Supplementary Material for: Transitional Adaptation of Pretrained Models for Visual Storytelling

Youngjae Yu[1]*, Jiwan Chung[2]*, Heeseung Yun[2], Jongseok Kim[3], Gunhee Kim[2]
[1]Allen Institute for AI, [2]Seoul National University, [3]Violet
{yj.yu, jiwanchung, heeseung.yun, js.kim}@vision.snu.ac.kr, gunhee@snu.ac.kr

Abstract

We provide the details of implementation and experiments that are not fully described in the main paper.

The outline of this material is as follows.

- *Implementation Details*
  - *Computing Infrastructure*
  - *Random Seeds*
  - *Computational Efficiency*
- *Additional Experiments*
  - *Fill-in-the-Blank QA*
  - *Randomly Initialized Backbones*
- *AMT user interface*
- *Additional examples*

1. Implementation Details

1.1. Computing Infrastructure

With the GPT-2-small model as the language generator, TAPM includes 751M parameters in total. The model takes approximately 30 minutes per epoch for training using a single NVIDIA TITAN RTX GPU.

We here summarize some information about computing infrastructure for our experiments.

- GPU: NVIDIA TITAN RTX
- CPU: Intel(R) Xeon(R) E5-2650 CPU
- OS : Ubuntu 16.04 LTS OS.
- RAM: SAMSUNG DDR4 8G
- Operating System: Ubuntu 16.04

*Equal Contribution

Table 1. Mean and standard deviations of TAPM using random seed $[0-4]$. Note that we fix the random seed to 0 in all other experiments.

	LSMDC			VIST		
Stats	C	M	R	C	M	R
mean	15.50	8.55	20.23	8.26	34.02	29.70
std	0.33	0.05	0.12	0.17	0.08	0.06

Table 2. The number of parameters and GFLOPs.

Models	GFLOPs (G)	Params (M)
TAPM	5.766	62.3
-A	5.761	60.3

Table 3. Results on Fill-in-the-Blank QA task in LSMDC 2017.

Models	Accuracy
JsFusion [4]	45.52
Cross-Modal BERT -TAPM	50.10
Cross-Modal BERT +TAPM	52.53

- Names and versions of relevant software libraries and frameworks: python $\geq$ 3.6 and PyTorch $\geq$ 1.3

All pretrained transformers are from the huggingface implementations (`https://github.com/huggingface/transformers`). See the source code for more details.

1.2. Random Seeds

Table 1 shows that the performance of TAPM is stable across several random seeds.

1.3. Computational Efficiency

Table 2 shows the number of parameters and GFLOPs (floating point operations) for training. Since the adaptation module (A) requires only 4 FC layers ($f_v^p, f_s^p, f_v^f, f_s^f$), it does not significantly affect computation complexity and training time. The adaptation module is not used for the inference time, so the inference time and complexity of TAPM and TAPM-A are exactly the same. Please note that our adaptation module does not contribute to the complexity of model inference.

