The Effects of Pretraining in Video-Guided Machine Translation

Ammon Shurtz, Lawry Sorenson, Stephen D. Richardson
Brigham Young University
Provo, UT
{acshurtz, lawrysorenson, srichardson}@byu.edu

Abstract

We propose an approach that improves the performance of VMT (Video-guided Machine Translation) models, which integrate text and video modalities. We experiment with the MAD (Movie Audio Descriptions) dataset, a new dataset which contains transcribed audio descriptions of movies. We find that the MAD dataset is more lexically rich than the VATEX (Video And TEXt) dataset (the current VMT baseline), and we experiment with MAD pretraining to improve performance on the VATEX dataset. We experiment with two different video encoder architectures: a Conformer (Convolution-augmented Transformer) and a Transformer. Additionally, we conduct experiments by masking the source sentences to assess the degree to which the performance of both architectures improves due to pretraining on additional video data. Finally, we conduct an analysis of the transfer learning potential of a video dataset and compare it to pretraining on a text-only dataset. Our findings demonstrate that pretraining with a lexically rich dataset leads to significant improvements in model performance when models use both text and video modalities.

Keywords: Machine Translation, Statistical and Machine Learning Methods, Word Sense Disambiguation

1. Introduction

Video-guided Machine Translation (VMT) is a subfield of Multimodal Machine Translation research that aims to translate spoken or written language aligned with video frames into another language (Wang et al., 2019). This task is challenging due to the complexity of video data and the need to integrate multiple modalities such as audio, visual, and textual cues. Multimodal Machine Translation has traditionally focused on combining images and text to obtain a more accurate translation while recently, VMT has had more attention from the research community (Chen et al., 2022; Gu et al., 2021; Li et al., 2022). Because of this, there is much room for improvement in the accuracy and efficiency of VMT models.

In this work, we address the challenges of VMT by experimenting with different datasets and a novel model architecture. Specifically, we focus on the MAD (Movie Audio Descriptions) dataset (Soldan et al., 2022). The MAD dataset contains transcribed audio descriptions of movies that are typically used for visually impaired individuals to understand the visual elements of a movie. To our knowledge, this dataset has not yet been exploited for VMT tasks and presents an opportunity to improve upon the current state-of-the-art models. We also employ the OpenSubtitles dataset as a large, text-only dataset to assess the potential of text-only pretraining for multimodal systems (Lison and Tiedemann, 2016).

In addition to these datasets, we propose a model architecture that takes advantage of a Conformer (Gulati et al., 2020), which is a convolutional transformer. The Conformer is designed to process variable-length sequences efficiently and has shown to be effective in a wide range of speech and language tasks such as Automatic Speech Recognition (ASR). By incorporating the Conformer into the VMT model architecture, we aim to enhance the performance of the model and contribute to the development of more effective multimodal machine translation systems.

These VMT models receive multimodal input in the form of tokenized source text to be translated and sub-sampled frames of videos (at 5 frames per second) in the form of pre-extracted features. Target text translation predictions are output by each model.

All code used in our experiments can be found at this Github repository[1].

2. Related Works

Multimodal machine translation was first explored in a shared task (Specia et al., 2016) where source language captions were translated into a target language using an image to increase the accuracy of the translation. Many other works have explored using image context to guide translation. One such work found that utilizing a decoder that attends to the image and source text separately outperformed decoders that attend to them together (Calixto et al., 2017).

The primary benchmark for Video-guided Machine Translation is the VATEX (Video And TEXt) dataset (Wang et al., 2019). This dataset consists of clips taken from YouTube videos with labels created by crowdsourced workers. Each video clip in the VATEX dataset has ten corresponding la-

[1] https://github.com/byu-matrix-lab/vmt-pretraining

LREC-COLING 2024, pages 15888–15898

bels in both English and Chinese, five of which are parallel translations from English to Chinese. It also contains pre-extracted video features using the pretrained I3D model by Carreira and Zisserman (2017). I3D was trained on the DeepMind Kinetics human action video dataset (Kay et al., 2017). VATEX was the subject of the 2020 VMT Challenge.

The winning model for this challenge was submitted by Hirasawa et al. (2020), which is based on keyframe-based feature extraction and positional encoding for video contexts. Their model uses hierarchically-attentive RNNs as its base. This model outperformed transformer-based architectures by using an ensemble of video features. However, the video context tended to only marginally improve the performance.

