End-to-End Transformer Based Model for Image Captioning (Supplementary Materials)

Anonymous submission: 8053

Anonymous submission

Supplementary Materials

More Experimental Details

Source Code The source code for the construction of our PureT is provided as supplementary material, which can be found in folder `PureT_source_code(Excerpt)/models`. We will release trained models and the full source code.

Experimental Results on MSCOCO The detailed captions generated by our models reported in Table 1 and Table 2 of the main paper are provided as supplementary material, organized in the following format:

```
[
{
  "image_id": 391895,
  "caption": "a man walking a dirt bike on a dirt road"
},
...
]
```

which include:

- Offline test results of single model: `MSCOCO_results/Singlemodel_offline_results`
- Offline test results of ensemble model: `MSCOCO_results/Ensemblemodel_offline_results`
- Online test results of ensemble model: `MSCOCO_results/Ensemblemodel_online_results`

Experimental Environments We use Intel Xeon E5-2698 v4 2.2 GHz (20-Core) with Tesla V100 for our experiments. The operating system is Ubuntu Desktop Linux OS 18.04, and the main software environments include Python 3.7.4 and PyTorch 1.5.1.

Visualization of Attention Weights In order to facilitate the intuitive observation of the transformation of attention area during the caption generation process, we visualize the attention weights of Cross MSA Module in decoder. Specifically, we visualize the Cross MSA module in the last block of decoder at each timestep. The size of attention weights is [8, 12, 12], where 8 is the number of heads, we calculate the average values of the 8 heads to obtain the final attention weights α with the size of [12, 12] for visualization. We only retain the top-20 weight values and set the remaining values to 0 for better visualization.

Furthermore, we need to scale α from [12, 12] to the original size [H, W] of the input image to obtain a brightness mask. This can be implemented by applying the `resize`$(\cdot)$ and `pyramid_expand`$(\cdot)$ methods of `skimage.transform` in Python.

Additional Ablation Study

To quantify the influence of different features extracted by different backbone models, we adopt different image captioning models, as baseline models and ablate them with different configurations of backbone models as shown in Table 6. The baseline models include: $\mathcal{M}^2$ Transformer (Cornia et al. 2020), X-Transformer (Pan et al. 2020) and standard Transformer (Vaswani et al. 2017). The backbone models include: Faster R-CNN (Ren et al. 2017) in conjunction with ResNet-101, which is adopted in (Anderson et al. 2018); Faster R-CNN in conjunction with ResNeXt-101, which is adopted in (Jiang et al. 2020); SwinTransformer (Liu et al. 2021).

As we can see, grid features extracted by SwinTransformer can achieve significant performance improvement compared with region features extracted by ResNet-101 and grid features extracted by ResNeXt-101.

In terms of $\mathcal{M}^2$ Transformer and X-Transformer, the backbone models of ResNet-101 and ResNeXt-101 have similar performance. The backbone model of SwinTransformer comprehensively improves scores of all metrics, which boosts the CIDEr score more than 3.7% in $\mathcal{M}^2$ Transformer especially. Note that the backbone model with $N = 3$ has a better performance than $N = 6$ in X-Transformer, which indicates the superiority of SwinTransformer in image captioning and allows us to explore more tiny and efficient models and apply it in more actual scenes. In terms of standard Transformer, the backbone model of SwinTransformer reaches an excellent performance and is even better than $\mathcal{M}^2$ Transformer and X-Transformer in scores of METEOR, CIDEr and SPICE. In terms of our PureT, the backbone of SwinTransformer also achieves a better performance than ResNeXt-101.

In general, in our extensive experiments, we find that the backbone models of CNN (e.g. Faster RCNN in conjunction with ResNet-101 or ResNeXt-101) are more suitable for us-

