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

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

Hong Kong University of Science and Technology

Contact Information:
Department of Electronic & Computer Science
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong, China

Email: tyuah@connect.ust.hk
wdaiai@connect.ust.hk

Introduction

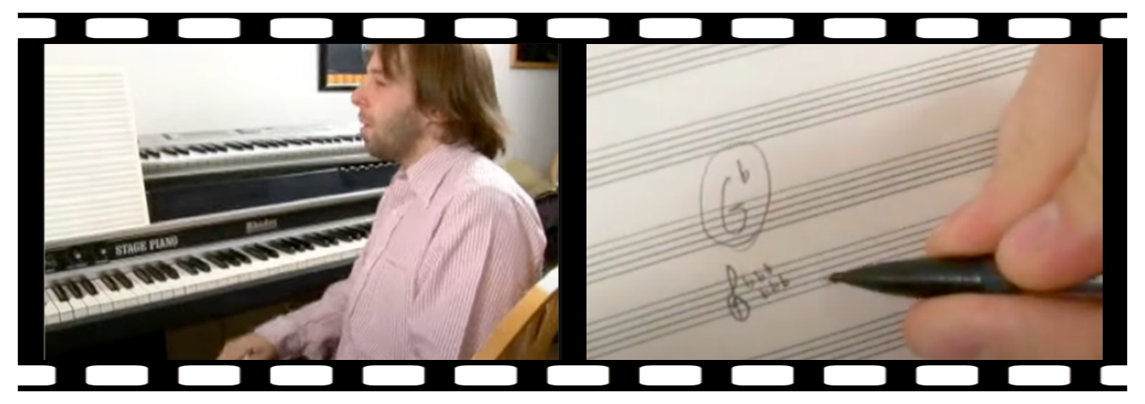
Multimodal abstractive summarization (MAS) aims to take advantage of data from multiple modalities and provides a short, concise and readable textual summary to let users quickly acquire their essential information.

Recently, large-scale generative pre-trained language models (GPLMs) have shown remarkable performance on abstractive text summarization. However, leveraging and adapting GPLMs to MAS is still an unexplored research direction.

Our contributions in this work are threefold:

- To the best of our knowledge, we are the first to inject visual information into text-only GPLMs, and to use it for the MAS task.
- We systematically study two research questions: 1) how to inject visual information into GPLMs without hurting their generation ability; 2) where is the optimal place in GPLMs to inject the visual information?
- Our model significantly outperforms the state-of-the-art model on the How2 dataset, and the injected visual guidance contributes 83.6% of the overall improvement.

Video Frames



Transcript: 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)***

Reference Summary: learn how to read and write music intervals for improving your playing and improvisational skills on the **piano** in this free video clip series.

Summary from *Transcript* (BART): learn tips on how to read and write intervals on sheet music in this free video clip on music theory and music lessons.

Summary from *Transcript+Video* (VG-BART): learn how to sight read in **the key of g flat** for improving your playing and improvisational skills on the **piano** in this free video clip series.

