Vision Guided Generative Pre-trained Language Models for Multimodal Abstractive Summarization

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

CAiRE
Centre for Artificial Intelligence Research

Tiezheng Yu*, Wenliang Dai*, Zihan Liu, Pascale Fung

Hong Kong University of Science and Technology
Centre for Artificial Intelligence Research (CAiRE)

Introduction

What is

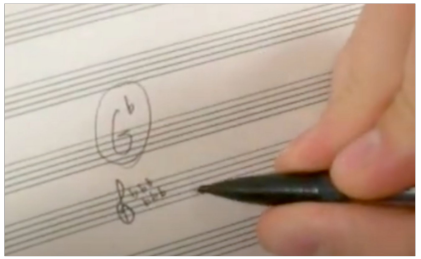
Multimodal Abstractive Summarization (MAS)?

Introduction

What is multimodal abstractive summarization (MAS)?

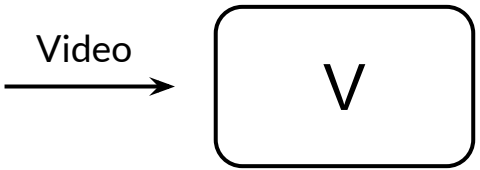
Multimodal abstractive summarization (MAS) is a task that aims to summarize data with multiple modalities and provide a short, concise and readable textual summary to let users quickly acquire the essential information about the video data.

Example

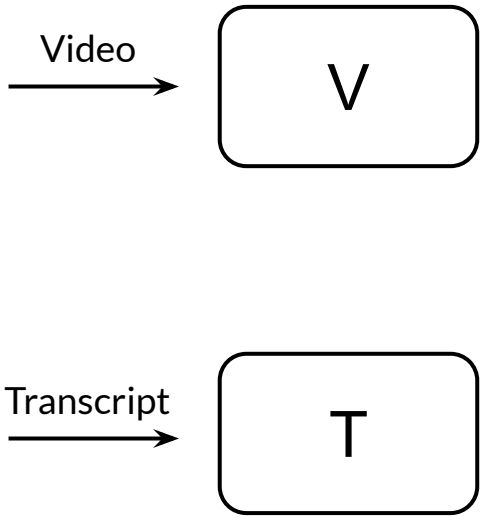
Video Frames	Video Transcript	Ground Truth Summary
 ⋮  ⋮	So now we are going to go over some basics sheet music readings for the key of g flat major. so you noticed the key of g flat, when you are reading real books, there is going to be a treble cleft here. it is going to have 6 flats 1, b flat, e flat, a flat, d flat, g flat and c flat. so 6 flats equals key of g flat. [...] so if you have a flat and there is a natural sign, play the a. so go through the scale and you've got g flat, a flat, d flat, c, flat, d flat, e flat and f, so f is your only 9 flat note in the scale. ***(No mention of the piano)***	Learn how to read and write music intervals for improving your playing and improvisational skills on the ***piano*** in this free video clip series.