Baseline Models	Backbone	Feat. Type	Feat. Size	N	B-1	B-2	B-3	B-4	M	R	C	S
$\mathcal{M}^2$ Transformer	ResNet-101	Region	(10−100)	3†	80.8	-	-	39.1	29.2	58.6	131.2	22.6
	ResNeXt-101	Grid	7×7	3‡	80.8	-	-	38.9	29.1	58.5	131.7	22.6
	SwinTransformer	Grid	12×12	3	81.8	66.8	52.6	40.5	29.6	59.9	135.4	23.3
X-Transformer	ResNet-101	Region	(10−100)	6†	80.9	65.8	51.5	39.7	29.5	59.1	132.8	23.4
	ResNeXt-101	Grid	7×7	6‡	81.0	-	-	39.7	29.4	58.9	132.5	23.1
	SwinTransformer	Grid	12×12	6	81.4	66.3	52.0	39.9	29.5	59.5	133.7	23.4
	SwinTransformer	Grid	12×12	3	81.9	66.7	**52.3**	40.1	29.6	59.6	134.8	23.4
standard Transformer	ResNet-101	Region	(10−100)	3	80.0	64.9	50.5	38.7	29.0	58.6	130.1	22.9
	ResNeXt-101	Grid	7×7	3‡	81.2	-	-	39.0	29.2	58.9	131.7	22.6
	ResNeXt-101	Grid	12×12	3	80.8	65.8	51.4	39.4	29.4	59.2	132.8	23.2
	SwinTransformer	Grid	12×12	3	81.6	66.5	52.0	39.8	29.9	59.6	136.4	23.8
PureT	ResNeXt-101	Grid	12×12	3	80.7	65.9	51.7	39.9	29.2	59.1	131.8	23.0
	SwinTransformer	Grid	12×12	3	**82.1**	**67.3**	52.0	**40.9**	**30.2**	**60.1**	**138.2**	**24.2**

Table 6: Performance comparison of different configuration of backbone models. ResNet-101 and ResNeXt-101 indicate Faster R-CNN in conjunction with them respectively. Region features extracted by ResNet-101 have adaptive size of 10 to 100. Grid features extracted by ResNeXt-101 can be extracted in the size of 12×12 or 7×7 by average pooling as need. Grid features (SwinTransformer) are extracted in the size of 12×12. N denotes the number of encoder and decoder blocks, superscript † indicates that the results are from the respectively official paper and ‡ indicates that the results are from (Luo et al. 2021), and other results come from our experiments.

ing LSTM or Transformer with non-standard MSA (e.g. X-Transformer) as decoder, and the backbone of SwinTransformer is more suitable for using Transformer with standard MSA (e.g. $\mathcal{M}^2$ Transformer, standard Transformer and our PureT) as decoder. Therefore, we intend to explore a lighter and simpler Transformer-based model in our future work.

Additional Visualization Examples

Figure 5 gives a comparison of three extra visualization examples of the standard Transformer and our PureT. It can be seen that both standard Transformer and our PureT can pay attention to the correct area when generating words. But our PureT can catch more fine-grained details and generate more accurate and descriptive captions.

Furthermore, we illustrate more examples of captions generated by standard Transformer, $\mathcal{M}^2$ Transformer and our PureT, as shown in Figure 6, Figure 7 and Figure 8. The red marked words indicate wrong information or incomplete sentence, and the green marked words indicate more fine-grained details. For example, $\mathcal{M}^2$ Transformer generates an incomplete sentence "a boat in the water with a mountain in the" for the 4-th image from left of Figure 6, while our PureT generates "a small boat in a large lake with mountains in the background" and contains more detailed adjectives "small" and "large"; $\mathcal{M}^2$ Transformer generates an incorrect sentence "a group of men playing tennis on a field" for the 6-th image from right of Figure 6, in which the splicing of "three pictures" is incorrectly recognized as "a group of people". Our model generates a correct sentence "three pictures of men playing tennis in a field". For another example in the 8-th image from right of Figure 7, $\mathcal{M}^2$ Transformer generates "a young boy standing in front of a refrigerator" and incorrectly identifies the "door" as "refrigerator", standard Transformer generates a simple and correct sentence "a young boy standing in front of a door", and our PureT generates "a young boy wearing a tie standing in front of a door" with more details information "wearing a tie".

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the CVPR*, 6077–6086.

Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2020. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the CVPR*, 10575–10584.

Jiang, H.; Misra, I.; Rohrbach, M.; Learned-Miller, E. G.; and Chen, X. 2020. In Defense of Grid Features for Visual Question Answering. In *Proceedings of the CVPR*, 10264–10273. IEEE.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv preprint arXiv:2103.14030*.

Luo, Y.; Ji, J.; Sun, X.; Cao, L.; Wu, Y.; Huang, F.; Lin, C.; and Ji, R. 2021. Dual-level Collaborative Transformer for Image Captioning. In *Proceedings of the AAAI*, 2286–2293. AAAI Press.

Pan, Y.; Yao, T.; Li, Y.; and Mei, T. 2020. X-Linear Attention Networks for Image Captioning. In *Proceedings of the CVPR*, 10968–10977.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6): 1137–1149.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Proceedings of the NIPS*, 5998–6008.

