The Summary of ”Multimodal Transformer with Multi-View Visual Representation for Image Captioning” [1] Paper

Furkan Gül

I. Motivation

Thanks to recent developments in deep learning, computer vision and natural language processing communities have a great chance to increase their capability to a higher level of performance. Beyond the advances in those two fundamental eras of deep learning, recent achievements at the combination of computer vision and natural language processing have enabled several multimodal learning tasks to be resolved such as image-text matching, image captioning, and visual question and answering (VQA). Image captioning intends to describe an input image in terms of its contents by a natural language sentence. In addition to recognizing objects in an image, it must infer the relationships between those objects into a natural language sentence. That's why image captioning is considered a challenging task. Up until now, the best performance for image captioning task had been obtained by inserting an attention mechanism into the traditional encoder-decoder model framework (Seq2seq model for machine translation).

Despite the greater success in image captioning, there exists some limitations in the recently proposed approaches. These can be counted as the following: 1- current attention models limited to object-to-word (inter-modal) interactions while ignoring the object-to-object and word-to-word (intra-modal) interactions; 2- since current models for image captioning are generally shallow, they can not be successful at fully capturing the complex relations among visual objects; 3- all visual objects in an image could not be recognized by the region-based visual features which in return results in insufficient visual representations for creating correct captions for images. In the paper, they try to overcome the first two limitations by extending the machine translation transformer model [2] to a Multimodal Transformer (MT) model for image captioning. For the last limitation, they add multi-view feature learning into the standard Multimodal Transformer.

II. Approach

A. Multimodal Transformer: Image Encoder & Caption Decoder

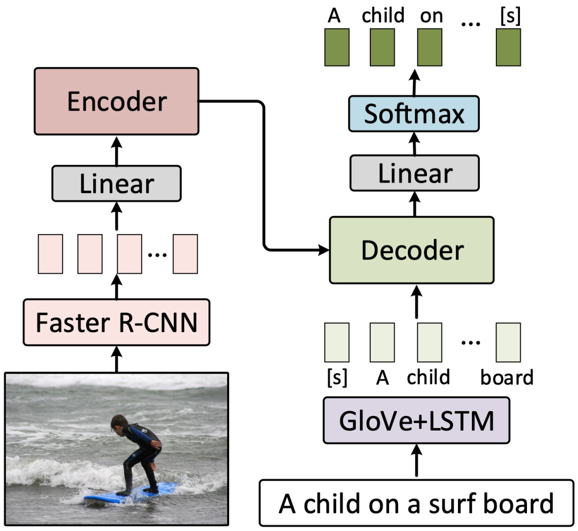


Fig. 1. MT model for image captioning. It includes an image encoder to learn self-attended visual features, and a caption decoder to generate the image caption from the attended visual features [1].

They suggest an MT framework for image captioning. It includes an image encoding part and a textual decoding part. In the image encoder part, an input image is processed through a pre-trained Faster-RCNN [3] on Visual Genome dataset [4] to extract a deep region-based representation of images. Then, those extracted visual features are given to the encoder part as an input to capture the attended visual representation by using a self-attention manner. Those attended visual representations in the result of this process can be called single-view features since there is only one object detector which is Faster-RCNN. In the caption decoder part, using the attended visual representations from the encoder, textual captions for images are generated. Each word in the input caption is transformed into a vector embedding by using the 300-D GloVe vector embedding [5] which is pre-trained on a large corpus. In other words, the caption for the input image together with its attended image features obtained by using both Faster-RCNN and self-attention mechanism are given into the caption model to obtain the word with the largest probability among all other words in the vocabulary by using softmax cross-entropy loss. The model framework for Multimodel Transformer can be seen in Figure 1.

B. Multi-View Image Encoders

In order to increase the representation power of the Multimodel Transformer model, the multi-view image representation is adopted into the image encoder part. Since the visual objects identified with different detectors are naturally unaligned which causes the whole process of learning the correspondence over different views to be more challenging, they propose two different multi-view image encoder models: the Unaligned Multi-View (UMV) image encoder and the Aligned Multi-View (AMV) image encoder.

