With a Little Help from your own Past: Prototypical Memory Networks for Image Captioning

Supplementary Material

Manuele Barraco[1] Sara Sarto[1] Marcella Cornia[1] Lorenzo Baraldi[1] Rita Cucchiara[1,2]
[1]University of Modena and Reggio Emilia, Modena, Italy [2]IIT-CNR, Pisa, Italy
{name.surname}@unimore.it

In the following, we present additional materials about PMA-Net. In particular, we provide additional experimental results using model ensembling, visualizations of memory attention scores on sample images from the COCO dataset, and qualitative results on all considered datasets.

1. Additional Experimental Results

As a complement of the experiments reported in the main paper, in Table 1 we report the performance of PMA-Net using an ensemble of four models on the Karpathy test split, in comparison with ensembles built with a base Transformer using the same visual features and with related methods from the literature. PMA-Net confirms its effectiveness showing an improvement over the other competitors and the baseline version.

	B-1	B-2	B-3	B-4	M	R	C	S
GCN-LSTM [9]	80.9	-	-	38.3	28.6	58.5	128.7	22.1
SGAE [8]	81.0	-	-	39.0	28.4	58.9	129.1	22.2
AoANet [3]	81.6	-	-	40.2	29.3	59.4	132.0	22.8
$\mathcal{M}^2$ Transformer [2]	82.0	-	-	40.5	29.7	59.5	134.5	23.5
X-Transformer [7]	81.7	66.8	52.6	40.7	29.9	59.7	135.3	23.8
DLCT [6]	82.2	-	-	40.8	29.9	59.8	137.5	23.3
COS-Net [4]	83.5	69.1	54.9	42.9	**30.8**	61.0	143.0	**24.7**
Transformer	83.9	69.1	54.6	42.3	30.2	60.7	141.3	23.6
PMA-Net	**84.9**	**70.4**	**56.1**	**43.8**	30.7	**61.6**	**145.9**	24.1

Table 1. Comparison with the state of the art on the COCO Karpathy test split using an ensemble of models.

2. Visualizations

In addition to the aggregate visualization of the attentive distributions shown in the main paper on the entire COCO test set, in Figure 1 we show some examples of the usage of the prototypical memories on sample captions in a non-aggregated way. The attention score for each word is computed in the same manner: we calculate the relative average of the attention scores related to the memory slots over the totality of the attention scores for each layer, then we average over all the layers. These visualizations show how memories are employed during caption generation in different scenarios, *i.e.* for retrieving further details on specific concepts, describing actions, and using more appropriate nouns and verbs with respect to other approaches. As also shown in the main paper, the prototype memories are in general employed during the generation of the entire sentence.

3. Qualitative Results

Finally, we report qualitative results on different datasets. In particular, Figure 2 shows sample results from the COCO Karpathy test split, obtained from PMA-Net and the baseline version, while additional samples are reported on the robust test split of COCO [5] in Figure 3 and on the validation split of the nocaps [1] dataset in Figure 4.

We observe that, on average, PMA-Net generates more detailed and correct descriptions with respect to a vanilla Transformer architecture. Further, the proposed approach showcases better performances when describing pairs of objects which do not appear in the training set and when describing out-of-domain objects.

References

[1] Harsh Agrawal, Karan Desai, Xinlei Chen, Rishabh Jain, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019. 1

[2] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-Memory Transformer for Image Captioning. In *CVPR*, 2020. 1

[3] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on Attention for Image Captioning. In *ICCV*, 2019. 1

[4] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *CVPR*, 2022. 1

Figure 1. Sample captions generated on the COCO dataset, together with the magnitude of attention on prototype memories over time.

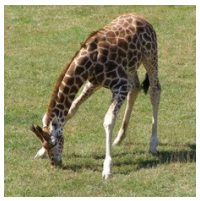
[5] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural Baby Talk. In *CVPR*, 2018. 1

[6] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-Level Collaborative Transformer for Image Captioning. In *AAAI*, 2021. 1

[7] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-Linear Attention Networks for Image Captioning. In *CVPR*, 2020. 1

[8] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-Encoding Scene Graphs for Image Captioning. In *CVPR*, 2019. 1

[9] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring Visual Relationship for Image Captioning. In *ECCV*, 2018. 1



Ground-truth: A giraffe bending over while standing on green grass.
Transformer: A giraffe bending down to eat the grass.
PMA-Net: A giraffe bending down to eat grass in a field.

Ground-truth: A ship in the water sailing past the city in the background.
Transformer: A boat in a large body of water.
PMA-Net: A boat in the water near a city.

Ground-truth: A small black dog playing with a frisbee.
Transformer: A black dog running with a frisbee in its mouth.
PMA-Net: A dog playing with a frisbee in the grass.