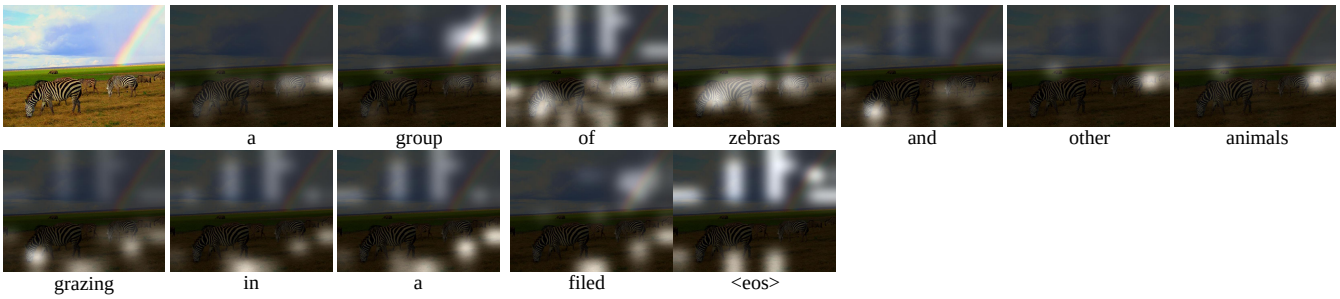
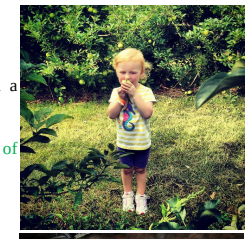


Figure 5: Visualization of attention heatmap during caption generation process of standard Transformer and our PureT.

image_id: 400453
$\mathcal{M}^2$ **Transformer:** a woman sitting at a table with a plate of food
Transformer: a woman sitting at a table eating a salad
PureT: a woman sitting at a table drinking a glass of wine

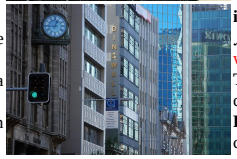


image_id: 526955
$\mathcal{M}^2$ **Transformer:** a young boy standing in the grass holding a tennis racket
Transformer: a young girl standing in a field of apples
PureT: a little girl standing in a field eating a piece of fruit

image_id: 2295
$\mathcal{M}^2$ **Transformer:** a black and white cat sitting on a wooden table
Transformer: a cat sitting on top of a wooden bowl
PureT: a black and white cat sticking its head in a wooden bowl

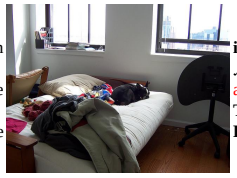
image_id: 268885
$\mathcal{M}^2$ **Transformer:** a cat laying on top of a computer keyboard
Transformer: two cats laying on top of a computer keyboard
PureT: two cats laying on top of a desk next to a keyboard

image_id: 5123
$\mathcal{M}^2$ **Transformer:** a woman standing in front of a house on a street
Transformer: a woman sitting on a bench in front of a house
PureT: a woman standing on the corner of a street in front of a house

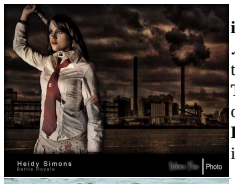


image_id: 92488
$\mathcal{M}^2$ **Transformer:** a clock on the side of a city street with a
Transformer: a traffic light in front of a building with a clock
PureT: a traffic light in front of a tall building with a clock on it

image_id: 156532
$\mathcal{M}^2$ **Transformer:** a boat in the water with a mountain in the
Transformer: a boat in the water with a mountain in the background
PureT: a small boat in a large lake with mountains in the background



image_id: 241503
$\mathcal{M}^2$ **Transformer:** a dog laying on a bed in a room with a
Transformer: a dog laying on a bed in a room
PureT: an unmade bed with a dog sleeping on it



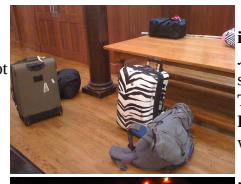
image_id: 213773
$\mathcal{M}^2$ **Transformer:** a woman wearing a tie standing next to a body of water
Transformer: a woman wearing a tie standing in front of a city
PureT: a woman wearing a white shirt and tie standing in front of a city

image_id: 72873
$\mathcal{M}^2$ **Transformer:** a bedroom with a bed and a window in a
Transformer: a bedroom with a bed and a window
PureT: a hospital room with a bed and a window

image_id: 467540
$\mathcal{M}^2$ **Transformer:** a woman in a boat with an umbrella on a
Transformer: a man sitting in a boat with an umbrella
PureT: a man sitting on a small boat with an umbrella in the water

