Meshed-Memory Transformer for Image Captioning
Supplementary Material

Marcella Cornia Matteo Stefanini Lorenzo Baraldi Rita Cucchiara
University of Modena and Reggio Emilia
{name.surname}@unimore.it

In the following, we present additional material about our $\mathcal{M}^2$ Transformer model. In particular, we provide additional training and implementation details, further experimental results, and visualizations.

1. Additional implementation details

Decoding optimization. As mentioned in Sec. 3.3, during the decoding stage computation cannot be parallelized over time as the input sequence is iteratively built. A naive approach would be to feed the model at each iteration with the previous $t-1$ generated words, $\{w_0, w_1, ..., w_{t-1}\}$ and sample the next predicted word w_t after computing the results of each attention and feed-forward layer over all timesteps. This in practice requires to re-compute the same queries, keys, values and attentive states multiple times, with intermediate results depending on w_t being recomputed $T-t$ times, where T is the length of the sampled sequence (in our experiments T is equal to 20).

In our implementation, we revert to a more computationally friendly approach in which we re-use intermediate results computed at previous timesteps. Each attentive layer of the decoder internally stores previously computed keys and values. At each timestep of the decoding, the model is fed only with w_{t-1}, and we only compute queries, keys and values depending on w_{t-1}.

In PyTorch, this can be implemented by exploiting the `register_buffer` method of `nn.Module`, and creating buffers to hold previously computed results. When running on a NVIDIA 2080Ti GPU, we found this to reduce training and inference times by approximately a factor of 3.

Vocabulary and tokenization. We convert all captions to lowercase, remove punctuation characters and tokenize using the spaCy NLP toolkit[1]. To build vocabularies, we remove all words which appear less than 5 times in training and validation splits. For each image, we use a maximum number of region feature vectors equal to 50.

Model dimensionality and weight initialization. Using 8 attentive heads, the size of queries, keys and values in each head is set to $d/8 = 64$. Weights of attentive layers are initialized from the uniform distribution proposed by Glorot *et al*. [3], while weights of feed-forward layers are initialized using [4]. All biases are initialized to 0. Memory vectors for keys and values are initialized from a normal distribution with zero mean and, respectively, $1/d_k$ and $1/m$ variance, where d_k is the dimensionality of keys and m is the number of memory vectors.

[1] `https://spacy.io/`

Memories	B-1	B-4	M	R	C	S
No memory	80.4	38.3	29.0	58.2	129.4	22.6
20	80.7	38.9	29.0	58.4	129.9	22.7
40	**80.8**	**39.1**	**29.2**	**58.6**	**131.2**	22.6
60	80.0	37.9	28.9	58.1	129.6	22.5
80	80.0	38.2	29.0	58.3	128.9	**22.9**

Table 1: Captioning results of $\mathcal{M}^2$ Transformer using different numbers of memory vectors.

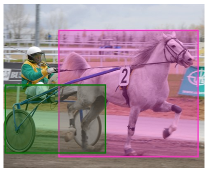
Layers	B-1	B-4	M	R	C	S
2	80.5	38.6	29.0	58.4	128.5	**22.8**
3	**80.8**	**39.1**	**29.2**	**58.6**	**131.2**	22.6
4	**80.8**	38.6	29.1	58.5	129.6	22.6

Table 2: Captioning results of $\mathcal{M}^2$ Transformer using different numbers of encoder and decoder layers.

2. Additional experimental results

Memory vectors. In Table 1, we report the performance of our approach when using a varying number of memory vectors. As it can be seen, the best result in terms of BLEU, METEOR, ROUGE and CIDEr is obtained with 40 memory vectors, while 80 memory vectors provide a slightly superior result in terms of SPICE. Therefore, all experiments in the main paper are carried out with 40 memory vectors.

Encoder and decoder layers. To complement the analysis presented in Sec. 4.3, we also investigate the performance of the $\mathcal{M}^2$ Transformer when changing the number of encoding and decoding layers. Table 2 shows that the best



Constraints: horse; cart.

Transformer: A horse pulling a cart down a street.
$\mathcal{M}^2$ **Transformer:** A white horse pulling a man in a cart.

