The Effects of Pretraining on Video Guided Machine Translation

Ammon Shurtz, Lawry Sorenson, Stephen D. Richardson

LREC-COLING 2024

BYU

Research Questions:

How can pretraining strategies improve Video-Guided Machine Translation (VMT)?

What modalities make the biggest impact for pretraining strategies in VMT?

BYU

What is Video-Guided Machine Translation?

- Video-Guided Machine Translation provides video context to aid in translation.
- This can help the model disambiguate between potential meanings of the sentence.

BYU

VaTeX Dataset

- Approx. 30,000 YouTube video clips labeled in English and Chinese by Amazon Mechanical Turk.
- 10 labels per video in both languages, 5 of which are translations of each other.
- Used for a shared task (VMT Challenge held at the ALVR Workshop) in 2020.

BYU

VaTeX Example

A group of dogs of varying breeds pull a man in a sled across the finish line.

https://www.youtube.com/watch?v=UZLSnJ2OBYA&ab_channel=grandylady

BYU

VaTeX Example

Original Sentence:

A group of dogs of varying breeds pull a man in a sled across the finish line.

Masked input sentence:

A group _____ of varying _____ pull ___ man ___ a sled across the finish line.

Produced translation:

一群不同品种的狗拉着一个乘坐雪橇的人穿过了终点线。

Backtranslated by Google Translate:

A pack of dogs of various breeds pulls a man on a sled across the finish line.

BYU

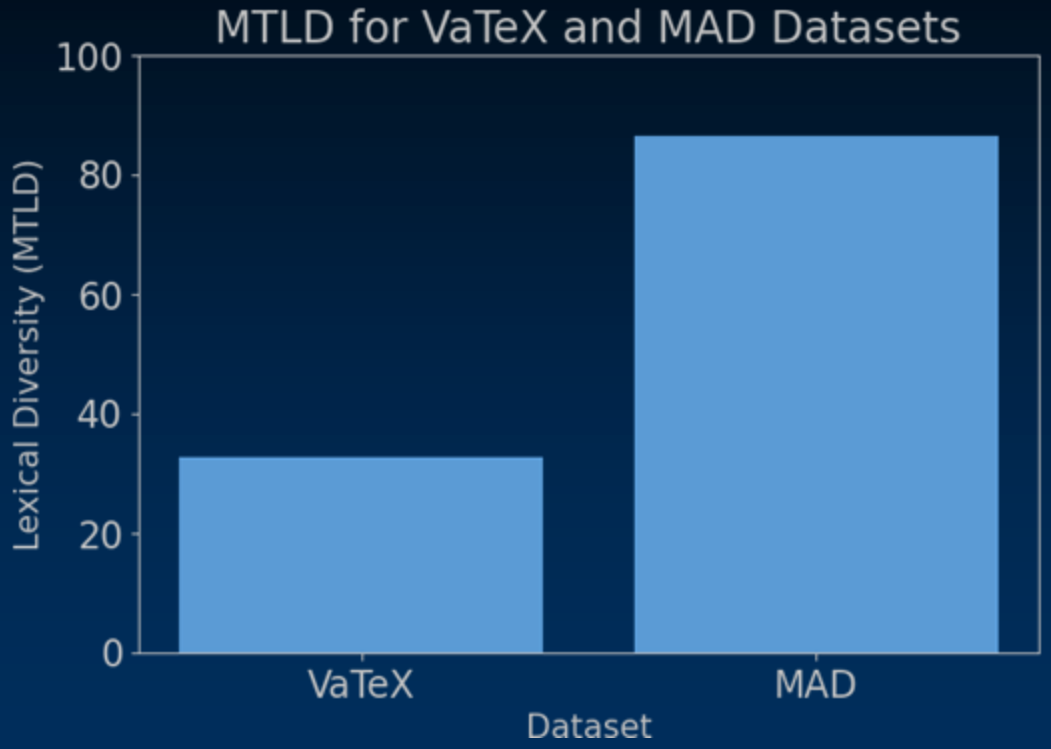
Movie Audio Description (MAD) Dataset

- 600 Movies ~ 180k sentences
- Transcribed English Audio Descriptions (not dialogue)
- We created synthetic Chinese translations using Google Translate.
- Higher quality and more variety than the VaTeX dataset