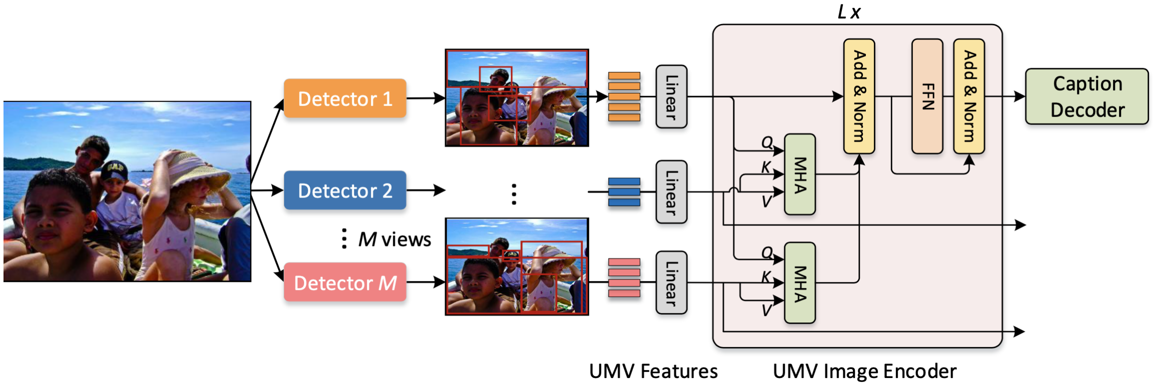


Fig. 2. Framework of the unaligned multi-view image encoder model [1].

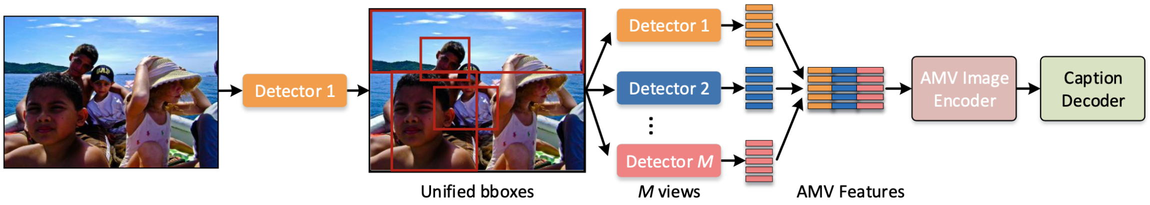


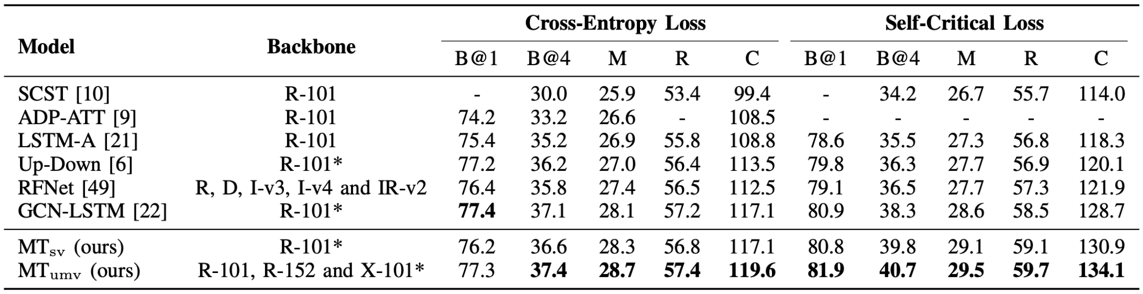
Fig. 3. Framework of the aligned multi-view image encoder model [1].

As it can be seen from Figure 2; UMV image encoder takes an image as input and extracts unaligned multi-view features from different detectors in parallel, then those UMV features are fed into UMV image encoder model to give the output of the attended features with adaptive alignment learning to the caption decoder as an input (which is exactly the same as the one for single-view features mentioned in earlier).

When the aligned multi-view image encoder model is considered as it is in Figure 3, AMV image encoder takes an image as input and regards outputs of different object detectors as the multiple views. To generate the aligned multi-view features, M object detectors are chosen to predict unified bounding boxes for objects in an input image. Then, those bounding boxes are used to extract AMV features. Finally outputted AMV features are fed into AMV image encoder model to give the attended features to the caption decoder as an input.