Recent work by Yang et al. (2022) has found that the short and straightforward labels in the VATEX dataset are able to produce accurate translations alone, without relying on the video for support. They propose that future VMT datasets should focus on the problem of live translation of languages with different structures such that incomplete utterances can be complemented using video context. In particular, they suggest creating VMT datasets that address linguistic particularities such as word order and gender marking.

The transcriptions of movie audio descriptions in the MAD dataset (Soldan et al., 2022) represent another source of language that is strongly correlated with video (Rohrbach et al., 2017). These transcriptions present large contexts for machine translation tasks, making them also relevant for context-aware NMT (Neural Machine Translation) tasks, such as lexical cohesion (Lyu et al., 2022). The MAD dataset only contains transcriptions for English, and is presented primarily for video captioning tasks. MAD video features were extracted using CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021).

State-of-the-art MT (Machine Translation) models are based on the Transformer architecture (Vaswani et al., 2017). The Conformer was proposed by Gulati et al. (2020) for the task of Automatic Speech Recognition (ASR). It takes advantage of the attention mechanism while also using convolutions in order to extract audio features (or in our case, visual features) to be attended to. The authors reported state of the art performance, and competitive performance with a smaller model. It has not been previously applied to the task of video-guided machine translation as a video encoder, though it has been used in Speech Translation (Jia et al., 2021).

In March 2023, Han et al. (2023) released MAD-v2, an enhanced iteration of the MAD (Movie Audio Descriptions) dataset. This updated version exhibits a reduced level of noise compared to its predecessor, potentially leading to improved outcomes in various applications. The dataset's improved quality holds promise for generating more accurate and reliable results. In July 2023, Kang et al. (2023) released BigVideo, a large, multimodal VMT dataset which includes a test set that contains ambiguous words which can be resolved by video context. We concluded this work before the release of either of these datasets, therefore leave the use of them to future work.

3. Methodology

3.1. Data

In our experiments, we utilize MAD[2], OpenSubtitles[3], and VATEX[4] for training. We only evaluate on the VATEX data set.

We focus on the English to Chinese VMT task, since that is the language pair of the VATEX dataset. However, as the MAD dataset is only in English, we pre-translate the English transcriptions to Chinese using Google translate (Google, 2023). The initial translations are generated without the benefit of visual context. Despite the flaws of Google Translate's machine translations, we opt to use them for the sake of increasing the amount and lexical diversity of both source and target language pretraining data. Furthermore, the machine-translated text aligns with the video embeddings from the MAD dataset, which proves valuable in providing contextual information during training. We note that this technique of training on machine-translated sentences has been previously used for knowledge distillation (Kim and Rush, 2016).

To control for noise in the dataset due to potential inaccuracies of the English ASR transcriptions and Google's machine translations, we run the reference-less metric COMET-QE (Rei et al., 2020) on the translation pairs and filter out training data pairs with a score equal to zero, keeping only segments with a positive score. The MAD dataset is thus reduced to approximately 69% of it's original size, filtering out 87,053 of the 280,183 sentences.

The MAD dataset was originally created for video captioning tasks, and the names of movie characters were excluded from its validation and test sets. A purely visual model would have no context for predicting names (the names are replaced with the token "SOMEONE"). For the task of VMT, the source sentence *can* provide the character names that will be translated into the target language. Because of this, we rely solely on the MAD training set, incorporating character names instead of the

[2] https://github.com/Soldelli/MAD

[3] https://opus.nlpl.eu/OpenSubtitles-v2018.php

[4] https://eric-xw.github.io/vatex-website/index.html

	Train		Validation		Test	
	Videos	Text	Videos	Text	Videos	Text
MAD	183,130	183,130	10,000	10,000	-	-
VATEX	25,991	129,955	1,500	7,500	1,500	7,500
OpenSubtitles	-	10,643,121	-	560,165	-	-

Table 1: Size of each dataset by number of unique videos and text segments.

generic "SOMEONE" tokens. We randomly selected 10,000 sentences from the filtered MAD training set to use as a validation set when training.

Since the VATEX dataset did not publicly release translations for the test set, we follow the example of Li et al. (2022) and Chen et al. (2020) in randomly splitting the validation set equally to provide validation and test datasets. For our experiments, we only draw from the five aligned English-Chinese translation pairs for each video clip in the VATEX dataset, ignoring the other segments in each language that are not translations of one another. Each video clip and the five aligned translation pairs associated with it are only in the validation set or the test set, but not both.

We used the OpenSubtitles English to Chinese data for text-only pretraining experiments. We used the data as is, and randomly selected five percent of the segments to use in the validation set when pretraining.