Constraints: monkey.

Transformer: A brown bear sitting on a rock monkey.
$\mathcal{M}^2$ **Transformer:** A small monkey sitting on a rock in the grass.

Constraints: bookcase.

Transformer: A woman holding a bookcase in a store.
$\mathcal{M}^2$ **Transformer:** A woman holding a book in front of a bookcase.

Constraints: bee; lavender.

Transformer: A bee lavender of purple flowers in a field.
$\mathcal{M}^2$ **Transformer:** A field of lavender purple flowers with bee.

Constraints: flag.

Transformer: A red kite with a flag in the sky.
$\mathcal{M}^2$ **Transformer:** A red and white flag flying in the sky.

Constraints: rabbit.

Transformer: A cat sitting on the rabbit with a cell phone.
$\mathcal{M}^2$ **Transformer:** A rabbit sitting on a table next to a person.

Figure 1: Sample nocaps images and corresponding predicted captions generated by our model and the original Transformer. For each image, we report the Open Images object classes predicted by the object detector and used as constraints during the generation of the caption.

	SPICE	Obj.	Attr.	Rel.	Color	Count	Size
Up-Down [2]	21.4	39.1	10.0	6.5	11.4	18.4	3.2
Transformer	21.1	38.6	9.6	6.3	9.2	17.5	2.0
$\mathcal{M}^2$ **Transformer**	**22.6**	**40.0**	**11.6**	**6.9**	**12.9**	**20.4**	**3.5**

Table 3: Breakdown of SPICE F-scores over various subcategories.

performance is obtained with three encoding and decoding layers, thus confirming the initial findings on the base Transformer model. As our model can deal with a different number of encoding and decoding layers, we also experimented with non symmetric encoding-decoding architectures, without however noticing significant improvements in performance.

SPICE F-scores. Finally, in Table 3 we report a breakdown of SPICE F-scores over various subcategories on the "Karpathy" test split, in comparison with the Up-Down approach [2] and the base Transformer model with three layers. As it can be seen, our model significantly improves on identifying objects, attributes and relationships between objects.

3. Qualitative results and visualization

Figures 2 and 3 show additional qualitative results obtained from our model in comparison to the original Transformer and corresponding ground-truth captions. On average, the proposed model shows an improvement in terms of caption correctness and provides more detailed and exhaustive descriptions.

Figures 4, 5, and 6, instead, report the visualization of attentive states on a variety of sample images, following the approach outlined in Sec. 4.6 of the main paper. Specifically, the Integrated Gradients approach [5] produces an attribution score for each feature channel of each input region. To obtain the attribution of each region, we average over the feature channels, and re-normalize the obtained scores by their sum. For visualization purposes, we apply a contrast stretching function to project scores in the 0-1 interval.

4. Novel object captioning

Figure 1 reports sample captions produced by our approach on images from the nocaps dataset. On each image, we compare to the baseline Transformer and show the constraints provided by the object detector. Overall, the $\mathcal{M}^2$ Transformer is able to better incorporate the constraints while maintaining the fluency and properness of the generated sentences.

Following [1], we use an object detector trained on Open Images [2] and filter detections by removing 39 Open Images classes that contain parts of objects or which are seldom mentioned. We also discard overlapping detections by removing the higher-order of two objects based on the class hierarchy, and we use the top-3 detected objects as constraints based on the detection confidence score. As mentioned in Sec. 4.5 and differently from [1], we do not consider the plural forms or other word phrases of object classes, thus taking into account only the original class names. After decoding, we select the predicted caption with highest probability that satisfies the given constraints.

[2] Specifically, the `tf_faster_rcnn_inception_resnet_v2_atrous_oidv2` model from the Tensorflow model zoo.

GT: A man milking a brown and white cow in barn.
Transformer: A man is standing next to a cow.
$\mathcal{M}^2$ **Transformer:** A man is milking a cow in a barn.