III. Dataset & Results

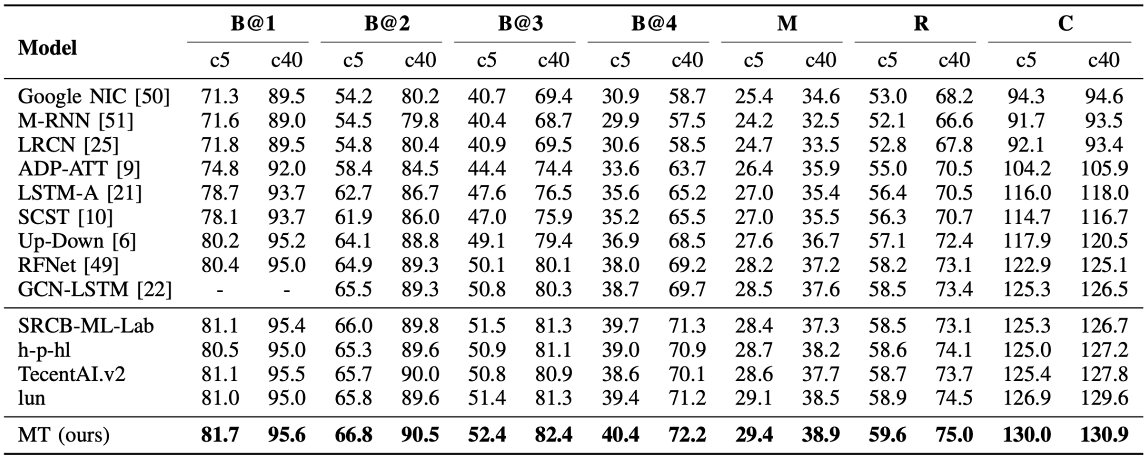
MSCOCO [6] dataset is used in their experiments to evaluate the suggested captioning model. There are at least 5 sentence descriptions for every image in this dataset. They use Karparthy split [7] which is extensively preferred for reporting results in previous works. For both the validation and test phases with MSCOCO dataset, 5000 images are used. They convert all caption sentences to lower case, tokenize on white space, and discard words that have less than five occurrences or do not exist in the pre-trained GloVe vocabulary. Thus, the vocabulary size for their experiments is fixed and 9,343. For the evaluation of the caption model, following metrics are used: CIDEr [8], METEOR [9], ROUGE-L [10], and BLEU [11].



Model	Backbone	Cross-Entropy Loss					Self-Critical Loss				
		B@1	B@4	M	R	C	B@1	B@4	M	R	C
SCST [10]	R-101	-	30.0	25.9	53.4	99.4	-	34.2	26.7	55.7	114.0
ADP-ATT [9]	R-101	74.2	33.2	26.6	-	108.5	-	-	-	-	-
LSTM-A [21]	R-101	75.4	35.2	26.9	55.8	108.8	78.6	35.5	27.3	56.8	118.3
Up-Down [6]	R-101*	77.2	36.2	27.0	56.4	113.5	79.8	36.3	27.7	56.9	120.1
RFNet [49]	R, D, I-v3, I-v4 and IR-v2	76.4	35.8	27.4	56.5	112.5	79.1	36.5	27.7	57.3	121.9
GCN-LSTM [22]	R-101*	**77.4**	37.1	28.1	57.2	117.1	80.9	38.3	28.6	58.5	128.7
$\mathrm{MT}_{\mathrm{sv}}$ (ours)	R-101*	76.2	36.6	28.3	56.8	117.1	80.8	39.8	29.1	59.1	130.9
$\mathrm{MT}_{\mathrm{umv}}$ (ours)	R-101, R-152 and X-101*	77.3	**37.4**	**28.7**	**57.4**	**119.6**	**81.9**	**40.7**	**29.5**	**59.7**	**134.1**

Fig. 4. Single-model image captioning performance on the MSCOCO Karpathy test split. The methods marked with * denote using the bottom-up-attention visual features from a pre-trained Faster R-CNN model. R, D, I-v3, I-v4 and IR-v2 denotes the ResNet, DenseNet, Inception-v3, Inception-v4 and Inception-ResNet-v2 model, respectively. [1]

For the mult-view image encoder part, they trained three Faster-RCNN models (*i.e.* the number of views, M, is 3) with different backbones: ResNet-101 [12], ResNet-152 [12], and ResNeXt-101 [13]. The hyperparameters of the MT models are as follows: 1- the dimentionality of input image features and input caption features are 2048 and 512 respectively, 2- latent dimensionality of multi-head attention model (MHA) is 512, the number of head is 8, and latent dimensionality of each head is 64. 3- the number of attention blocks in the encoder and decoder parts are 1, 2, 4, 6, and 8. For the optimization, Adam [14] is used with a batch size of 10.