Table 1 gives the number of videos and translated segments for each dataset. Note that the filtered MAD dataset adds many video contexts for training beyond what is available solely in the VATEX dataset, while the OpenSubtitles dataset provides exponentially more text segments.

Since Yang et al. (2022) concluded that the simpler sentences in the VATEX dataset do not require the videos for more accurate translations, we explore the lexical richness of the MAD dataset compared to VATEX, believing that models will benefit from training on data with greater lexical diversity. We use two lexical diversity metrics implemented in the LexicalRichness package (Shen, 2022) for comparisons: Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis, 2010) and Mean Segmental Type-Token Ratio (MSTTR) (Johnson, 1944). Alongside MAD and VATEX, we incorporate two supplementary datasets for comparison: MSR-VTT (Xu et al., 2016), which was previously evaluated against VATEX by Wang et al. (2019), and the OpenSubtitles English to Chinese corpus (Lison and Tiedemann, 2016), serving as a large text-only corpus baseline.

Table 2 provides lexical richness scores. Consistent with the findings of Wang et al. (2019), the MSR-VTT scores demonstrate lower MTLD and MSTTR scores compared to VATEX. However, both OpenSubtitles and MAD are more lexically diverse than VATEX, with MAD obtaining the highest scores among all the datasets.

	MTLD	MSTTR
MAD	86.515	0.709
OpenSubtitles	64.473	0.688
VATEX	32.724	0.527
MSR-VTT	26.739	0.507

Table 2: Comparison of lexical richness scores Measure of Textual Lexical Diversity (MTLD) and Mean Segmental Type-Token Ratio (MSTTR) across four datasets: Movie Audio Descriptions (MAD), OpenSubtitles, Video And TEXt (VATEX), and Microsoft Research Video to Text (MSR-VTT)

Due to the copyright of the movies in the MAD dataset, Soldan et al. (2022) do not include the raw videos. Instead, the authors include features generated by CLIP, which represents images and text jointly in the same space (Radford et al., 2021). Like in the MAD dataset, Wang et al. (2019) do not provide the raw videos in the VATEX dataset. Instead, they offer pre-generated features and YouTube video IDs. The provided features are generated by the pretrained I3D (Inflated 3D Convnet) model, which inflates a 2D ConvNet for classification into a temporal third dimension (Carreira and Zisserman, 2017).

Because of the disparity between the provided I3D features of VATEX and the CLIP features of MAD, we retrieve the raw videos from the VATEX dataset collected using the YouTube video IDs and then extract the CLIP features from those videos using the pretrained CLIP model[5] which was used on MAD. We use these CLIP features for all experiments. We remove any segments for which the video is no longer available[6], leaving us with 88.0% of the training set, and 88.6% of the validation set. These features are available in our Github repository[7].

[5] https://huggingface.co/openai/clip-vit-large-patch14

[6] We downloaded the VATEX dataset videos in June 2023

[7] https://github.com/byu-matrix-lab/vmt-pretraining/releases/tag/vatex-clip-features

3.2. Architecture

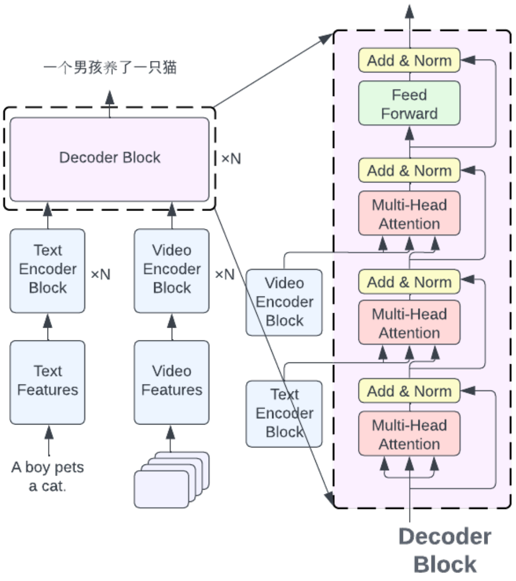

Figure 1: The high-level architecture used for each model. On the left, the overall architecture mainly consists of N text encoder blocks, N video encoder blocks, and N decoder blocks. The decoder block is shown in greater detail on the right side of the figure. It is similar to work by Vaswani et al. (2017) but contains an added attention layer for video attention. It takes a source sentence and video (at 5 frames per second) as input and it outputs the sentence in the target language.

Each model used in our experiments is an encoder-decoder model utilizing a doubly attentive Transformer as inspired by (Calixto et al., 2017). The decoder separately attends to the video encoding, the source text encoding, and itself. This architecture enables control of what context is passed to the decoder, as both encoder blocks have the option of being included or excluded during the various experiments. Figure 1 shows the overall architecture of our model.