Methodology

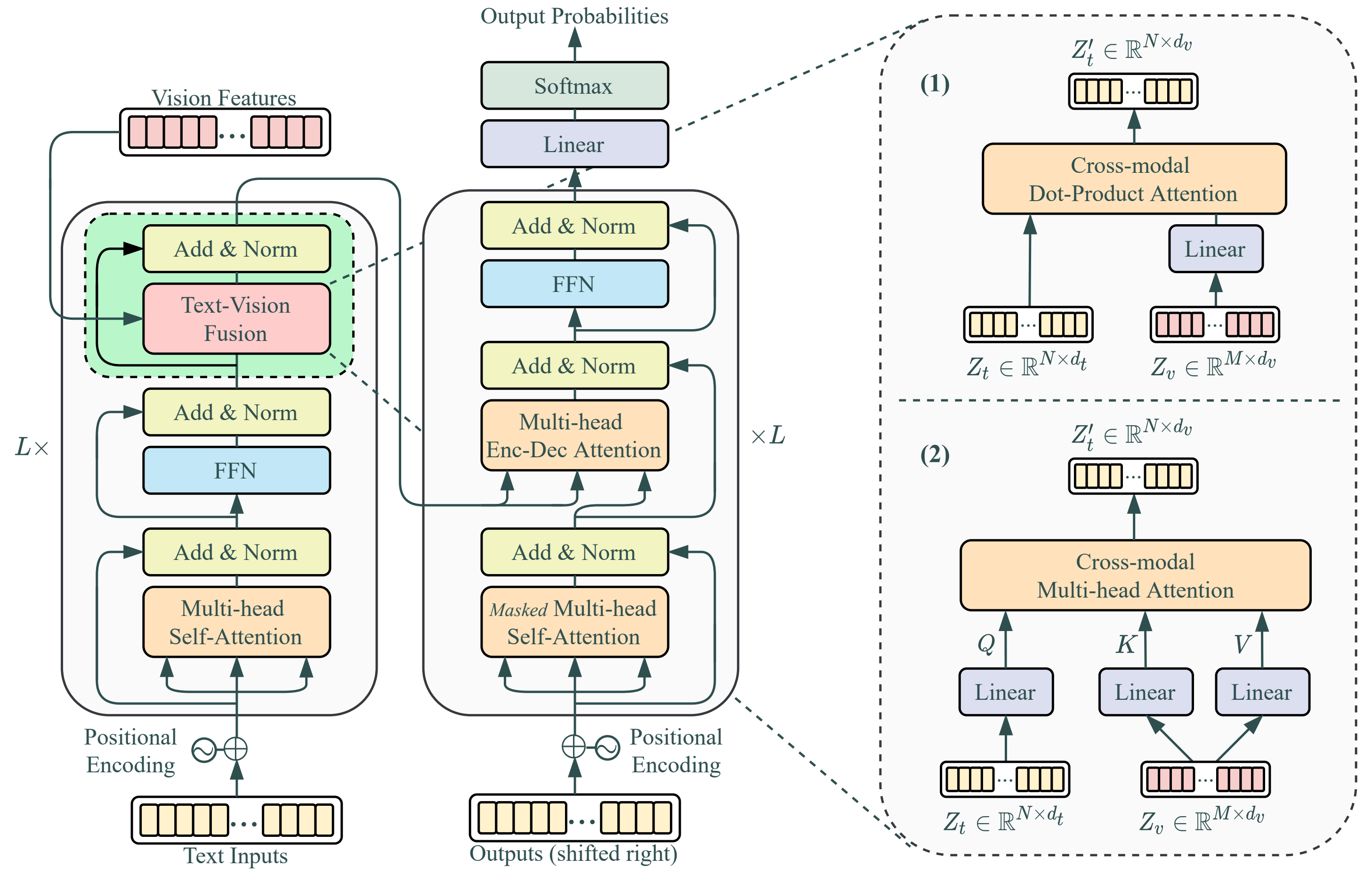

Output Probabilities
Softmax
Linear
Vision Features
Add & Norm
Text-Vision Fusion
Add & Norm
FFN
Add & Norm
Multi-head Self-Attention
$L\times$
Positional Encoding
Text Inputs
Add & Norm
FFN
Add & Norm
Multi-head Enc-Dec Attention
Add & Norm
Masked Multi-head Self-Attention
$\times L$
Positional Encoding
Outputs (shifted right)
(1)
$Z_t' \in \mathbb{R}^{N\times d_v}$
Cross-modal Dot-Product Attention
Linear
$Z_t \in \mathbb{R}^{N\times d_t}$
$Z_v \in \mathbb{R}^{M\times d_v}$
(2)
$Z_t' \in \mathbb{R}^{N\times d_v}$
Cross-modal Multi-head Attention
Q
K
V
Linear
Linear
Linear
$Z_t \in \mathbb{R}^{N\times d_t}$
$Z_v \in \mathbb{R}^{M\times d_v}$

Figure 1: An overview of our proposed VG GPLMs. It is built based on the Transformer-based Seq2Seq GPLMs (*left*). To inject visual information, we insert add-on sub-layers (*the green dashed block*) by mainly leveraging two kinds of attention-based text-vision fusion mechanism (*right*): 1) Cross-modal Dot-Product Attention; and 2) Cross-modal Multi-head Attention.