Ground-truth: A man bending down to fix his motorcycle in a parking lot.
Transformer: A man on a dirt motorcycle on a road.
PMA-Net: A man leaning on a motorcycle in a parking lot.

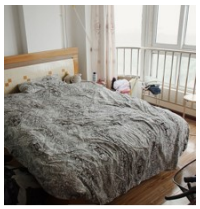
Ground-truth: A woman sitting in a chair and drinking from a wine glass.
Transformer: A woman sitting on a couch drinking a glass.
PMA-Net: A woman sitting on a couch holding a glass of wine.

Ground-truth: An elephant can be seen through a barbed wire fence.
Transformer: An elephant standing next to a wire fence.
PMA-Net: An elephant standing behind a barbed wire fence.

Ground-truth: A girl playing with a red frisbee outside at the park.
Transformer: A woman throwing a frisbee in a park.
PMA-Net: A woman throwing a red frisbee in a park.

Ground-truth: Well cooked rice and vegetables on a white plate.
Transformer: A plate of food with rice and rice on a.
PMA-Net: A plate of food with rice and broccoli on a table.

Ground-truth: A black and white photo of a castle at night.
Transformer: A large building with a clock tower on top of.
PMA-Net: A black and white photo of a building with a clock.



Ground-truth: A bed sitting on a hard wood floor.
Transformer: A bedroom with a bed and a desk.
PMA-Net: A bedroom with a large bed and a window.

Ground-truth: Four people compete on horseback playing playing polo.
Transformer: A group of men on horses playing soccer on a beach.
PMA-Net: A group of men playing polo on the beach.

Ground-truth: Many oranges have been placed inside a bowl.
Transformer: A white plate of four oranges on a table.
PMA-Net: A white bowl of oranges on a table.

Ground-truth: A man with a skateboard that is jumping in the air.
Transformer: A man is doing a trick on a skateboard.
PMA-Net: A man flying through the air while riding a skateboard.

Ground-truth: A vase filled with yellow and red flowers.
Transformer: A vase of flowers sitting on a table.
PMA-Net: A vase filled with red and yellow flowers on a table.

Ground-truth: A bus is parked near a bus stop on a street.
Transformer: Two buses parked on the side of a street.
PMA-Net: A blue and white bus parked at a bus stop.

Ground-truth: A man in a suit carefully adjusts his tie.
Transformer: A man in a suit tying his tie.
PMA-Net: A man adjusting his tie in a suit.

Figure 2. Qualitative results on sample images from the COCO Karpathy test split.

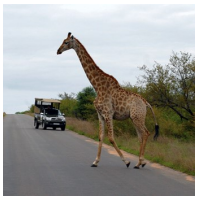
Ground-truth: A man riding a red motorcycle down a street with a dog on back of it.
Transformer: A man riding on the back of a red motorcycle.
PMA-Net: A man riding a motorcycle with a dog on the back.

Ground-truth: The group of people lay on the beach near a parked bicycle.
Transformer: A group of people sitting on top of a sandy beach.
PMA-Net: A group of people sitting on a beach next to a bike.

Ground-truth: A piece of dessert and fork are on the plate.
Transformer: A slice of pie sitting on top of a white plate.
PMA-Net: A slice of cheesecake on a plate with a fork.

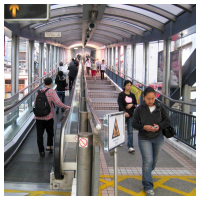
Ground-truth: A dog is lying on the carpet of the living room.
Transformer: A living room with couches and a coffee table.
PMA-Net: A living room filled with furniture and a dog laying on a rug.

Ground-truth: A man on a cell phone waits with luggage by a train track.
Transformer: A man standing on a train platform next to luggage.
PMA-Net: A man talking on a cell phone while standing next to a train track.



Ground-truth: A giraffe is crossing the street in front of a car.
Transformer: A giraffe walking down a road next to a truck.
PMA-Net: A giraffe crossing a road next to a vehicle.

Figure 3. Qualitative results on sample images from the robust COCO test set.



Transformer: A group of people walking down a subway train.
PMA-Net: A group of people walking up an escalator.

Transformer: A man is standing in front of a bike.
PMA-Net: A man taking a picture of himself in front of a mirror.

Transformer: A woman is washing a baby on a machine.
PMA-Net: A woman holding a baby in a hospital room.

Transformer: A pirate ship on the water near a.
PMA-Net: A metal barrel sitting next to a body of water.

Transformer: A man in a suit and tie holding a microphone.
PMA-Net: A man in a suit and tie holding a remote.

Transformer: A man is on a boat in the water.
PMA-Net: A man standing on a boat holding a fish.

Figure 4. Qualitative results on sample images from the nocaps validation set.