A Transformer (Vaswani et al., 2017) text encoder is used in each experiment. We experiment with both a Conformer (Gulati et al., 2020) video encoder and a Transformer video encoder. The Conformer encoder is a convolution-augmented Transformer which takes advantages of a convolutional neural network's ability to exploit local features while maintaining the Transformer's ability to exploit global features.

We use six encoder and decoder layers for all experiments. The CLIP features of a 5-frame-per-second video are the input for the video encoder. We use the CLIP L14 features, which have a size of 768 dimensions. We add a projection layer before the video encoder that projects the features onto 512 dimensions to match the d_model shape of the text encoder and decoder. This projection layer is exchanged after pretraining on MAD to match the size of the VATEX dataset input. The output of the decoder is a translated text segment.

We employed two baseline models for comparison. The first baseline model has a text-only Transformer-based architecture (Vaswani et al., 2017) that does not incorporate video features. In addition, we utilized the code provided by Wang et al. (2019) to train their multimodal sequence-to-sequence model, which we refer to as the VATEX baseline. The VATEX baseline consists of LSTM encoders and decoders, enabling the model to process both textual and visual information.

All of our models are implemented using HuggingFace and PyTorch. We use BARTForConditionalGeneration (Lewis et al., 2020) as the base for the Transformer and Wav2Vec2Conformer (Wang et al., 2020) for the Conformer implementation.

See additional experiment details in Appendix A.

3.3. Experiments

In our experiments, we assess the performance of each model using two metrics: BLEU, calculated with SacreBLEU (Post, 2018), and COMET (Rei et al., 2022). For computing the BLEU scores, we employ default SacreBLEU parameters along with Chinese tokenization. To evaluate using COMET, we utilize the latest COMET model, wmt22-comet-da. We conduct statistical significance testing on the COMET scores using their dedicated statistical testing tool, which employs the Paired T-Test and bootstrap methods proposed by Koehn (2004).

Although the original VMT task was evaluated using only BLEU scores, we evaluated the proposed architectures and the VATEX baseline using both BLEU and COMET because recent work has shown the latter to be superior (Freitag et al., 2022). We note that the wmt22-comet-da model has been fine tuned on the English to Chinese translation evaluation task from the WMT 20 DA (direct assessment) data (Rei et al., 2022). This COMET model produces scores between 0 and 1, where higher scores indicate better performance.

In our initial experiment, we compared the performance of each model architecture when pretrained on MAD or OpenSubtitles with their performance when trained from scratch on the VATEX dataset.

Following the approach of Yang et al. (2022), we conducted experiments with both pretrained and non-pretrained models, considering different proportions of masked tokens in the source sentence. We use two approaches to masking: randomly selecting tokens and masking the end of the sentence. Specifically, we masked 15% and 30% of each

Model	COMET	BLEU
VATEX (baseline)	0.7227	30.4
Text-only Transformer	0.7302	29.7
Text-only Transformer - MAD pretrained	0.7569*	34.8
Transformer Video Encoder	0.7337	30.3
Transformer Video Encoder - MAD pretrained	0.7573*	**35.0**
Conformer Video Encoder	0.7344	29.4
Conformer Video Encoder - MAD pretrained	**0.7590***	34.6

Table 3: English to Chinese results of each model architecture on the VATEX test set, when trained solely on the VATEX train set vs. when pretrained on MAD and then fine-tuned on VATEX. COMET and BLEU scores are shown. Models marked with an asterisk (*) have significantly different ($p < 0.05$) COMET scores compared to other models of the same architecture. In this test, there was no significant difference between our architectures when trained on the same data.

source sentence and compared the resulting BLEU and COMET scores. Masking was accomplished by replacing tokens in the source with a <mask> token. This approach aligns with the suggestion of Yang et al. (2022) to focus on simultaneous translation tasks where the complete source sentence is unavailable. We apply masking both at training time and inference time for these experiments.

We note that Yang et al. (2022) has already conducted an extensive error analysis for VMT translations of masked sentences using the VATEX dataset. We chose to focus our experiments on comparing the effects of pretraining, and we do not include a similar analysis of our own results.

In each experiment, we evaluated the following models:

- VATEX (baseline)
- Text-only Transformer
- Transformer Video Encoder
- Conformer Video Encoder

We also evaluated pretraining on OpenSubtitles in our experiments to assess the effect of pretraining a multimodal system on a large text-only dataset.