As exhibited in Figure 1, we insert a third sub-layer (the green dashed block) into each encoder layer, which contains the text-vision fusion mechanism and also a residual connection followed by a layer normalization. We propose two types of text-vision fusion mechanism: **1) Cross-modal Dot-product Attention; and 2) Cross-modal Multi-head Attention**, as shown on the right-hand side of the figure. Given the textual input $Z_t \in \mathbb{R}^{N\times d_t}$ and visual input $Z_v \in \mathbb{R}^{M\times d_t}$, the fusion mechanism produces vision guided output $Z_t' \in \mathbb{R}^{N\times d_t}$ that has a same dimension as the textual input, which allows the continual stacking of layers.

In addition, we also explore the effects of using a forget gate in the text-vision fusion. The forget gate can potentially remove redundant and noisy information from the video features, which also helps the model to learn to discard needless visual information to retain its pre-trained text generation ability.

Results

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 denote ROUGE, BLEU, METEOR, CIDEr and Content F1 by R, B, M, C and CF respectively.

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.

From Table 1, we observe that both text-only T5 and BART outperform all the baseline models by a large gap owe to their pre-trained text generation ability. Moreover, BART is even better than all previous multimodal models trained on transcript and video.

The visual guidance consistently boosts the performance of T5 and BART by a large step. As shown in Table 2, our best model `VG-BART+FG+VTF` with the cross-modal multi-head attention surpasses the previous state-of-the-art model (MFFG) by 5.7 ROUGE-1, 5.3 ROUGE-2, and 5.1 ROUGE-L scores. The visual guidance contributes 83.6% of the overall improvement on average of all ROUGE scores.

How to Inject Visual Information

We mainly adopt two text-vision fusion mechanisms to inject visual information, the cross-modal dot-product attention and multi-head attention. As shown in Table 1, for the VG-BART model, these two fusion mechanisms consistently improve its performance on all metrics by a comparable margin. To ensure the visual features really help in the learning and our add-on layers aid the understanding of them, we conduct further experiments by replacing the visual features in the input data with random noise of the same dimension and sequence length.

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

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

As depicted in Table 3, VG GPLMs with random noise as visual features achieve similar or slightly worse performance compared to the text-only GPLMs. This shows the effectiveness of our method to keep GPLMs' text generation ability.

Furthermore, compared to the dot-product attention based fusion, the multi-head fusion is better at retaining GPLMs' performance, which again demonstrates its superiority.

Where to Inject Visual Information

Encoder Layer						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 and decoder of our best model ✓ 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						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).

As discussed in Section , one of the main challenges of building VG GPLMs is to find the optimal location to inject the visual information (i.e., the text-vision fusion). A sub-optimal location might lead to a less effective modality fusion and even hurt the GPLMs' original text generation ability. As GPLMs have a stack of layers in the encoder and also the decoder, we explore this problem from two aspects: 1) which single layer has the best fusion effect; and 2) does multiple times of fusion help GPLMs to understand the visual information better?

As depicted in Table 4 and 5, firstly, we enumerate each single layer in the encoder and decoder of our best model (`VG-BART+FG+VTF`) to perform the text-vision fusion. In terms of ROUGE scores, we can clearly tell that injecting visual information into the encoder can generally boost the model's performance by a large step, while injecting into the decoder only shows negligible improvement. Furthermore, in the encoder, we observe that injecting at a higher layer (closer to the encoder output) brings more improvement. Instead, in the decoder, there is no clear pattern showing the influence of injecting location.

Secondly, we conduct multiple times of fusion in the encoder's different locations. We observe that when fusing at all encoder layers simultaneously, the model converges to a much worse performance. We conjecture that this causes the catastrophic forgetting of the pre-trained knowledge in GPLMs. We find that fusing at the last several layers (e.g., 5 and 6) in the encoder is able to further improve the summarization performance.

Conclusion and Future Work

In this paper, we propose two types of attention mechanisms for the text-vision fusion and interaction by inserting attention-based add-on layers to GPLMs: 1) Cross-modal Dot-product Attention; and 2) Cross-modal Multi-head Attention. Experimental results show that multi-head attention is more robust than the dot-product attention and higher layers of the encoder is the optimal place. 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.