BYU

Lexical Richness

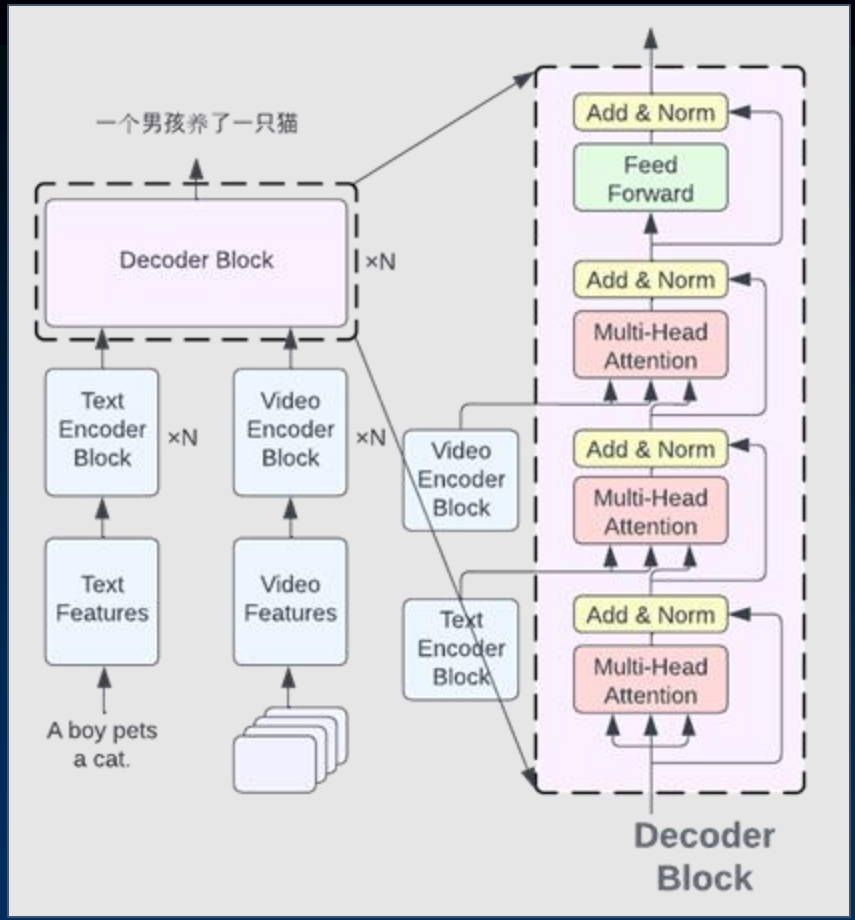
MTLD (Measure of Textual Lexical Diversity) reflects the amount of variety of words in text.



BYU

Architecture

Transformer architecture with separate video and text encoders, decoder has dual attention.