The VATEX baseline was trained using the code provided by its authors with all the default model parameters.[8] All models were trained to convergence using half of the split VATEX validation set, as described in Section 3.1 above. See Appendix A for more training details.

4. Results

4.1. MAD Pretraining

We evaluated the performance of the Conformer and Transformer video encoder architectures by comparing their COMET and BLEU scores with those of the VATEX baseline and the text-only Transformer baseline. Additionally, we evaluated the effectiveness of each model architecture when pretrained on MAD video and text data, followed by fine-tuning on VATEX. Results are shown in Table 3.

The VATEX baseline model produces output with the lowest COMET score, which has a statistically significant difference from all MAD pretrained COMET score results. Each model pretrained on MAD outperforms the same model trained only on VATEX based on both BLEU and COMET scores (increasing by an average of 5 BLEU points and a COMET score of 0.025). The Transformer and Conformer video encoder COMET scores do not show any statistically significant difference between each other.

4.2. Masked Source

We proceeded to evaluate the previously mentioned models when masking source sentences according to the percentages mentioned previously (see section 3.3). For simplicity, results of all masking experiments are presented in Table 4, which also includes the delta improvement in COMET and BLEU scores when comparing pretraining on the MAD dataset to no pretraining. Random masking and end masking results are separated.

As can be seen in Table 4, the overall improvement gained from pretraining on MAD decreases as the percentage of masked tokens increases for all model architectures. We also observe that when masking the end of source segments, the text-only transformer model has the largest COMET and BLEU performance gain compared to the two model architectures that take video context into account. We hypothesize that the video context is adding noise, which is harming the models' ability predict the masked token translations. Because

[8] https://github.com/eric-xw/Video-guided-Machine-Translation

End Masking	No Masking		15% Masking		30% Masking	
	COMET	BLEU	COMET	BLEU	COMET	BLEU
Text-only Transformer	0.7302	29.7	0.7018	26.7	0.6786	23.6
Text-only Transformer - MAD	0.7569	34.8	0.7275	30.5	0.6985	26.7
Δ	**0.0267**	5.1	**0.0257**	3.8	**0.0199**	**3.1**
Transformer Video Encoder	0.7337	30.3	0.7129	26.5	0.6941	24.8
Transformer Video Encoder - MAD	0.7573	35	0.7275	30.5	0.7011	27.1
Δ	0.0236	4.7	0.0146	**4**	0.007	2.3
Conformer Video Encoder	0.7344	29.4	0.7157	27.2	0.6985	25
Conformer Video Encoder - MAD	0.759	34.6	0.7375	31.1	0.7147	27.4
Δ	0.0246	**5.2**	0.0218	3.9	0.0162	2.4

Random Masking	No Masking		15% Masking		30% Masking	
	COMET	BLEU	COMET	BLEU	COMET	BLEU
Text-only Transformer	0.7302	29.7	0.711	25.6	0.6792	21.3
Text-only Transformer - MAD	0.7569	34.8	0.7349	30.8	0.7092	25.4
Δ	**0.0267**	5.1	0.0239	**5.2**	**0.03**	**4.1**
Transformer Video Encoder	0.7337	30.3	0.7056	25.9	0.7028	22.9
Transformer Video Encoder - MAD	0.7573	35	0.7345	30.4	0.7064	25.4
Δ	0.0236	4.7	**0.0289**	4.5	0.0036	2.5
Conformer Video Encoder	0.7344	29.4	0.7161	26.1	0.7007	23
Conformer Video Encoder - MAD	0.759	34.6	0.7414	31.1	0.7267	26.6
Δ	0.0246	**5.2**	0.0253	5	0.026	3.6

Table 4: English to Chinese language direction. Δ Change in COMET/BLEU Score by video encoder architecture with masked source sentences. Rows with "MAD" indicate the MAD pretrained experiments. For random masking, masked tokens are picked uniformly from across the source sentence, while for end masking, the final tokens in the sentence are masked. The highest changes in score are bolded. For all models, pretraining consistently improved model performance. We note that the first column, No Masking, contains the same results as Table 3, but they are included here for reference.

Model	COMET	BLEU
Text-only - Text-only MAD pretrained	0.7569	34.8
Text-only - Text-only OpenSubtitles pretrained	0.7662*	36.4
Transformer Video Encoder - Text-only MAD pretrained	0.7576	34.6
Transformer Video Encoder - MAD with Video pretrained	0.7573	35.0
Transformer Video Encoder - Text-only OpenSubtitles pretrained	0.7675*	**36.8**
Conformer Video Encoder - Text-only MAD pretrained	0.7582	34.8
Conformer Video Encoder - MAD with Video pretrained	0.7590	34.6
Conformer Video Encoder - Text-only OpenSubtitles pretrained	**0.7684***	36.5

Table 5: English to Chinese results comparing pretraining on text-only datasets with pretraining using video context. Models marked with an asterisk (*) have significantly different ($p < 0.01$) COMET scores compared to other models of the same architecture. Significance scores were computed with Tukey HSD. (Tukey, 1949)

of these results, we conducted an ablation study to investigate the overall effect of pretraining with video (see Section 4.5).