Model	B@1		B@2		B@3		B@4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Google NIC [50]	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6
M-RNN [51]	71.6	89.0	54.5	79.8	40.4	68.7	29.9	57.5	24.2	32.5	52.1	66.6	91.7	93.5
LRCN [25]	71.8	89.5	54.8	80.4	40.9	69.5	30.6	58.5	24.7	33.5	52.8	67.8	92.1	93.4
ADP-ATT [9]	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9
LSTM-A [21]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
SCST [10]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	65.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-Down [6]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet [49]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
GCN-LSTM [22]	-	-	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SRCB-ML-Lab	81.1	95.4	66.0	89.8	51.5	81.3	39.7	71.3	28.4	37.3	58.5	73.1	125.3	126.7
h-p-hl	80.5	95.0	65.3	89.6	50.9	81.1	39.0	70.9	28.7	38.2	58.6	74.1	125.0	127.2
TecentAI.v2	81.1	95.5	65.7	90.0	50.8	80.9	38.6	70.1	28.6	37.7	58.7	73.7	125.4	127.8
lun	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
MT (ours)	**81.7**	**95.6**	**66.8**	**90.5**	**52.4**	**82.4**	**40.4**	**72.2**	**29.4**	**38.9**	**59.6**	**75.0**	**130.0**	**130.9**

Fig. 5. Real-time leaderboard of the state-of-the-art solutions on the online MSCOCO test server (April 21st, 2019). The first split shows the published solutions while the second split shows the unpublished ones. [1]

In Figure 4, they reports their results together with the SCST [15] , ADP-ATT [16], LSTM-A [17], Up-Down [18] and GCN-LSTM [19] on the Karpathy test split. $\boldsymbol{MT_{sv}}$, $\boldsymbol{MT_{amv}}$, and $\boldsymbol{MT_{umv}}$ are stands for the single-view MT model, the aligned multi-view MT model, and the unaligned multi-view MT model respectively. As it can be seen in Figure 4, $\boldsymbol{MT_{umv}}$ model outperforms all SOTA models and becomes the new SOTA single-model for Karpathy split in terms of all evaluation metrics.

In the Figure 5, they report the performance of seven MT ensembled models (the $\boldsymbol{MT_{sv}}$, $\boldsymbol{MT_{amv}}$, and $\boldsymbol{MT_{umv}}$ models with different random seeds) trained on the official MSCOCO evaluation server. At the time of submission, 21 April 2010, they take the lead in the leader board by outperforming all other submissions on the test server in terms of all metrics.

IV. A Couple of Words

I could not get into the details of the equations of Multimodal Transformer model since it is not fully covered in the class. But I will refer to this section at a later time. The next paper to be examined and summarized will be ”EfficientDet: Scalable and Efficient Object Detection” [20] from 2020 since it may be the backbone for the image encoder of my model implementation.

References

[1] J. Yu, Jing Li, Zhou Yu, and Q. Huang. Multimodal transformer with multi-view visual representation for image captioning. *ArXiv*, abs/1905.07841, 2019.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[3] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[4] R. Krishna, Yuke Zhu, O. Groth, J. Johnson, Kenji Hata, J. Kravitz, Stephanie Chen, Yannis Kalantidis, L. Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016.

[5] Jeffrey Pennington, R. Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[6] Tsung-Yi Lin, M. Maire, Serge J. Belongie, J. Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014.

[7] A. Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676, 2017.

[8] Ramakrishna Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.

[9] Michael J. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT@ACL*, 2014.

[10] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*, 2004.

[11] Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[12] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[13] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[15] Steven J. Rennie, E. Marcheret, Youssef Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195, 2017.

[16] Jiasen Lu, Caiming Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3250, 2017.

[17] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and T. Mei. Boosting image captioning with attributes. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4904–4912, 2017.

[18] Peter Anderson, X. He, C. Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[19] Ting Yao, Yingwei Pan, Yehao Li, and T. Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.

[20] Mingxing Tan, R. Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10778–10787, 2020.