Previous Work

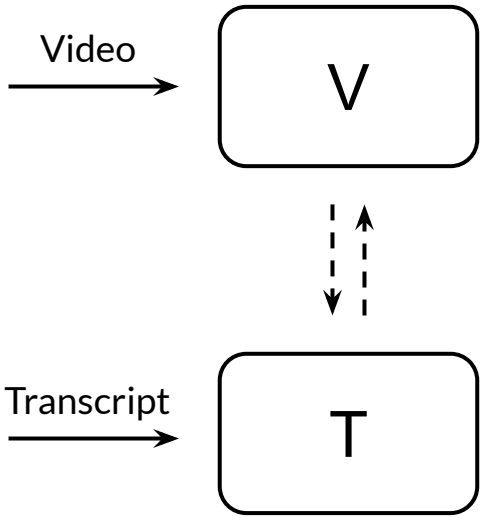
Previous Work



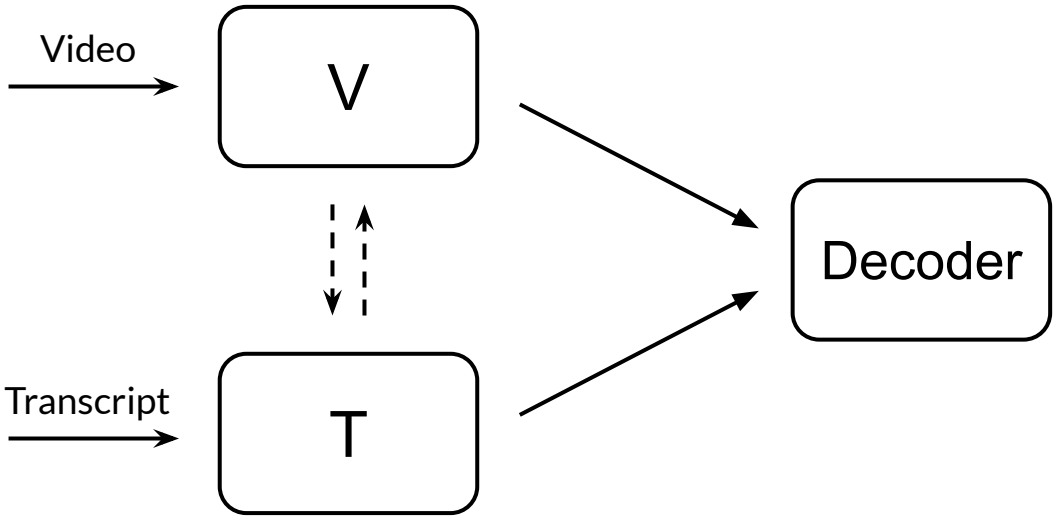
Previous Work



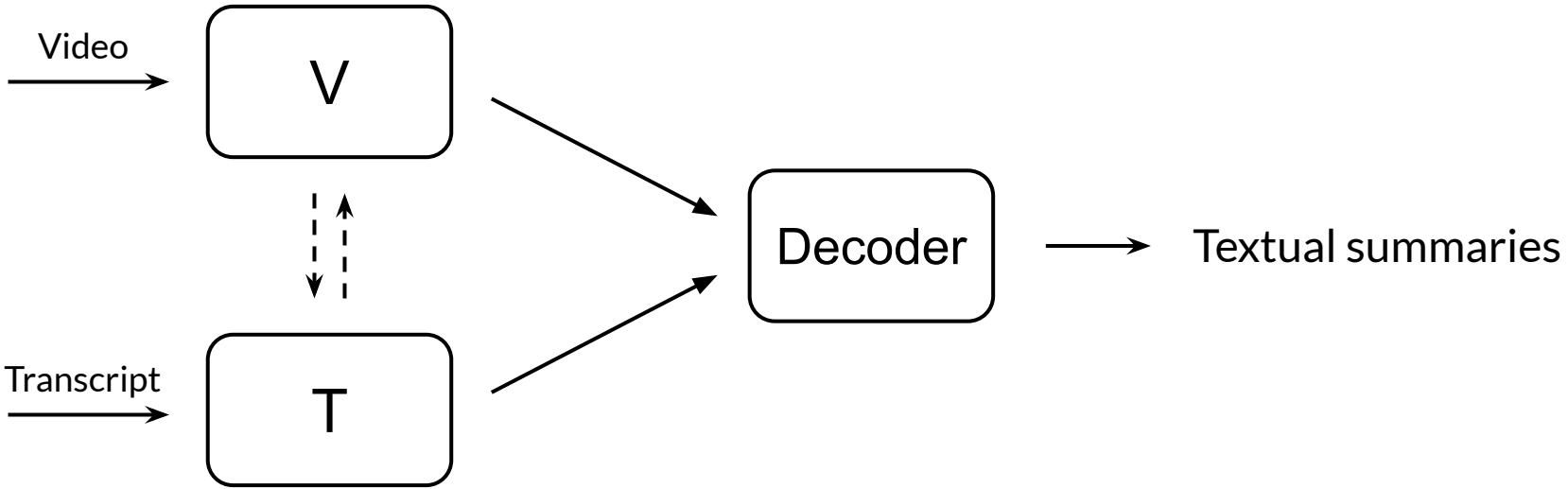
Previous Work



Previous Work



Previous Work



Model Pre-training

Generative Large Pre-trained Language Models (GPLMs)