4.3. Text-Only Pretraining

In order to see the effect of video data pretraining, we compare different pretraining datasets for the English to Chinese task: MAD video and text, MAD text only, and OpenSubtitles text. Results are shown in Table 5.

We also conduct reference-based human ranking evaluations of the same pretraining groups for the Chinese to English direction. The evaluation analyzed Chinese to English translations due to the local availability of native English speakers. This

Type of Masking	No Pretraining	MAD with video	Text-only MAD	OpenSubtitles
No Masking	2.4550	2.0600*	2.0150*	**1.8500***
Mask 30 End	2.5050	2.4150	2.3650	**2.1200***
Mask 30 Rand	2.5700	2.5050	2.2600	**2.1200***

Table 6: Chinese to English results of reference-based ranking human evaluation comparing different pretraining strategies. Sentences were rated from 1 to 4, where 1 is the best and 4 is the worst, and ties were allowed. Results marked with an asterisk (*) have significantly different ($p < 0.01$) scores compared to models without any pretraining.

assessment is performed for models trained with no masking, the last 30% of the sentence masked, or 30% of the words in the sentence randomly masked. For each of the listed groups, we randomly select 50 different segments translated by both Conformer Video Encoder models and Transformer Video Encoder models for a total of 100 segments per group. We combined the two model architecture predictions due to all validation and test sets resulting in similar scoring across all automatic metrics. We had two native English speakers evaluate sentences produced by models with the four pretraining strategies. The evaluators were given the target English sentence and asked to rate the translations by how well they conveyed the meaning of the original sentence. The averages of these ratings by category are presented in Table 6.

On the OpenSubtitles dataset for English to simplified Chinese (which has around 11M segments), we found significant performance gains for all models. The Conformer video encoder model, pretrained on OpenSubtitles text and fine-tuned on the VATEX video and text training set, achieved the highest COMET score on our VATEX test set and had the best human evaluation ranking in Table 6. This result highlights the efficacy of pretraining with a substantial text corpus and shows the advantages for a multimodal machine translation system when it is fine-tuned on its distinct modalities, such as video and text. Additionally, pretraining with MAD video and text data does not show significant improvements over pretraining with just MAD text data.

4.4. Related Work Comparison Validity

Although Hirasawa et al. (2020) achieved the top-performing model in the 2020 VMT challenge, we were unable to directly compare our results with theirs. This limitation arose due to the unavailability of their model code and the utilization of a private test set that differs from our own. However, it is important to note that we did employ the same validation set for evaluation purposes, although we sampled only half of the validation set compared to their usage of the entire set. The remaining half of the validation set was reserved as our independent test set.

Architecture	ΔCOMET	ΔBLEU
Comformer	0.0013	0.0000
Transformer	-0.0165	-0.8571*

Table 7: Ablation study of the effects of pretraining with CLIP embeddings from MAD compared to training with only the text from MAD. Negative values indicate that including the videos decreased performance. The asterisk (*) indicates that the transformer BLEU score is significant with $p < 0.01$.

4.5. Ablation Study

We ran an ablation study to isolate the effects of including the CLIP embeddings when pretraining on the MAD dataset. The average change in the COMET and BLEU scores across all tests by video encoder architecture is depicted in Table 7. Note that both average effects tend to be negative or close to zero.

As there was not a significant effect from pretraining with CLIP video embeddings, the pretraining improvements may be attributed to the additional text from the MAD dataset.

5. Discussion

In all of our tests there was not a significant difference between the Conformer and Transformer video encoders. Since our models process precomputed video features, it would be interesting to see if there is any advantage to using the Conformer on raw video input, where we would expect the convolutions to be more useful.

While it is not possible to conduct a direct comparison between our results and the winning model of the 2020 VMT challenge, it is worth noting that our text-only transformer model, pretrained on OpenSubtitles, achieved a BLEU score of 36.4 without any video context on its validation set. This score is within 0.1 BLEU of the winning model's reported validation score of 36.48 (Hirasawa et al., 2020), with the assumption that the random subset of the validation set we used is representative of the full validation set. This strongly supports the findings of Yang et al. (2022), which concludes that the text in the VATEX dataset is sufficiently simple to translate without video context.

