Modulated Self-attention Convolutional Network for VQA

Jean-Benoit Delbrouck[1]   **Antoine Maiorca**[1]   **Nathan Hubens**[1,2]   **Stéphane Dupont**[1]

[1] ISIA Lab, Polytechnic Mons, Belgium   [2] Telecom SudParis, France

{jean-benoit.delbrouck, antoine.maiorca, nathan.hubens, stephane.dupont}@umons.ac.be

Abstract

As new data-sets for real-world visual reasoning and compositional question answering are emerging, it might be needed to use the visual feature extraction as a end-to-end process during training. This small contribution aims to suggest new ideas to improve the visual processing of traditional convolutional network for visual question answering (VQA). In this paper, we propose to modulate by a linguistic input a CNN augmented with self-attention. We show encouraging relative improvements for future research in this direction.

1   Introduction

Problems combining vision and natural language processing such as visual question answering (VQA) is viewed as an extremely challenging task. Visual attention-based neural decoder models [1, 2] have been widely adopted to solve such tasks. All recent works pushing the state-of-the-art in VQA are using the so-called bottom-up attention features [3]. It consists of multiple pre-extracted features corresponding to local regions of interest in an image. Therefore, it is not possible to modulate the entire visual pipeline to capture relationships or make comparisons between widely separated spatial regions according to the question. For example, questions of the new GQA data-set [4] tackles reasoning skills such as object and attribute recognition, transitive relation tracking, spatial reasoning, logical inference and comparisons. This contribution aims to bring back the features extraction process end-to-end in the learning and to modulate the convolutional network by a linguistic input though self-attention.

2   Related work

Two main approaches have helped pushing further the state-of-the-art in Visual Question Answering: the co-attention module and the bottom up image features. Our work also takes inspiration in the recent attention augmented Convolutional Networks.

Image features   Most early results have used a VGG [5] or a ResNet [6] CNN pretrained on ImageNet [7] to extract visual features from images. Recently, richer bottom-up attention features [3] have been proposed. It consists of pre-extracted from a Faster R-CNN (that outputs bounding boxes of interest) in conjunction with ResNet-101 (that acts as a feature extractor for each selected bounding box). They are widely used in the most recent works as they have shown great success for VQA and image captioning.

Co-attention   Attention networks in multimodal learning provide an efficient way to utilize given visual information selectively. However, the computational cost to learn attention distributions for every pair of multimodal input channels is prohibitively expensive. To solve this problem, a co-attention learning framework is proposed [8] to reduce the expensive computational cost to learn attention distributions for every pair of multimodal input channels. An improvement to the

co-attention method is introduced by [9] that consists of two steps, a self-attention for a question embedding and the question-conditioned attention for a visual embedding. However, these co-attention approaches use separate attention distributions for each modality, neglecting the interaction between the modalities what we consider and model. To tackle this problem, [10] propose bilinear attention networks that find bilinear attention distributions to effectively utilize vision-language information. Finally, [11] propose a Deep Modular Co-Attention Networks composed of multiple blocks of two basic attention units : a self-attention unit (inspired by machine translation [12]) and the guided-attention (GA) unit to model the dense inter-modal interactions. The authors also concatenate the res5c features from ResNet-152 to the bottom up features from [3].

Attention augmented CNN   Over the years, multiple attention mechanisms have been proposed for visual tasks to address the weaknesses of convolutions. For instance, Squeeze-and-Excitation networks [13] reweight feature channels using signals aggregated from entire feature maps while CBAM [14] module sequentially infers attention maps along two separate the channel and spatial independently. More recent attention augmented networks produces additional feature maps by using Attention Augmented Convolution [15] or recalibrates the feature maps via addition [16].

3   Attention augmented Residual Network

3.1   Residual Network

ResNets [17] are built from residual blocks:

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}, \{\boldsymbol{W}_i\}) + \boldsymbol{W}_s \boldsymbol{x}$$

Here, $\boldsymbol{x}$ and $\boldsymbol{y}$ are the input and output vectors of the layers considered. The function $\mathcal{F}(\boldsymbol{x}, \{\boldsymbol{W}_i\})$ is the residual mapping to be learned. For an example, if we consider two layers, $\mathcal{F} = \boldsymbol{W}_2 \sigma(\boldsymbol{W}_1 \boldsymbol{x})$ where σ denotes ReLu function. The operation $\mathcal{F} + \boldsymbol{x}$ is the shortcut connection and consists of an element-wise addition. Therefore, the dimensions of $\boldsymbol{x}$ and $\mathcal{F}$ must be equal. When this is not the case (e.g., when changing the input/output channels), the $\boldsymbol{W}_s$ matrix performs a linear projection by the shortcut connections to match the dimension. Finally, it performs a last second nonlinearity after the addition (i.e., $\sigma(y)$). A group of blocks are stacked to form a stage of computation. The general ResNet architecture starts with a single convolutional layer followed by 4 stages of computation.