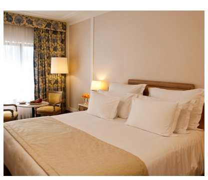
GT: A woman with blue hair and a yellow umbrella.
Transformer: A woman is holding an umbrella.
$\mathcal{M}^2$ **Transformer:** A woman with blue hair holding a yellow umbrella.

GT: Several zebras and other animals grazing in a field.
Transformer: A herd of zebras are standing in a field.
$\mathcal{M}^2$ **Transformer:** A herd of zebras and other animals grazing in a field.

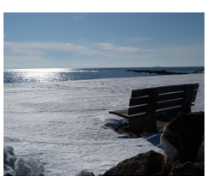
GT: A woman who is skateboarding down the street.
Transformer: A woman walking down a street talking on a cell phone.
$\mathcal{M}^2$ **Transformer:** A woman standing on a skateboard on a street.

GT: Some people are standing in front of a red food truck.
Transformer: A group of people standing in front of a bus.
$\mathcal{M}^2$ **Transformer:** A group of people standing outside of a food truck.

GT: A little girl is eating a hot dog and riding in a shopping cart.
Transformer: A little girl sitting on a bench eating a hot dog.
$\mathcal{M}^2$ **Transformer:** A little girl sitting in a shopping cart eating a hot dog.



GT: A hotel room with a well-made bed, a table, and two chairs.
Transformer: A bedroom with a bed and a table.
$\mathcal{M}^2$ **Transformer:** A hotel room with a large bed with white pillows.



GT: A empty bench on a snow covered beach.
Transformer: Two benches sitting on a beach near the water.
$\mathcal{M}^2$ **Transformer:** A bench sitting on the beach in the snow.

GT: A man riding a small pink motorcycle on a track.
Transformer: A man is riding a red motorcycle.
$\mathcal{M}^2$ **Transformer:** A man riding a pink motorcycle on a track.

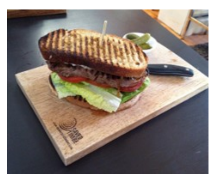
GT: A man in a red Santa hat and a dog pose in front of a Christmas tree.
Transformer: A Christmas tree in the snow with a Christmas tree.
$\mathcal{M}^2$ **Transformer:** A man wearing a Santa hat with a dog in front of a Christmas tree.

GT: Several people standing outside a parked white van.
Transformer: A group of people standing outside of a bus.
$\mathcal{M}^2$ **Transformer:** A group of people standing around a white van.

GT: A truck sitting on a field with kites in the air.
Transformer: A group of cars parked in a field with a kite.
$\mathcal{M}^2$ **Transformer:** A white truck is parked in a field with kites.

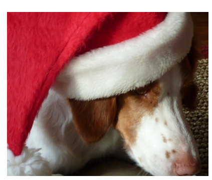
GT: Orange cat walking across two red suitcases stacked on floor.
Transformer: An orange cat sitting on top of a suitcase.
$\mathcal{M}^2$ **Transformer:** An orange cat standing on top of two red suitcases.

GT: A boat parked in a field with long green grass.
Transformer: A field of grass with a fence.
$\mathcal{M}^2$ **Transformer:** A boat in the middle of a field of grass.



GT: A grilled sandwich sits on a cutting board by a knife.
Transformer: A sandwich sitting on top of a wooden table.
$\mathcal{M}^2$ **Transformer:** A sandwich on a cutting board with a knife.

GT: An open toaster oven with a glass dish of food inside.
Transformer: An open suitcase with food in an oven.
$\mathcal{M}^2$ **Transformer:** A toaster oven with a tray of food inside of it.



GT: A brown and white dog wearing a red and white Santa hat.
Transformer: A white dog wearing a red hat.
$\mathcal{M}^2$ **Transformer:** A dog wearing a red and white Santa hat.

GT: Three people sit on a bench looking out over the water.
Transformer: Two people sitting on a bench in the water.
$\mathcal{M}^2$ **Transformer:** Three people sitting on a bench looking at the water.

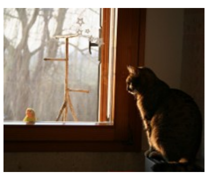
Figure 2: Additional sample results generated by our approach and the original Transformer, as well as the corresponding ground-truths.