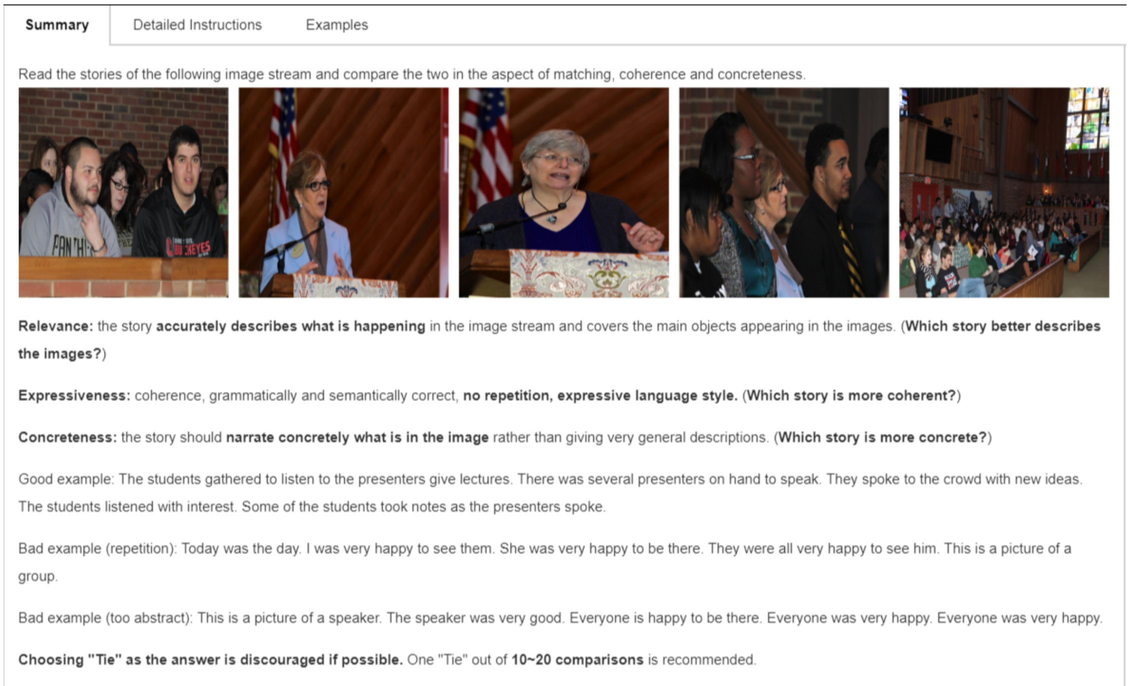


Figure 1. The AMT Instruction for the turkers for the VIST model comparison.

Table 4. Comparison between not pretrained language models on LSMDC 2019 public test set. C, M and R denotes CIDEr, METEOR and ROUGE-L, respectively. All evaluations are on the sentence level.

	No Adaptation			Adaptation (No split-training)			Adaptation (split-training)		
Models	C	M	R	C	M	R	C	M	R
Baseline [3]	11.90	8.25	-	-	-	-	-	-	-
LSTM-Scratch	5.13	6.77	19.34	3.67	5.95	18.51	7.90	7.70	19.45
QRNN-Scratch	1.48	5.65	16.29	3.01	5.73	17.13	7.05	7.25	18.58
GPT2-Scratch	4.17	5.94	16.97	4.01	6.03	17.18	12.68	8.27	20.08
GPT-2	**14.54**	**8.27**	**19.89**	**14.28**	**8.34**	**19.71**	**15.37**	**8.41**	**20.21**

2. Additional Experiments

2.1. Fill-in-the-Blank QA

We explore the generalizability of TAPM on another type of task. In Table 3 we test TAPM with a videoQA task, specifically Fill-in-the-Blank QA task of LSMDC2017, beyond the sequential caption generation tasks in the original paper. The results show that our approach achieves the state-of-the-art performance for another multimodal task.

2.2. Randomly Initialized Backbones

Additionally, we explore how TAPM affects randomly initialized language models. In Table 4, we test three randomly initialized language generators; LSTM-Scratch, QRNN-Scratch [1] and GPT-2-Scratch. As with pretrained language models, adaptation with split-training consistently improves caption quality across all language models. Even when there is no pretrained language information to adapt to, self-supervision may enhance robustness [2] and hence generalization in sparse-signal datasets such as LSMDC.

3. AMT user interface

In our main paper, we conduct our human evaluation to compare different models' outputs on Amazon Mechanical Turk (AMT). Figure 1,2,3 respectively shows the user interfaces for AMT instruction and human evaluation layouts for VIST and LSMDC 2019.

4. Additional examples

We provide additional examples to compare TAPM variants and with selected baselines qualitatively. Figure 4,5 are from LSMDC 2019 experiments, while Figure 6,7 are from VIST tests.