3.2   Self-attention layer

A self-attention layer can be placed in-between two layers in a residual block. Let's consider the image features from the previous layer: $\boldsymbol{x} \in \mathbb{R}^{C \times N}$. Here, C is the number of channels and N is the number of feature locations of features.

Features x are first transformed into two feature spaces $\boldsymbol{f}, \boldsymbol{g}$, where $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{W}_f \boldsymbol{x}, \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{W}_g \boldsymbol{x}$. We use $\boldsymbol{f}$ and $\boldsymbol{g}$ to compute the square matrix $N \times N$ of attention weights β:

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})}, \text{ where } s_{ij} = \boldsymbol{f}(\boldsymbol{x}_i)^T \boldsymbol{g}(\boldsymbol{x}_j) \tag{1}$$

With $\beta_{j,i}$ indicating the extent to which the model attends to the ith location when synthesizing the jth location. The output of the attention layer is $\boldsymbol{o} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_j, \ldots, \boldsymbol{o}_N) \in \mathbb{R}^{N \times C}$, where

$$\boldsymbol{o}_j = \sum_{i=1}^{N} \beta_{j,i} \boldsymbol{k}(\boldsymbol{x}_i), \boldsymbol{k}(\boldsymbol{x}_i) = \boldsymbol{W}_k \boldsymbol{x}_i \tag{2}$$

In the above formulation, $\boldsymbol{W}_g \in \mathbb{R}^{C \times \overline{C}}, \boldsymbol{W}_f \in \mathbb{R}^{C \times \overline{C}} \boldsymbol{W}_k \in \mathbb{R}^{C \times C}$ are the learned weight matrices, which are implemented as 1×1 convolutions.

Additionally, the output is multiplied by a factor γ and we add back the input feature map.

$$\boldsymbol{y}_i = \gamma \boldsymbol{o}_i + \boldsymbol{x}_i \qquad (3)$$

γ is a learnable scalar and it is initialized as 0. Introducing the learnable γ allows the visual network to first rely on the cues in the local neighborhood and the language model to converge normally. As the gamma goes up, the model will gradually learn long range interactions as required for a task such as VQA.

4 Linguistic modulation

In this section, we described two methods to modulate a pretrained ResNet through the self attention modules. We want to enable the linguistic embedding to manipulate the self-attention mechanism. We decide to do so through the parameter γ and the attention weights β.

γ **modulation**   Given the last hidden state $\boldsymbol{h}_T$ of the RNN encoding the question, we output a modulated γ^{m} given by :

$$\gamma^h = \boldsymbol{W}_h \boldsymbol{h}_T \qquad (4)$$

where $\boldsymbol{W}_h \in \mathbb{R}^{|\boldsymbol{h}_T| \times 1}$. Recall that in the previous section, γ in equation 3 is unique for any training examples. Here, we output a dedicated scalar for every example in the batch. We rewrite equation 3 as :

$$\boldsymbol{y}_i = \gamma_i^h \boldsymbol{o}_i + \boldsymbol{x}_i \qquad (5)$$

β **modulation**

We define new linguistic and visual features spaces $\boldsymbol{p}, \boldsymbol{q}$ where $\boldsymbol{p}(\boldsymbol{h}) = \boldsymbol{W}_p \boldsymbol{h}$ and $\boldsymbol{q}(\boldsymbol{x}) = \boldsymbol{W}_q \boldsymbol{x}$. Both feature spaces are used to compute a new set of attention weights β^h in the following manner:

$$\beta_{j,i}^h = \frac{\exp\left(s_{ij}\right)}{\sum_{i=1}^{N} \exp\left(s_{ij}\right)}, \text{ where } s_{ij} = \boldsymbol{q}\left(\boldsymbol{x}_i\right)^T \boldsymbol{p}\left(\boldsymbol{h}_j\right) \qquad (6)$$

Where $\beta_{j,i}^h$ indicating the extent to which the model attends to the ith spatial location of feature map $\boldsymbol{x}$ when synthesizing the jth dimension of the hidden state. The hidden state h_T is a vector therefore $j = 1$ and $\beta^h \in \mathbb{R}^{N \times 1}$. We apply this set of betas to the output $\boldsymbol{o}$ defined in equation 2. This modulation can be seen as a linguistic spatial attention on the visual self attention: each column β_j is re-weighted by the scalar β_j^h.