- GPT/GPT-2/GPT-3
- UniLM
- BART
- T5
- …

Model Pre-training

Vision-Language (VL) Pre-trained Models

- Classification
  - LXMERT, VLBERT, VideoBERT, UNITER, CLIP, ALIGN, etc.
- Text generation with image input
  - VL-BART/T5, E2E-VLP, SIMVLM, etc.

Model Pre-training

Vision-Language (VL) Pre-trained Models

- Classification
  - LXMERT, VLBERT, VideoBERT, UNITER, CLIP, ALIGN, etc.
- Text generation with image input
  - VL-BART/T5, E2E-VLP, SIMVLM, etc.

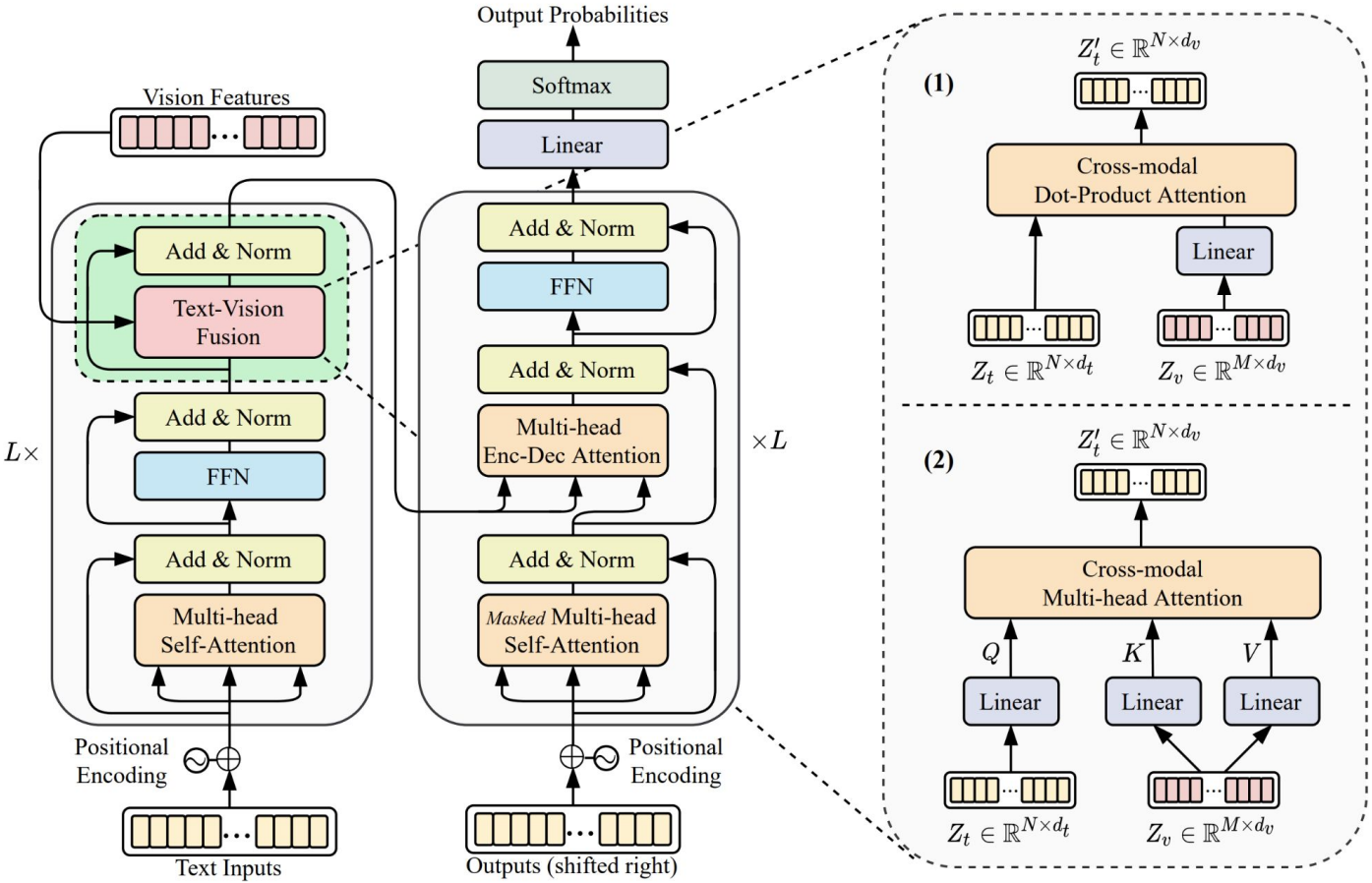
Problem:
No VL pre-trained models yet for text generation with video+text input.