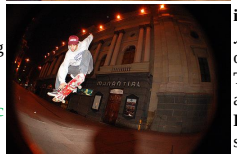
image_id: 248616
$\mathcal{M}^2$ **Transformer:** a group of men playing tennis on a field
Transformer: two pictures of a man hitting a tennis ball with a tennis racket
PureT: three pictures of men playing tennis in a field

image_id: 79860
$\mathcal{M}^2$ **Transformer:** a man sitting at a table eating a hot dog
Transformer: a man sitting at a table holding a hot dog
PureT: a man sitting at a table holding a large hot dog



image_id: 378506
$\mathcal{M}^2$ **Transformer:** a person laying on the floor next to a suitcase
Transformer: a pile of luggage sitting on top of a table
PureT: a group of luggage sitting on the floor next to a wooden table

image_id: 145873
$\mathcal{M}^2$ **Transformer:** a large construction truck driving down a road with a
Transformer: a truck with a crane on the back of it
PureT: a crane on a truck on a highway with a traffic light



image_id: 392640
$\mathcal{M}^2$ **Transformer:** a man doing a trick on a skateboard on a ramp
Transformer: a man doing a trick on a skateboard in the air
PureT: a man flying through the air while riding a skateboard

image_id: 531602
$\mathcal{M}^2$ **Transformer:** a young boy brushing his teeth in a bathroom
Transformer: a young boy brushing his teeth with a toothbrush
PureT: a young boy brushing his teeth with an electric toothbrush

image_id: 300066
$\mathcal{M}^2$ **Transformer:** a plane is sitting on the runway with a man
Transformer: a large airplane sitting on top of an airport runway
PureT: an airplane sitting on the runway with a mountain in the background

Figure 6: More examples of captions generated by standard Transformer, $\mathcal{M}^2$ Transformer and our PureT. The red marked words indicate wrong information or incomplete sentence, and the green marked words indicate more fine-grained details.

image_id: 541478
$\mathcal{M}^2$ **Transformer:** a man wearing glasses and a tie holding a sign
Transformer: a man wearing glasses and a tie holding a cell phone
PureT: a man wearing glasses and a shirt and tie holding a sign

image_id: 520012
$\mathcal{M}^2$ **Transformer:** a tray with a sandwich and a cup of coffee on a
Transformer: a tray with plates of food and a cup of coffee
PureT: a plate of pancakes with vegetables and a cup of coffee

image_id: 534162
$\mathcal{M}^2$ **Transformer:** a woman in a tank top is looking at her cell phone
Transformer: two women sitting on a couch looking at cell phones
PureT: two women sitting on a black couch looking at their cell phones

image_id: 557731
$\mathcal{M}^2$ **Transformer:** a pizza sitting on top of a wooden board with a
Transformer: a pizza sitting on top of a wooden table
PureT: a pizza sitting on top of a wooden table with a bottle of wine

image_id: 408621
$\mathcal{M}^2$ **Transformer:** a green bus is parked in front of a building
Transformer: a bus parked in front of a building
PureT: a green and yellow bus parked in front of a building

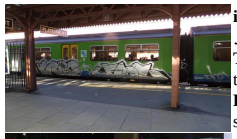
image_id: 335800
$\mathcal{M}^2$ **Transformer:** a traffic light with a street sign on a pole
Transformer: a street sign on the side of a pole with a building
PureT: a street sign and a flag on a pole in front of a tall building

image_id: 27186
$\mathcal{M}^2$ **Transformer:** a little girl holding a nintendo wii game controller
Transformer: a little girl holding a wii remote in her hand
PureT: a little girl in a pink dress holding a nintendo wii game controller

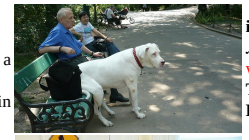
image_id: 579664
$\mathcal{M}^2$ **Transformer:** a bunch of bananas in baskets in a
Transformer: a bunch of bananas sitting on top of a table
PureT: a bunch of bananas sitting on top of a wooden box

image_id: 537991
$\mathcal{M}^2$ **Transformer:** a woman sitting on a couch looking at her cell phone
Transformer: a woman sitting in a chair holding two cell phones
PureT: a woman wearing headphones looking at her cell phone with a laptop

image_id: 69795
$\mathcal{M}^2$ **Transformer:** a blue vase with flowers in front of a window
Transformer: a blue vase with flowers sitting in front of a window
PureT: a blue vase with purple flowers sitting in front of a window