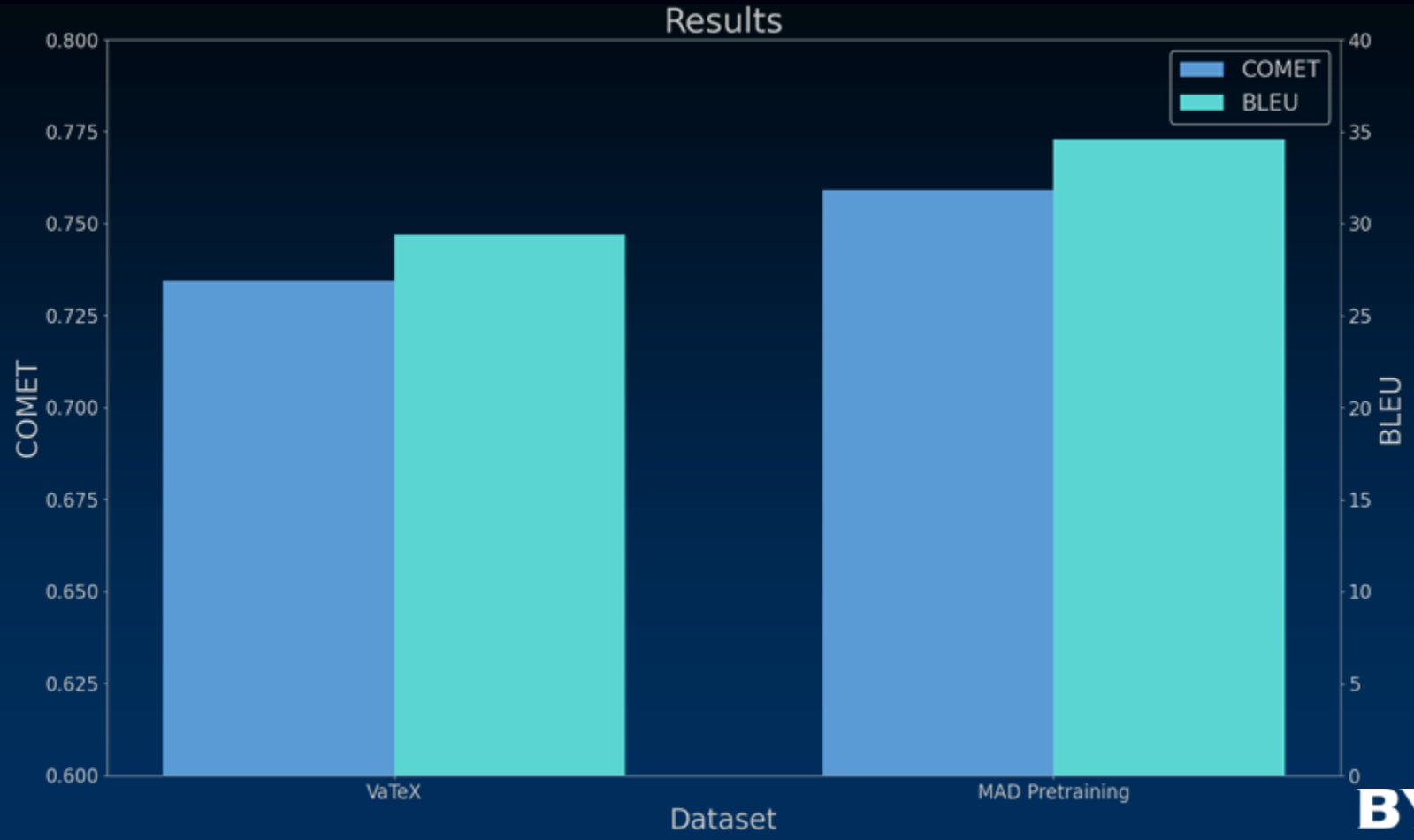


BYU

Experiments

- Try pretraining on the MAD dataset.
- Finetune on the VaTeX dataset.
- Test performance on the VaTeX dataset.

BYU


Results
COMET
BLEU
0.800
0.775
0.750
0.725
0.700
0.675
0.650
0.625
0.600
40
35
30
25
20
15
10
5
0
VaTeX
MAD Pretraining
Dataset

BYU

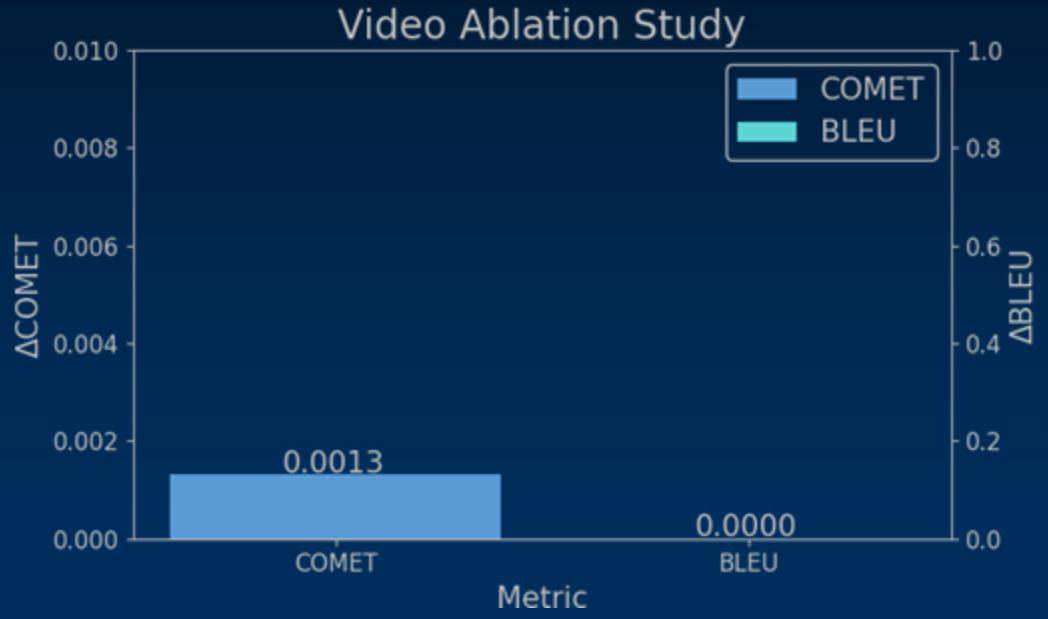
Video Ablation Study

How much do the videos really matter compared to pretraining with more text?