GT: Several people with skis on riding down a hill.
Transformer: A group of people on a ski lift in the snow.
$\mathcal{M}^2$ **Transformer:** A group of people skiing down a snow covered slope.

GT: Two girls sitting on some grass are looking at their cellphones.
Transformer: Two women sitting on the ground next to a woman.
$\mathcal{M}^2$ **Transformer:** Two women sitting on the grass looking at their cell phones.

GT: A bunch of apples and cider for sale on a table.
Transformer: A table with many boxes of food on it.
$\mathcal{M}^2$ **Transformer:** A table of apples for sale at an outdoor market.

GT: Motorcycle police officer riding through the city.
Transformer: A man riding a motorcycle down a street.
$\mathcal{M}^2$ **Transformer:** A police officer riding a motorcycle down a street.



GT: A cat watching a small bird through a window.
Transformer: Two cats sitting on a window sill looking out.
$\mathcal{M}^2$ **Transformer:** A cat looking out of a window at a bird.

GT: A bowl of bananas sitting on the kitchen table.
Transformer: A kitchen with a wooden table and chairs.
$\mathcal{M}^2$ **Transformer:** A kitchen with a bowl of bananas on the table.

GT: A yellow truck with graffiti painted on it.
Transformer: A yellow truck is parked in a parking lot.
$\mathcal{M}^2$ **Transformer:** A yellow truck with graffiti on it parked in a field.

GT: A bed that has two pillows and a quilt on it.
Transformer: A bed with white sheets and pillows.
$\mathcal{M}^2$ **Transformer:** A bed with a quilt on it in a room.

GT: A fire hydrant spraying water on a street.
Transformer: A red fire hydrant on the side of a street.
$\mathcal{M}^2$ **Transformer:** A red fire hydrant spraying water on a street.

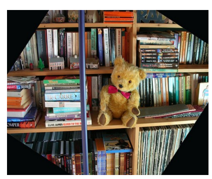
GT: Two women on a tennis court shaking hands.
Transformer: Two women standing on a tennis court.
$\mathcal{M}^2$ **Transformer:** Two women standing on a tennis court shaking hands.

GT: A dog sitting on the side of an area where boats are docked in the water.
Transformer: A black dog sitting on a boat in the water.
$\mathcal{M}^2$ **Transformer:** A dog sitting on a dock with boats in the water.

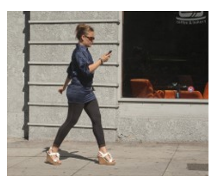
GT: A group of people riding a miniature train.
Transformer: A group of people sitting on a bench.
$\mathcal{M}^2$ **Transformer:** A group of people sitting on a small wooden train.

GT: A double decker bus is parked at a bus stop.
Transformer: A red double decker bus driving down a street.
$\mathcal{M}^2$ **Transformer:** A double decker bus parked at a bus stop.

GT: A white high speed train at a train station.
Transformer: A white airplane is parked on the runway at an airport.
$\mathcal{M}^2$ **Transformer:** A white high speed train at a train station.



GT: A stuffed teddy bear sitting on top of a book shelf.
Transformer: A book shelf filled with lots of books in a library.
$\mathcal{M}^2$ **Transformer:** A teddy bear sitting on top of a book shelf.



GT: A woman walking down the sidewalk with a cell phone in her hand.
Transformer: A woman talking on a cell phone.
$\mathcal{M}^2$ **Transformer:** A woman walking down a sidewalk while looking at her cell phone.

GT: A person in a yellow rain jacket standing next to a parking meter.
Transformer: A parking meter on the side of a street.
$\mathcal{M}^2$ **Transformer:** A person in a yellow jacket standing next to a parking meter.

GT: Three mountain goats standing on top of a mountain.
Transformer: Two goats sitting on a rock in a field.
$\mathcal{M}^2$ **Transformer:** Three mountain goats standing on a grassy hill.

Figure 3: Additional sample results generated by our approach and the original Transformer, as well as the corresponding ground-truths.