As can be seen across every experiment, pretraining with the video features does not seem to improve the models performance. We suspect that the video encoders employed were overly complex given that they started with high-level CLIP embeddings. Providing the raw videos to the video encoders could help improve the utilization of the video context for translation. As it is, we hypothesize the lack of improvement to be attributed to the substantial amount of noise present in the video data, which may pose challenges for the models in effectively discerning and learning relevant information from the video frames. This behavior could also have been caused by the VATEX dataset itself not needing additional video context to produce accurate translations.

We suggest that video context may show greater promise when translating from a less marked language to a more highly marked language by helping in the disambiguation of the more marked words and phrases. The English and Mandarin Chinese of VATEX may not be the most useful languages for evaluating the ability of video context to disambiguate words and phrases. We encourage further research in VMT between languages that would benefit more from the potential disambiguation provided by video context.

As a whole, our results show the potential of video-guided or multimodal translation systems to benefit from pretraining on large, varied text corpora. These corpora are simpler to build than multimodal corpora, yet can yield significant improvements in the final system. We believe that pretraining on a large text corpus and finetuning on a small multimodal dataset that has been cleaned and curated for variety will yield the best results when training VMT systems.

6. Conclusion

We have proposed an approach to Video-guided Machine Translation (VMT) that improves the performance of MT models that integrate text and video modalities. We evaluated the benefits of pretraining on the multimodal Movie Audio Descriptions (MAD) dataset and the large text-only OpenSubtitles corpus, finding that pretraining a VMT model on a large amount of text before finetuning to the specific video and text modalities yields significant improvement to both COMET and BLEU scores.

In our experiments, we showed that each proposed model, including the text-only baselines, outperforms the LSTM-based model employed by Wang et al. (2019) based on COMET scores. When pretraining on the multimodal MAD dataset, we see large gains in performance for each model architecture. We also showed that pretraining on text-only datasets can improve models that are fine-tuned on multiple modalities. In some cases, we observed greater improvements in COMET and BLEU scores when employing text-only pretraining as opposed to pretraining with both text and video. Possible reasons for this are discussed in section 5 above.

Masking source tokens can be useful for seeing how well a VMT model leverages video features. We show that for randomly masked tokens and end-of-sentence masking, both architectures that include video context generally achieve higher COMET and BLEU scores as the percentage of masked tokens increases. In the case of end-of-sentence masking, these results show the potential benefit of using a video encoder for simultaneous MT, where the incomplete translation of a live video could leverage the video context to generate a hypothesis translation.

7. Future Work

We suspect that the video encoders employed were overly complex given that they started with high-level CLIP embeddings. We would like to repeat experiments with a simpler video encoder to reduce the noise that the video adds to the translation. Using a conformer as a video encoder may benefit from being trained on raw video data rather than pre-computed video features, and experiments comparing the raw video and the pre-computed video features could yield informative results.

As mentioned in Section 2, the BigVideo dataset (Kang et al., 2023) has the potential to more accurately evaluate VMT systems and their ability to disambiguate semantically ambiguous segments. Evaluating pretraining methods on the BigVideo dataset and performing a comparative error analysis are important next steps.

Our research could be extended to several other domains, including multilingual video captioning, unsupervised machine translation, and simultaneous translation. We anticipate that VMT model architectures that excel in handling masked data will also yield superior performance in both unsupervised and simultaneous translation tasks. Furthermore, we believe that the updated MAD dataset (Han et al., 2023) will enable significant progress in exploring in these areas.

8. Limitations

In our assessment of the VATEX dataset, several errors and inconsistencies have been identified. Notably, a considerable portion of the VATEX dataset contains phrases such as "music playing in the background," "as a unseen voice narrates," "a person talks about," etc. These textual cues, unfortunately, could not be effectively leveraged by the video model due to the absence of corresponding audio data within the video features. Furthermore, several textual descriptions within the dataset did not faithfully depict the actual events in the videos. Consequently, we are unsure of the reliability of the VATEX dataset as a benchmark for assessing VMT models.

Addressing these limitations is crucial for refining the evaluation of VMT models. The utilization of a more refined version of VATEX, achieved through rigorous cleaning and verification, or the introduction of an entirely new dataset dependent on video context for disambiguation, holds the potential to be a more accurate and dependable metric for gauging the performance of VMT.