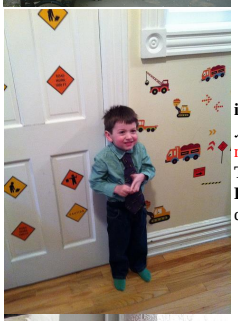


image_id: 337666
$\mathcal{M}^2$ **Transformer:** a train with graffiti on the side of a
Transformer: a train with graffiti on the side of it at a train station
PureT: a green and white train with graffiti at a train station

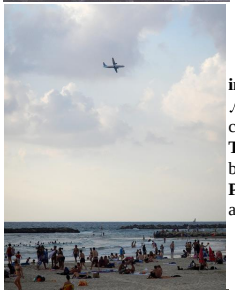


image_id: 334321
$\mathcal{M}^2$ **Transformer:** a white dog and a man on a bench with a
Transformer: a white dog sitting on a bench
PureT: a white dog standing next to a man on a bench

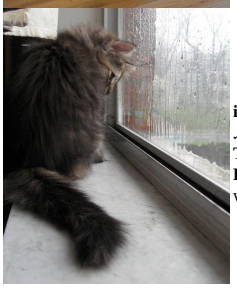
image_id: 71090
$\mathcal{M}^2$ **Transformer:** a white teddy bear laying on a window sill with a
Transformer: a black and white cat laying in front of a white teddy bear
PureT: a black and white cat laying in front of a large stuffed teddy bear



image_id: 482160
$\mathcal{M}^2$ **Transformer:** a young boy standing in front of a refrigerator
Transformer: a young boy standing in front of a door
PureT: a young boy wearing a tie standing in front of a door



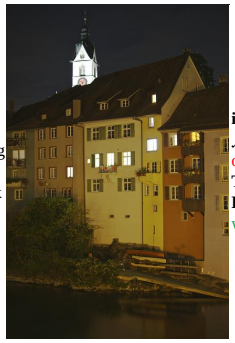
image_id: 28157
$\mathcal{M}^2$ **Transformer:** a large airplane flying over a crowded beach
Transformer: a large airplane flying in the sky over a beach
PureT: a large airplane flying over a crowd of people on a beach



image_id: 377000
$\mathcal{M}^2$ **Transformer:** a cat is looking out of a window
Transformer: a cat sitting on a window sill
PureT: a cat sitting on a window sill looking out a window

Figure 7: More examples of captions generated by standard Transformer, $\mathcal{M}^2$ Transformer and our PureT. The red marked words indicate wrong information or incomplete sentence, and the green marked words indicate more fine-grained details.

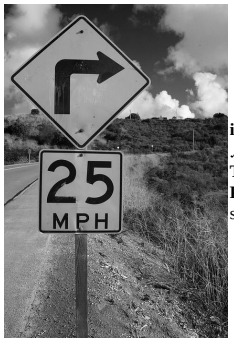
image_id: 360318
$\mathcal{M}^2$ **Transformer :** a brown bear sitting on a rock eating food
Transformer: a monkey eating a piece of food on a rock
PureT: a monkey sitting on a rock eating food



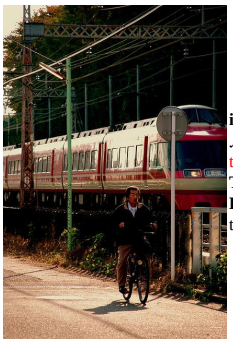
image_id: 442993
$\mathcal{M}^2$ **Transformer :** a building with a clock tower on top of a
Transformer: a building with a clock tower on top of it
PureT: a building with a clock tower in front of the water

image_id: 469618
$\mathcal{M}^2$ **Transformer :** a group of people sitting around a table with plates of food
Transformer: a group of people sitting at a table with plates of food
PureT: a group of people sitting at a picnic table with plates of food

image_id: 236068
$\mathcal{M}^2$ **Transformer :** a man in a top hat talking on a cell phone
Transformer: a man in a suit talking on a cell phone
PureT: a man wearing a top hat talking on a cell phone next to an old car



image_id: 444302
$\mathcal{M}^2$ **Transformer :** a street sign on the side of a road
Transformer: a street sign on the side of a road
PureT: a black and white photo of a street sign on the side of a road



image_id: 483108
$\mathcal{M}^2$ **Transformer :** a man riding a bike with a train in the
Transformer: a man riding a bike next to a train
PureT: a man riding a bike down the street next to a train

Figure 8: More examples of captions generated by standard Transformer, $\mathcal{M}^2$ Transformer and our PureT. The red marked words indicate wrong information or incomplete sentence, and the green marked words indicate more fine-grained details.