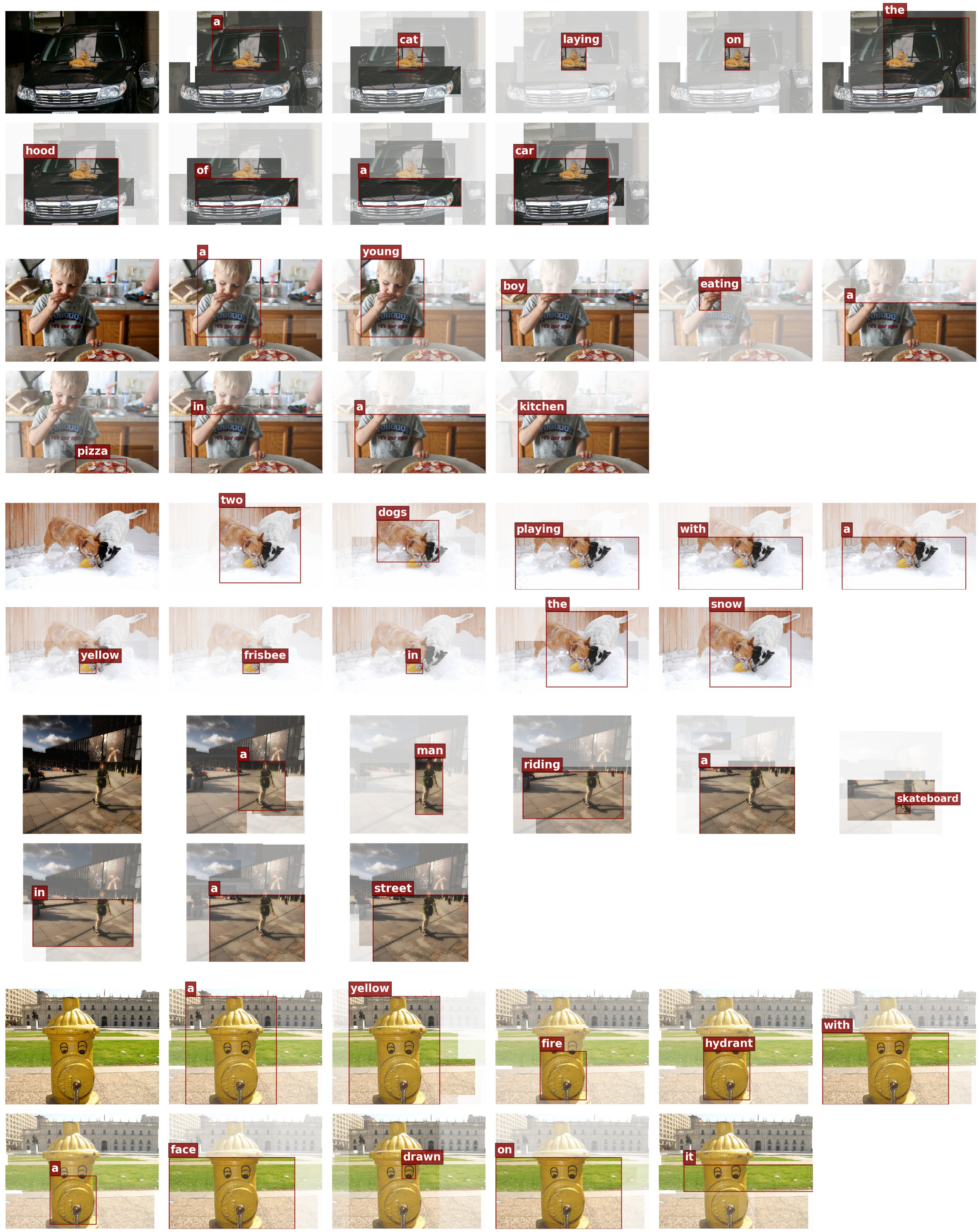


Figure 4: Visualization of attention states for sample captions generated by our $\mathcal{M}^2$ Transformer. For each generated word, we show the attended image regions, outlining the region with the maximum output attribution in red.