As stated previously, using Google Translate to generate translations for MAD could decrease each model's ability to translate accurately. Using human post-edited translations for the MAD dataset may improve pretraining results due to the higher quality translations in a post-edited dataset.

9. Ethical Considerations

We publish our findings with the goal of improving the quality of translations that can benefit from visual context, and we encourage the community to consider situations were Video-Guided Machine Translation can improve the availability of resources in global societies.

We also note that many of the datasets used in our research were gathered from existing content (movies and YouTube videos) without the express consent of the creators. This issue is widely prevalent when gathering datasets for machine learning research. In our case, we note that none of the datasets used distribute the original content, and that the models created do not compete with the content creators.

10. Bibliographical References

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. *arXiv preprint arXiv:1702.01287*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Shiyu Chen, Yawen Zeng, Da Cao, and Shaofei Lu. 2022. Video-guided machine translation via dual-level back-translation. *Knowledge-Based Systems*, 245:108598.

Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Google. 2023. Google translate. Accessed: April 24, 2023.

Weiqi Gu, Haiyue Song, Chenhui Chu, and Sadao Kurohashi. 2021. Video-guided machine translation with spatial hierarchical attention network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 87–92.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023. AutoAD: Movie description in context. In *CVPR*.

Tosho Hirasawa, Zhishen Yang, Mamoru Komachi, and Naoaki Okazaki. 2020. Keyframe segmentation and positional encoding for video-guided machine translation challenge 2020. *arXiv preprint arXiv:2006.12799*.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2021. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. *arXiv preprint arXiv:2107.08661*.

Wendell Johnson. 1944. Studies in language behavior: A program of research. *Psychological Monographs*, 56(2):1–15.

Liyan Kang, Luyang Huang, Ningxin Peng, Peihao Zhu, Zewei Sun, Shanbo Cheng, Mingxuan Wang, Degen Huang, and Jinsong Su. 2023. Bigvideo: A large-scale video subtitle translation dataset for multimodal machine translation. *arXiv preprint arXiv:2305.18326*.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Mingjie Li, Po-Yao Huang, Xiaojun Chang, Junjie Hu, Yi Yang, and Alex Hauptmann. 2022. Video pivoting unsupervised multi-modal machine translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Xinglin Lyu, Junhui Li, Shimin Tao, Hao Yang, Ying Qin, and Min Zhang. 2022. Modeling consistency preference via lexical chains for document-level neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6312–6326.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Matt Post. 2018. A call for clarity in reporting bleu scores. *WMT 2018*, page 186.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision*, 123(1):94–120.

Lucas Shen. 2022. LexicalRichness: A small module to compute textual lexical richness.

Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2022. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5026–5035.

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.

John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. 2020. fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *The IEEE International Conference on Computer Vision (ICCV)*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Zhishen Yang, Tosho Hirasawa, Mamoru Komachi, and Naoaki Okazaki. 2022. Why videos do not guide translations in video-guided machine translation? an empirical evaluation of video-guided machine translation dataset. *Journal of Information Processing*, 30:388–396.

A. Experiment Details

Table 8 presents the training hyperparameters used in all of our experiments. Every model was trained on an NVIDIA A100 80GB GPU. We use version 0.5.1 of the LexicalRichness package (Shen, 2022).

Table 9 presents the approximate average runtime of training and inference for each model. Note that the inference and training runtimes include pretraining on MAD, pretraining on OpenSubtitles, and training on VATEX.

Hyperparameter	Value
Video Encoder Architecture	Conformer, Transformer
Text Encoder Architecture	Transformer
Decoder Architecture	Transformer
Feed Forward Network Dim	2048
Attention Heads	8
Video Encoder Layers	6
Text Encoder Layers	6
Decoder Layers	6
Optimizer	Adam
Activation Function	Swish
Learning Rate	1e-4
Batch Size	32
Dropout Rate	0.1

Table 8: Hyperparameters of each proposed model: Conformer Video Encoder, Transformer Video Encoder, and Text-only Transformer

Model		Average Runtime
VATEX (Baseline)	Training	11.5 total hours
	Inference	-
Text-only	Training	10 total hours
	Inference	603 tokens/sec
Transformer	Training	13 total hours
	Inference	570 tokens/sec
Conformer	Training	12 total hours
	Inference	566 tokens/sec

Table 9: "Text-only" refers to the Text-only Transformer model, "Transformer" refers to the Transformer Video Encoder architecture, and "Conformer" refers to the Conformer Video Encoder architecture. The approximate average runtime includes pretraining on MAD or OpenSubtitles and training on VATEX, all to convergence. Inference is tokens per second. Training and Inference runtimes are reported separately.