0	0	0
-1.25	0	0	0	2	0	0	0
-0.75	1	1	1	3	0	2	0
-0.25	6	3	0	4	2	3	3
0.25	9	11	2	10	4	11	2
0.75	32	8	2	8	4	13	7
1.25	27	8	12	35	6	20	10
1.75	68	18	15	59	6	41	24
2.25	129	38	38	181	29	87	50
2.75	**202**	80	67	**242**	85	131	94
3.25	168	156	138	217	160	175	183
3.75	149	**192**	**189**	129	**189**	**205**	**215**
4.25	112	189	181	62	187	132	184
4.75	49	123	173	31	147	89	101
5.25	30	100	100	10	88	54	70
5.75	12	39	50	2	46	24	36
6.25	4	22	30	3	36	10	18
6.75	0	10	1	0	9	2	3
7.25	0	0	1	0	2	0	0
7.75	0	0	0	0	0	0	0
Count	1000.0	1000.0	1000.0	1000.0	1000.0	999.0	1000.0

Figure 3: Histogram of per sentence lag values for test 2016 English-French. Y axis shows mean values per bin. Bold highlights modes for each distribution.

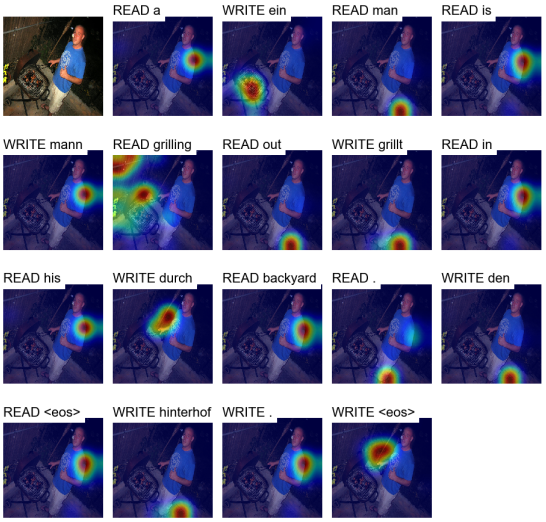


Figure 4: Visualisation of the agent attention and the corresponding actions over the source sentence from the test2016: 'A man is grilling out in his backyard.'

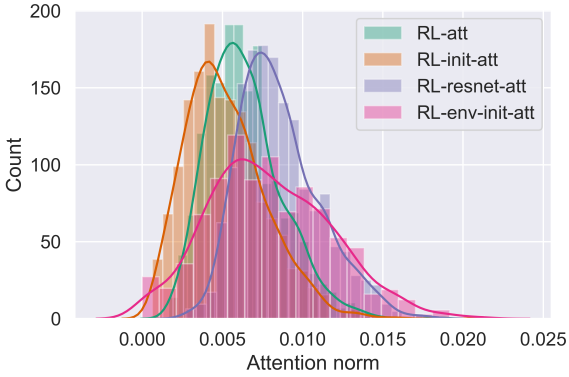


Figure 5: Distribution of attention norms for different agents with visual attention trained on the English→German dataset.

to the high variance of the estimator for sequence prediction, which increases sample complexity and impedes effective learning. On the other hand, the approaches with deterministic policies are simple and effective, as they are positively biased for language pairs that are close to each other. But the latter suffer from poor generalisation.

In the multimodal simultaneous machine translation setting, however, the variance of the estimator from RL models can be substantially reduced with to the presence of additional (visual) information.

Acknowledgments

The authors thank the anonymous reviewers for their useful feedback. This work was supported by the MultiMT (H2020 ERC Starting Grant No. 678017) project. The work was also supported by the Air Force Office of Scientific Research (under award number FA8655-20-1-7006) project. Andy Mingren Li was supported by the Imperial College London UROP grant.

References

Ashkan Alinejad, Maryam Siahbani, and Anoop Sarkar. 2018. Prediction improves simultaneous neural machine translation. In *Proceedings of the 2018 Con-*

ference on Empirical Methods in Natural Language Processing, pages 3022–3027, Brussels, Belgium. Association for Computational Linguistics.

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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Jacob Andreas, Dan Klein, and Sergey Levine. 2017. Modular multitask reinforcement learning with policy sketches. In *International Conference on Machine Learning*, pages 166–175.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020. Re-translation versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online. Association for Computational Linguistics.

Philip Arthur, Trevor Cohn, and Gholamreza Haffari. 2020. Learning coupled policies for simultaneous machine translation. *arXiv preprint arXiv:2002.04306*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *Computing Research Repository*, arXiv:1409.0473. Version 7.

Dzmitry Bahdanau, Felix Hill, Jan Leike, Edward Hughes, Arian Hosseini, Pushmeet Kohli, and Edward Grefenstette. 2018. Learning to understand goal specifications by modelling reward. *arXiv preprint arXiv:1806.01946*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

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

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? In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 627–633, Berlin, Germany. Association for Computational Linguistics.

Ozan Caglayan, Julia Ive, Veneta Haralampieva, Pranava Madhyastha, Loïc Barrault, and Lucia Specia. 2020. Simultaneous machine translation with visual context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2350–2361, Online. Association for Computational Linguistics.

Iacer Calixto, Desmond Elliott, and Stella Frank. 2016. DCU-UvA multimodal MT system report. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 634–638, Berlin, Germany. Association for Computational Linguistics.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405, Florence, Italy. Association for Computational Linguistics.

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

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA. Association for Computational Linguistics.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. 2018. Textworld: A learning environment for text-based games. In *Workshop on Computer Games*, pages 41–75. Springer.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2054–2063.

J. Deng, W. Dong, R. Socher, L. 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.

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, Copenhagen, Denmark. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Desmond Elliott and Ákos 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, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Prasoon Goyal, Scott Niekum, and Raymond J. Mooney. 2019. Using natural language for reward shaping in reinforcement learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2385–2391. International Joint Conferences on Artificial Intelligence Organization.

Alvin Grissom II, Naho Orita, and Jordan Boyd-Graber. 2016. Incremental prediction of sentence-final verbs: Humans versus machines. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 95–104, Berlin, Germany. Association for Computational Linguistics.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 603–611, Belgium, Brussels. Association for Computational Linguistics.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. 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, Valencia, Spain. Association for Computational Linguistics.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

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, Berlin, Germany. Association for Computational Linguistics.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

Aizhan Imankulova, Masahiro Kaneko, Tosho Hirasawa, and Mamoru Komachi. 2020. Towards multimodal simultaneous neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 594–603, Online. Association for Computational Linguistics.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In

Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73.

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, Vancouver, Canada. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. 2018. Low-latency neural speech translation. In *Proc. Interspeech 2018*, pages 1293–1297.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 290–298.

Koichiro Ryu, Shigeki Matsubara, and Yasuyoshi Inagaki. 2006. Simultaneous English-Japanese spoken language translation based on incremental dependency parsing and transfer. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 683–690, Sydney, Australia. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*.

Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2018. Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4213–4222.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019. 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, Hong Kong, China. Association for Computational Linguistics.