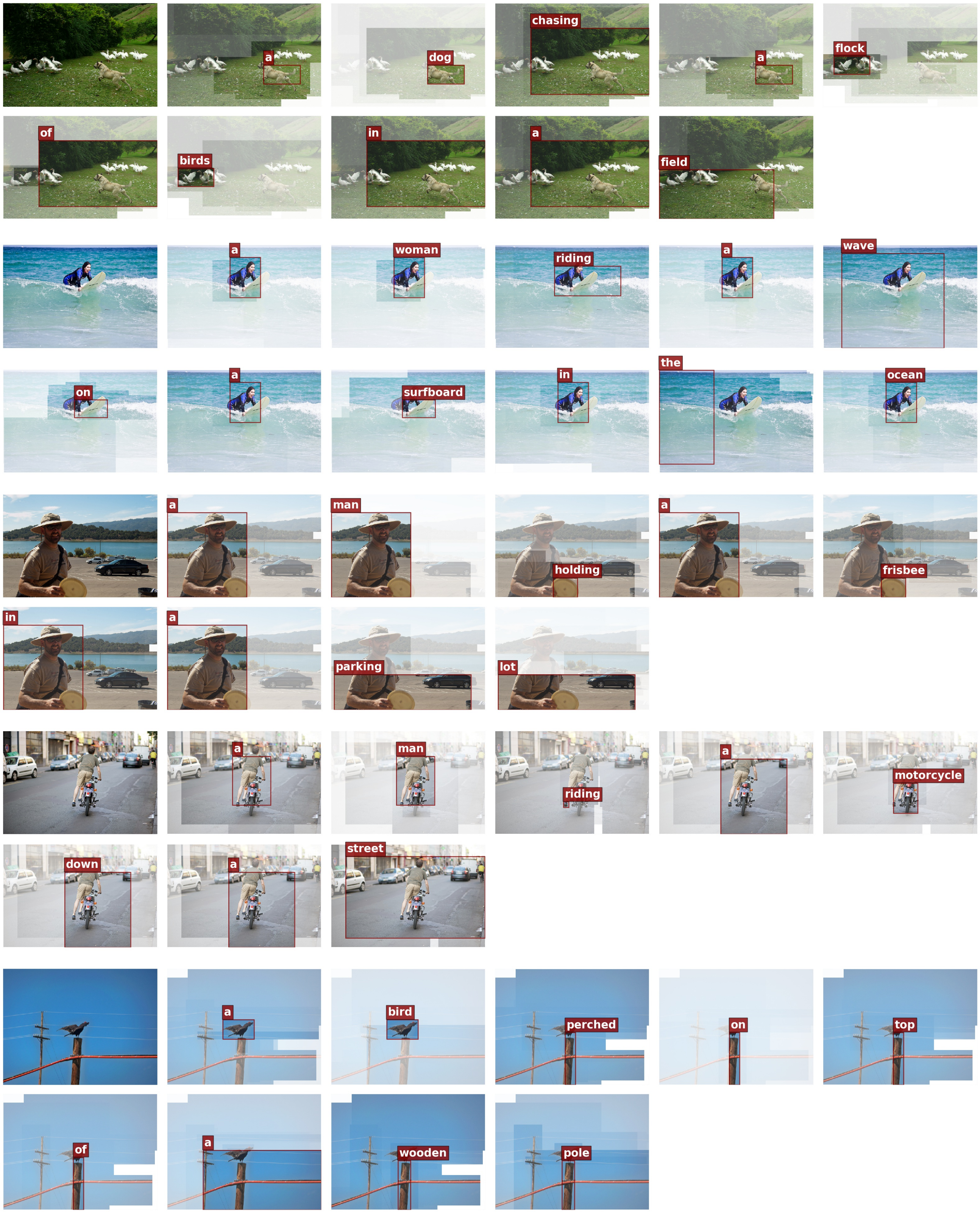


Figure 5: Visualization of attention states for sample captions generated by our $\mathcal{M}^2$ Transformer. For each generated word, we show the attended image regions, outlining the region with the maximum output attribution in red.