5 Experiments

Settings   We use the VQA v2.0 train and validation [18] consisting of 443,757 and 214,354 questions respectively over 123,287 images. As VQA model, we follow the Bottom-Up and Top-Down Attention from [3]. We replace the Up-Down features with a ResNet as visual features extractor (ResNet-34 for preliminary experiments and a ResNet-152 [17] for final results). The ResNet is pretrained on ImageNet [7] and frozen during training. Only the self-attention weights are training parameters. For each images, we extract features at the end of the third ResNet stage.

Learning parameters are trained with Adamax optimizer and a learning rate of $2e^{-2}$. In a block, the self-attention module is always placed between the last convolution and batch normalization layer of the said block.

Block and layer search   Thanks to the attention, the model can check that features in distant portions of the image are consistent with each other. Depending on where we place one or several attention modules, the model can compute affinity scores between low-level or high-level features in distant portions of the image. It is also worthy to note that attention module in early stages of ResNet is computationally expensive (i.e. the first stage has $N = 3136$ spatial locations).

ResNet-34	Eval set %	#param	ResNet-34	Eval set %	#param
Baseline (No SA)[3]	55.00	0M	SA (S: 3 - M: 1)	55.25	
SA (S: 1,2,3 - B: 1)	55.11		**SA (S: 3 - B: 3)**	**55.42**	0.082M
SA (S: 1,2,3 - B: 2)	55.17	0.107M	SA (S: 3 - B: 4)	55.33	
SA (S: 1,2,3 - B: 3)	**55.27**		SA (S: 3 - B: 6)	55.31	
			SA (S: 3 - B: 1,3,5)	55.45	0.245M
			SA (S: 3 - B: 2,4,6)	**55.56**	

Table 1: Experiments run on a ResNet-34. Numbers following S (stages) and B (blocks) indicate where SA (self-attention) modules are put. Parameters count concerns only SA and are in millions (M).

We empirically found that self-attention was the most efficient in the 3rd stage. It is also the less computationally expensive [1]. It is also beneficial to focus the last block of a stage rather than the first blocks. We notice small improvements relative to the baseline showing that self-attention alone does improve the VQA task.

Linguistic modulation We noticed that the current architecture was not able to learn the γ **modulation** properly: equation 4 would most of the time result in a large scalar or close to 1. when a sigmoid layer is added. Therefore, the self-attention weight is already too important at the beginning, skipping visual local connections are misleading the overall model training. However, we managed to show improvements with the β **modulation** with a ResNet-152. Though the improvement is slim, it is encouraging to continue researching into visual modulation

ResNet-34	Eval set %	#param	ResNet-152	Eval set %	#param
SA (S: 3 - B: 1,3,5)	**55.56**	0.245M	Baseline (No SA)[3]	57.10	0M
+ γ **modulation**	52.26	0.248M			
+ β **modulation**	55.32	1.573M	SA (S: 3 - B: 2,18,36)	58.05	3.932M
			+ β **modulation**	**58.35**	15.731M

Table 2: Experiments run on a ResNet-34 and 152 for the γ and β modulation.

It is worthy to note that the ResNet baseline of [3] reaches a 57.9% accuracy and was achieved with a ResNet-200. Our best model achieves 58.35% with a ResNet-152.

6 Conclusion

We showed that it is possible to improve the feature extraction procedure for the VQA task by adding self-attention modules in the different ResNet blocks. The improvement margins can be further improved by:

- **Using wider CNN**: Now that we include the feature extraction process end-to-end in the learning, we could pick wider CNN such as ResNeXt [19] or Wide-ResNet [20];
- **Better self-attention**: More sophisticated self-attention, such as multi-head attention in CNN [15, 11] could improve the overall model;
- **Better modulation**: The linguistic hidden state should directly be an input of the computation for self-attention. Other works have already considered using all the hidden states from the language model as the "key" K for visual self-attention [11].
- **Other modulation**: The self-attention modulation could be coupled with modulated batch-normalization as previously investigated in [21].

---

[1]Early stages requires less training parameters as C is smaller, but $f^T g$ computation is more expensive

References

[1] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.

[2] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137. IEEE Computer Society, 2015.

[3] 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*, pages 6077–6086, 2018.

[4] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[6] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[9] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018.

[10] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018.

[11] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[14] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[15] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. *arXiv preprint arXiv:1904.09925*, 2019.

[16] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363, 2019.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[20] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[21] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2017.