Methodology

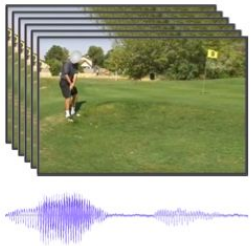
We propose an economical and practical method to leverage and adapt existing GPLMs to the MAS task.

- No need for pre-training
- Minimize the damage to GPLMs' text generation ability while enabling them to handle multimodal data.
- Has the potential to be extended to other multimodal generation tasks

Model Design



Experimental Settings



I'm very close to the green but I didn't get it on the green so now I'm in this grass bunker.

Eu estou muito perto do green, mas eu não pus a bola no green, então agora estou neste bunker de grama.

In golf, get the body low in order to get underneath the golf ball when chipping out of thick grass from a side hill lie.

Figure 1: How2 contains a large variety of instructional videos with utterance-level English subtitles (in bold), aligned Portuguese translations (in italics), and video-level English summaries (in the box). Multimodality helps resolve ambiguities and improves understanding.

Fig. 2 An Example from the how2 dataset[1] We use video, transprictions and summaries in our experiments.Training: 73,993; Validation: 2,965; Testing: 2,156

Video Feature Extraction

A 2048-dimensional feature representation is extracted for every 16 non-overlapping frames using a 3D ResNeXt-101 model [2]

GPLMs

BART-base
T5-base

Software and Hardware

pytorch-lightning
4 RTX 2080Ti

[1] Sanabria, Ramon, et al. "How2: A Large-scale Dataset for Multimodal Language Understanding." NeurIPS. 2018.
[2] Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh. "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018.

Content-F1

1. **Obtain alignment between the summaries and reference by METEOR toolkit**
2. **Remove the function words and task-specific stop words from the summaries and references**
3. **The remaining content words from the summaries and references are treated as two bags of words, and the F1 scores are calculated over the alignment.**

[1] Palaskar, Shruti, et al. "Multimodal Abstractive Summarization for How2 Videos." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.

Main Results

Input	Method	R-1	R-2	R-L	B-1	B-2	B-3	B-4	M	C	CF
Transcript	S2S*	58.6	40.6	53.8	55.2	45.6	39.9	35.8	27.6	2.35	-
	PG*	57.2	39.5	52.8	55.3	45.6	39.8	35.7	26.8	2.13	-
	TF*	59.0	41.0	54.3	56.6	46.7	40.8	36.6	27.7	2.30	-
	T5	62.8	45.0	57.5	60.5	50.4	44.2	39.6	30.6	2.76	61.7
	BART	64.0	46.4	58.9	62.4	52.6	46.4	42.0	31.7	2.97	63.9
Transcript +Video	HA (RNN)*	60.3	42.5	55.7	57.2	47.7	41.8	37.5	28.8	2.48	-
	HA (TF)*	60.2	43.1	55.9	58.6	48.3	43.3	38.1	28.9	2.51	-
	MFFG (RNN)†*	62.3	46.1	58.2	59.1	50.4	45.1	41.1	30.1	2.69	-
	MFFG (TF)*	61.6	45.1	57.4	60.0	50.9	45.3	41.3	29.9	2.67	-
	VG-T5 (Dot-product)	63.0	44.9	57.6	60.1	49.8	43.4	38.8	30.3	2.74	61.4
	VG-T5 (Multi-head)	63.3	45.3	58.0	60.7	50.8	44.7	40.2	31.0	2.86	62.8
	VG-BART (Dot-product)	66.1	49.3	61.2	**64.5**	**55.1**	**49.2**	**44.8**	**33.2**	**3.18**	66.9
	VG-BART (Multi-head)	**66.3**	**49.4**	**61.4**	64.1	54.8	48.9	44.6	33.1	**3.18**	**67.3**