BYU

Video Ablation Study

How much do the videos really matter compared to pretraining with more text?

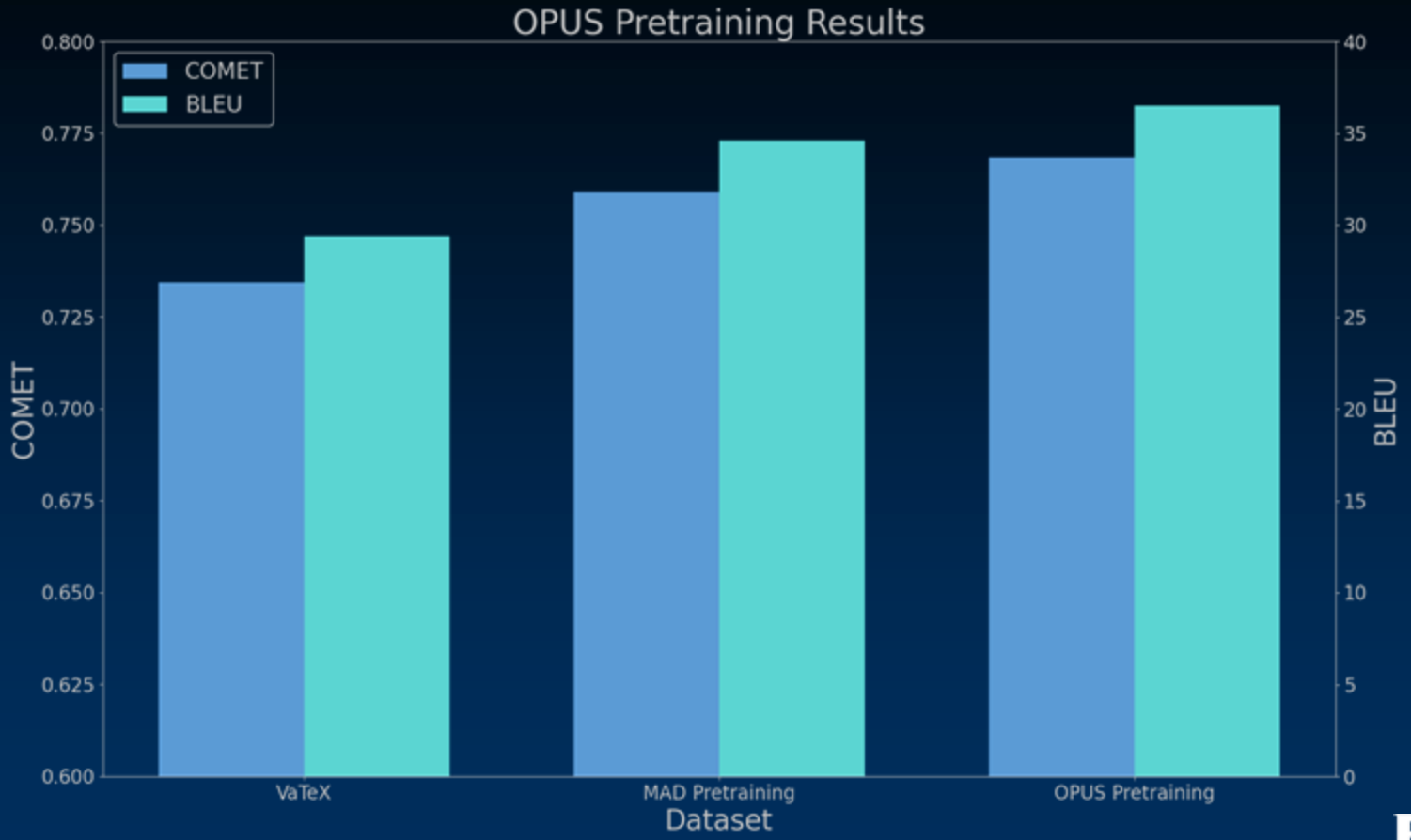


BYU

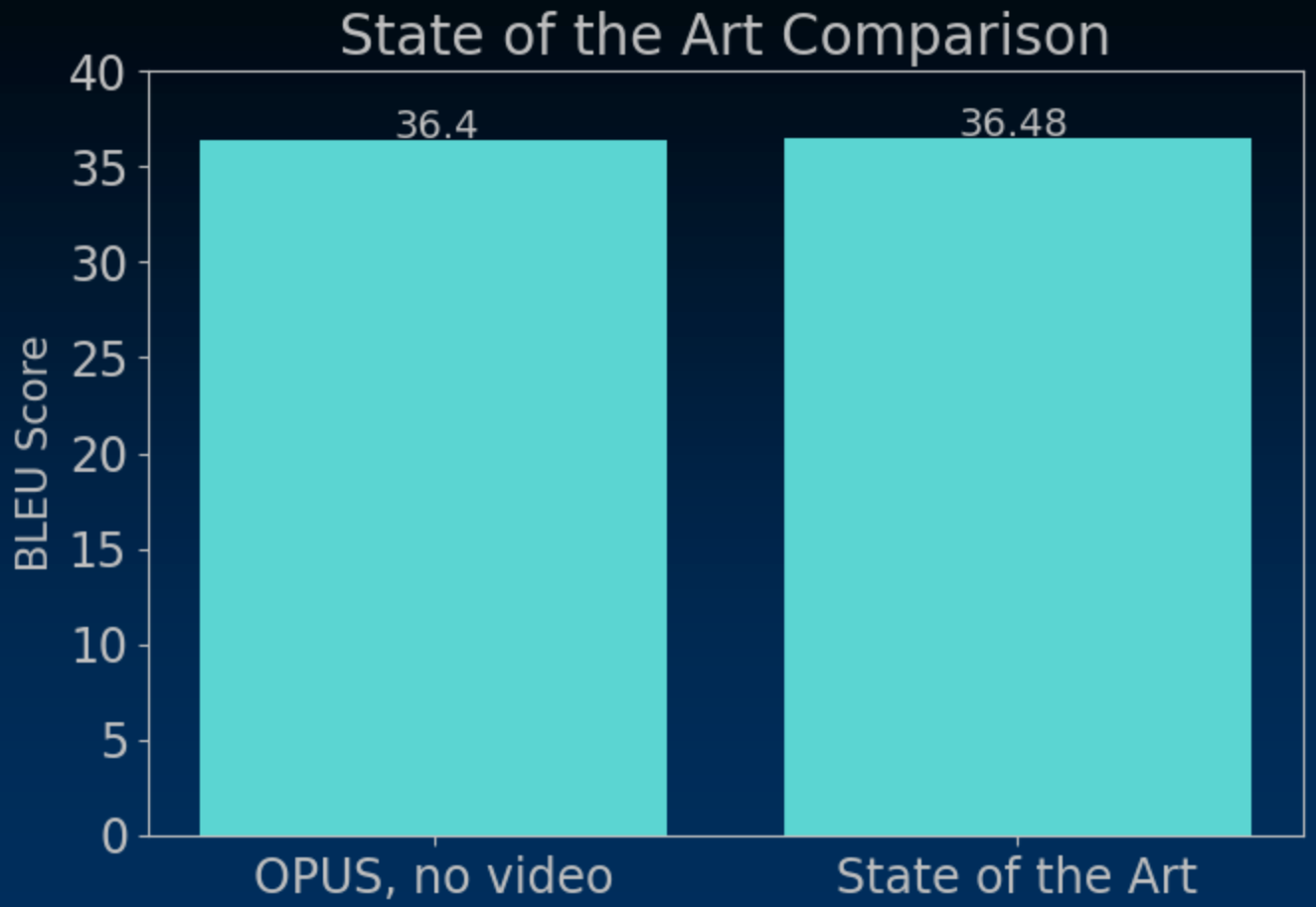
Added Dataset: OPUS Subtitles

- Collection of 10M parallel subtitles in English and Chinese
- Similar domain to VaTeX and MAD, but only dialogue

BYU


OPUS Pretraining Results
COMET
BLEU
0.800
0.775
0.750
0.725
0.700
0.675
0.650
0.625
0.600
40
35
30
25
20
15
10
5
0
VaTeX
MAD Pretraining
OPUS Pretraining
Dataset

BYU


State of the Art Comparison
BLEU Score
40
35
30
25
20
15
10
5
0
36.4
36.48
OPUS, no video
State of the Art

BYU

Conclusions

- Pretraining on diverse text significantly boosts performance.
- The VaTeX dataset sentences do not need video to be translated.

BYU

Future Work

- BigVideo dataset
- MAD 2.0 dataset
- Translate from morphologically simple to more complex languages
- Unsupervised alignment between languages

BYU