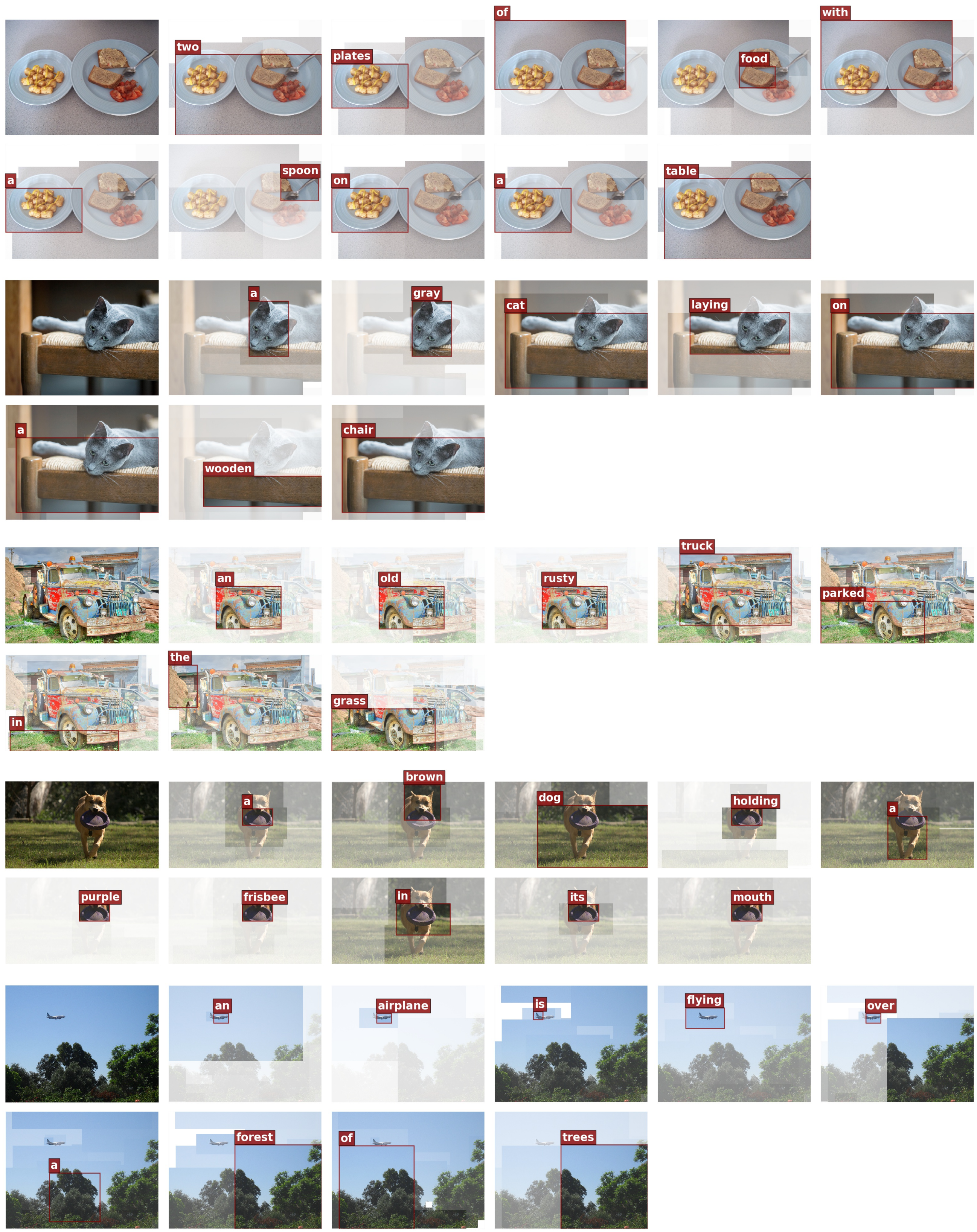


Figure 6: Visualization of attention states for sample captions generated by our $\mathcal{M}^2$ Transformer. For each generated word, we show the attended image regions, outlining the region with the maximum output attribution in red.

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 *Proceedings of the International Conference on Computer Vision*, 2019.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[3] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2010.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[5] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning*, 2017.