Table 1: Evaluation results of baselines and our proposed models on the How2 dataset. We compare the performance of using transcript only and transcript+video. The † indicates the previous state-of-the-art model. Results with * mark are taken from the previous work (Liu et al., 2020). We denote ROUGE, BLEU, METEOR, CIDEr and Content F1 by R, B, M, C and CF respectively.

[1] Liu, Nayu, et al. "Multistage Fusion with Forget Gate for Multimodal Summarization in Open-Domain Videos." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

How to Inject Visual Information

Input	Method	R-1	R-2	R-L
Transcript	T5	62.8	45.0	57.5
	BART	64.0	46.4	58.9
Transcript +Noise	VG-T5 (Dot-product)	62.5	43.9	57.0
	VG-T5 (Multi-head)	62.8	44.6	57.4
	VG-BART (Dot-product)	63.9	45.6	58.6
	VG-BART (Multi-head)	63.9	46.5	58.7
Transcript +Video	VG-T5 (Dot-product)	63.0	44.9	57.6
	VG-T5 (Multi-head)	63.3	45.3	58.0
	VG-BART (Dot-product)	66.1	49.3	61.2
	VG-BART (Multi-head)	**66.3**	**49.4**	**61.4**

Table 3: Results of using uniform noise to replace the visual features.

1. **Cross modal Dot-product attention and Multi-head attention are two effective ways to inject visual information.**
2. **Multi-head attention is a better approach to inject visual information.**

How to Inject Visual Information

Input	Method	R-1	R-2	R-L	B-1	B-2	B-3	B-4	M	C	CF
Transcript +Video	VG-BART (Multi-head)	66.3	49.4	61.4	64.1	54.8	48.9	44.6	33.1	3.18	67.3
	w/ FG	67.3	50.7	62.4	65.0	55.9	50.1	45.7	33.8	3.25	**72.5**
	w/ VTF	67.3	50.9	62.6	64.9	56.0	50.1	45.7	33.7	3.20	72.1
	w/ FG+VTF	**68.0**	**51.4**	**63.3**	**65.2**	**56.3**	**50.4**	**46.0**	**34.0**	**3.28**	69.7

Table 2: Further Evaluation of adding forget gate (FG) and visual transformer encoder (VTF) to our best model setting in Table 1 on the How2 dataset. VG-BART+FG+VTF largely surpasses the previous state-of-the-art model.

Model	R-1	R-2	R-L
MFFG	62.3	46.1	58.2
BART	64	46.4	58.9
Best	68	51.4	63.3

Table. 7 The ROUGE scores improvements

The visual guidance contributes 83.6% of the overall improvement on average of all ROUGE scores.