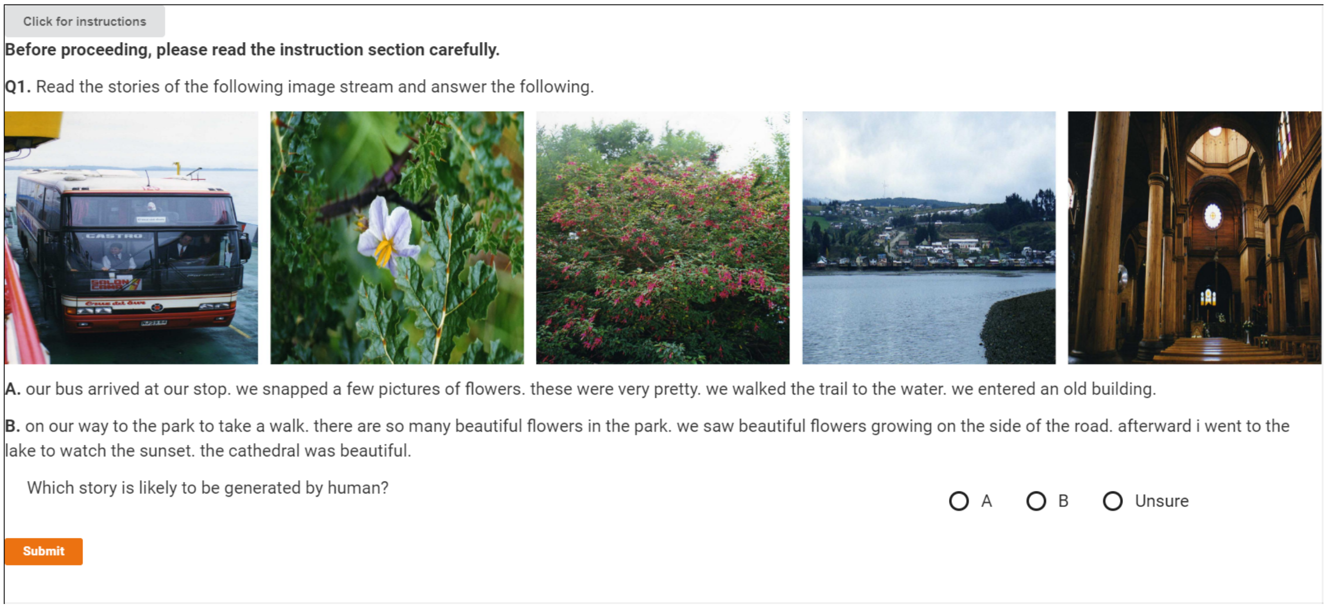


Figure 2. The AMT human evaluation layout for the VIST model comparison.

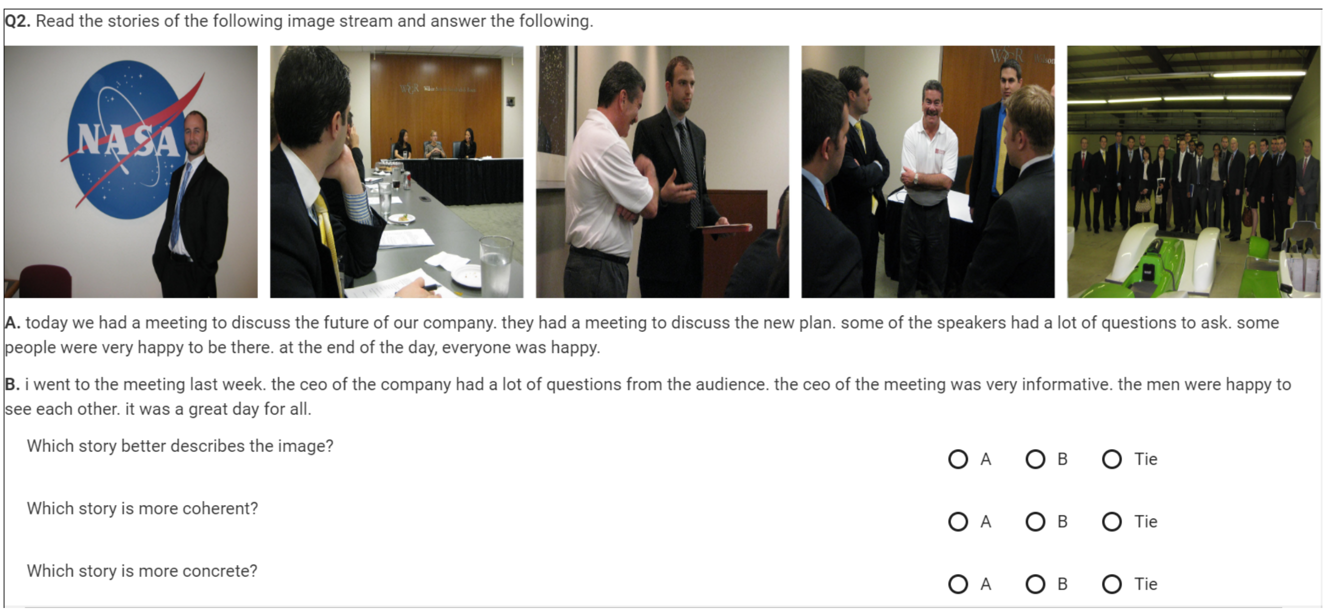


Figure 3. The AMT human evaluation layout for the LSMDC 2019 model comparison.