Where to Inject Visual Information

Encoder Layer (BART-base)						R-1	R-2	R-L
1	2	3	4	5	6			
✗	✗	✗	✗	✗	✗	64.0	46.4	58.9
✓	✗	✗	✗	✗	✗	66.7	49.9	61.8
✗	✓	✗	✗	✗	✗	67.0	50.5	62.2
✗	✗	✓	✗	✗	✗	67.3	50.8	62.4
✗	✗	✗	✓	✗	✗	67.4	50.9	62.6
✗	✗	✗	✗	✓	✗	67.4	50.8	62.5
✗	✗	✗	✗	✗	✓	67.7	51.3	63.0
✓	✓	✓	✓	✓	✓	60.4	43.4	55.8
✗	✓	✓	✓	✓	✓	64.1	47.0	59.3
✗	✗	✓	✓	✓	✓	65.3	49.2	60.0
✗	✗	✗	✓	✓	✓	67.5	50.9	62.7
✗	✗	✗	✗	✓	✓	**68.0**	**51.4**	**63.3**

Table 4: Performance of different text-vision fusion locations in the encoder of our best model (VG-BART+FG+VTF with cross-modal multi-head attention). ✓ indicates the occurrence of fusion at a certain layer and ✗ indicates non-occurrence. The first row is the result of BART using transcript only.

Decoder Layer (BART-base)						R-1	R-2	R-L
1	2	3	4	5	6			
✗	✗	✗	✗	✗	✗	64.0	46.4	58.9
✓	✗	✗	✗	✗	✗	64.6	47.1	59.6
✗	✓	✗	✗	✗	✗	**65.2**	**48.0**	**60.3**
✗	✗	✓	✗	✗	✗	64.9	46.9	59.6
✗	✗	✗	✓	✗	✗	64.8	46.9	59.7
✗	✗	✗	✗	✓	✗	64.3	46.6	59.1
✗	✗	✗	✗	✗	✓	64.4	46.7	59.0

Table 5: Performance of different fusion locations in the decoder of our best model (VG-BART+FG+VTF with cross-modal multi-head attention).

Injecting at a higher layer in Encoder (closer to the encoder output) brings more improvement.

Effects of the Forget Gate

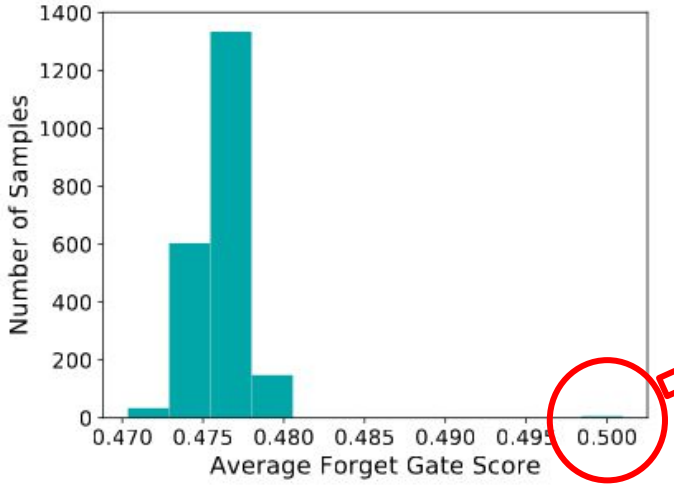


Figure 3: The distribution of average forget gate score on the How2 test set. The model is the VG-BART with dot-product attention.

Transcript: transcript not available **Summary from *Transcript* + *Video***: learn tips on how to write "cane" in chinese radicals with mandarin characters in the free video clip. get free foreign language lessons from an expert. **Reference Summary**: learn what ticks are in chinese calligraphy in this free video clip on languages and writing.

Table 6: An example from How2 testing dataset that has high forget gate score.

The model can still generate reasonable summary for it by paying more attention to the visual information.

Conclusion and Future Work

1. **Propose two types of attention mechanisms for the text-vision fusion and interaction by by inserting attention-based add-on layers to GPLMs: 1) Cross-modal Dot-product Attention; and 2) Cross-modal Multi-head Attention.**
2. **Experimental results show multi-head attention is more robust than the dot-product attention and higher layers of the encoder is the optimal place.**
3. **For future work, we believe that our analyses on the how and where to inject visual information into GPLMs can be applied to other multimodal tasks.**

Thank you for Listening

Check our code

https://github.com/HLTCHKUST/VG-GPLMs