References

[1] James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. Quasi-Recurrent Neural Networks. In *ICLR*, 2017. 2

[2] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*, 2019. 2

[3] Jae Sung Park, Marcus Rohrbach, Trevor Darrell, and Anna Rohrbach. Adversarial Inference for Multi-Sentence Video Description. In *CVPR*, 2019. 2

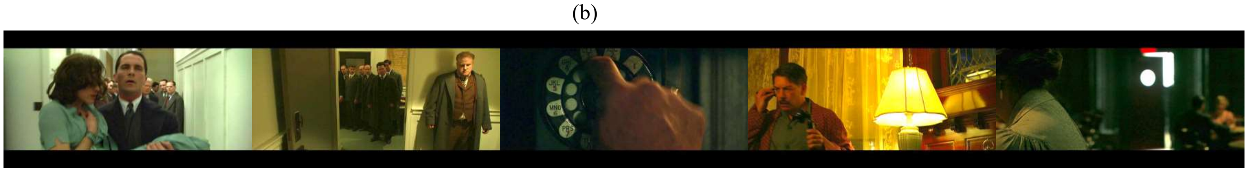
[4] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 2018. 1

GT	someone grooves.	someone touches the screen.	someone clicks the mouse.	someone points to someone.	someone laughs to the kids.
TAPM -A	someone smiles.	***someone clicks on a computer.***	someone looks at the computer.	someone smiles.	someone smiles.
TAPM -Split	someone smiles.	someone looks at the screen.	someone looks at the screen.	someone smiles.	someone smiles.
TAPM (Ours)	someone gives someone a thumbs up.	***someone clicks on a message on the screen: "access denied. ".***	someone sits at a computer.	someone gives a thumbs up.	someone smiles and nods.

(a)

GT	in the control room.	someone sneers at someone.	someone looks sharply at someone.	someone puts his arm around someone.	someone faces someone.
TAPM -A	someone sits on a couch.	someone looks at someone.	someone looks at someone.	someone looks at someone.	someone looks at someone.
TAPM -Split	someone sits on a couch.	someone looks at someone.	someone looks at someone.	someone looks at the camera.	someone looks at someone.
TAPM (Ours)	***someone sits at a desk in the security hub.***	someone and someone exchange a look.	someone and someone watch from their desks.	someone grabs the camera.	someone looks up at someone.

(b)



GT	now, someone nods to someone.	in the interrogation room, the burly agent stays put beside the open door.	someone dials a payphone.	in his home, someone answers.	now, a woman sits at a bar.
TAPM -A	someone looks at someone.	someone walks into the room.	someone is sitting on the couch.	someone sits on the bed, looking at the glass.	someone looks at someone.
TAPM -Split	someone pulls her into the room.	someone looks at someone.	someone takes a drink.	someone sits on the couch.	someone looks at someone.
TAPM (Ours)	someone and someone are in the corridor.	someone looks at someone.	**someone hangs up the phone.**	**someone answers the phone.**	someone looks up at someone.

(c)

Figure 4. The qualitative comparison between TAPM variants in the LSMDC 2019 dataset. Red indicates repetitions, blue/italic indicates interesting samples, and green/bold shows coherent sentences. In (a), TAPM tries to predict the message on the screen but nearly misses.

GT	another closes her eyes.	someone flips open another phone.	and pulls up its speed dial numbers.	someone appears at the top of the list.	someone glances at the executive, then presses a button.
TAPM -A	someone looks at someone.	someone walks to the door.	someone looks at the screen.	*someone reads the number.*	someone looks at the phone.
TAPM -Split	someone looks at someone.	someone looks at the camera.	someone looks at the screen.	*someone reads the number.*	someone looks at the phone.
TAPM (Ours)	someone and someone sit at the end of the long row of seats.	someone sees a man with a gun on a security guard.	someone sees a text message from someone.	*someone's phone shows a text message, "call me. ".*	someone answers his cell.

(d)

GT	someone glances toward the open front door, then takes a bite.	now, a man leads the woman and a plump brunette up the stairs.	the brunette lags behind.	now, a goateed man bicycles with the father.	a car darts out and nearly hits them.
TAPM -A	someone takes a sip of the coffee and then goes to the kitchen.	someone walks up to someone.	someone walks up to someone and starts to walk away.	someone rides a bike.	someone rides a bike.
TAPM -Split	someone pulls out a bottle of milk and starts to eat.	someone walks up to someone.	someone walks up to someone and starts to walk away.	someone rides a bike.	someone rides a bike.
TAPM (Ours)	someone eats a cookie.	someone and someone climb the stairs.	someone and someone run up to the fence.	**now, someone and someone ride on a bike.**	**someone rides up behind them.**

(e)

Figure 5. The qualitative comparison between TAPM variants in the LSMDC 2019 dataset. Red indicates repetitions, blue/italic indicates interesting samples, and green/bold shows coherent sentences. In (d), TAPM takes a wrong guess for the message on the cell phone.

GT	we decided to take a walk around town to see all the local stores.	we started bright and early in the morning, it was such a beautiful day.	we stopped for lunch at the location location and had a wonderful sandwhich.	. we then went to copperfield's books and read and browsed for awhile.	finally, on our way back we stopped in to popeyes for some delicious fried chicken.
TAPM -A	the city was a great place to go.	the view from the window was amazing.	the building was very beautiful.	the restaurant was very nice.	we had a great time.
TAPM -Split	*we went to the location location location location locatio*	the city is beautiful.	the building was very nice.	the sign was very nice.	the place was very nice.
TAPM (Ours)	we went to the city today to see what it was like.	the windows were very old and had a lot of history.	we saw a lot of interesting things.	there was a lot of cool things to see.	afterward i went back to my hotel.

(a)

GT	a little driver wanted to take them to their destination.	the church was small and held a few guests.	members of the party were young and old.	the rings were basic but significant.	photos taken provided memories for a lifetime.
TAPM -A	the car was a little too big for my liking.	the church was beautiful.	the flowers were beautiful.	the bride was very happy to be able to touch the ring.	the bride and groom were happy to be married.
TAPM -Split	the boy was so excited to get his new car.	the church was beautiful.	**the bride was so happy.**	the bride and groom were so happy.	the bride and groom were happy to be married.
TAPM (Ours)	i took my son to the park yesterday.	there were a lot of beautiful sights to see.	**her favorite part of the wedding was the umbrella.**	the bride's ring looked so delicate and beautiful.	they are now married and ready to go home.

(b)

Figure 6. The qualitative comparison between TAPM variants in the VIST dataset. Red indicates uninformative captions, blue/italic indicates language modelling failures, and green/bold shows coherent sentences. In (a), TAPM-Split shows a language modelling failure. Jointly training the adaptation loss with the generation loss could harm the language generation ability of the model. We see that full TAPM does not suffer from such issues. In (b), TAPM-Split and full TAPM try to describe the image within the context of wedding.

GT	the cruise ship look so majestic as it comes in the port.	friends are waiting to greet the passengers.	the port is all decked out for christmas, including this wonderful tree.	this lamp is so cool, i wonder where you find such a thing.	any of the passengers would like to get their picture taken professionally certainly have that option here.
AREL	we went on vacation to location.	we got to see a lot of people there.	*this is a picture of a tree.*	i had a great time there.	*this is a picture of a building.*
XE	we went to the location.	we took a trip to the local museum.	we saw a lot of flowers on the wall.	we saw a lot of interesting things to see.	we had a great time.
TAPM (Ours)	our cruise ship was ready for us.	the view from the deck was amazing.	i bought some flowers while i was there.	there was a lot of decorations there.	this sign was a great addition to the christmas tree.

(a)

GT	it was time for the halloween party and [male] the pirate was ready to go.	but [male] the pirate was very sad as [male] the balloon man showed up in a better costume.	[male] the balloon man had all kinds of fans.	a few that didn't see [male] the pirate were happy with [male] the balloon man.	but when [male] the pirate asked if [male] the balloon man can be his friend, everybody joined in too.
AREL	*this is a picture of a man.*	we had a lot of fun playing games.	there was a lot of fun dancing.	some of the performers were really cool.	*this is a picture of a group of people.*
XE	the halloween party was a lot of fun.	there were a lot of people there.	there were a lot of people there.	we had a great time.	at the end of the night, everyone had a great time.
TAPM (Ours)	i went to my friend's halloween party last night.	we played games and had fun.	some of the costumes were very creative.	after the party we all got together for a group photo.	it was a great halloween party and everyone had a great time.

(b)

Figure 7. The qualitative comparison of TAPM and the selected baselines in the VIST dataset. Red indicates uninformative or misaligned captions, and blue/italic indicates isolated sentences.