Highway Networks for Visual Question Answering

Aaditya Prakash and James Storer
Brandeis University
{aprakash,storer}@brandeis.edu

Abstract

We propose a version of highway network designed for the task of Visual Question Answering. We take inspiration from recent success of Residual Layer Network and Highway Network in learning deep representation of images and fine grained localization of objects. We propose variation in gating mechanism to allow incorporation of word embedding in the information highway. The gate parameters are influenced by the words in the question, which steers the network towards localized feature learning. This achieves the same effect as soft attention via recurrence but allows for faster training using optimized feed-forward techniques. We are able to obtain state-of-the-art[1] results on VQA dataset for Open Ended and Multiple Choice tasks with current model.

1. Introduction

While it is a common understanding in the field that deeper networks are better[3]; in practice training a very deep network for the same size data is prone to overfitting and vanishing gradients. Recently, some success has been observed by techniques which minimize the attenuation of the information by either learning to represent the residues[4] or by adding gating units which reinforce the original signal[12]. This allows successful training of network with thousands of layers.

For the task of visual question answering, we propose a network which alters the gating units to be feed by multi dimensional representation of the words from the question. This allows us to learn common feature space in a single end to end training. When the length of question is shorter than the number of layers, it can be either padded with zeros or by repeating the words from the question. We present our model and result from the VQA dataset[2].

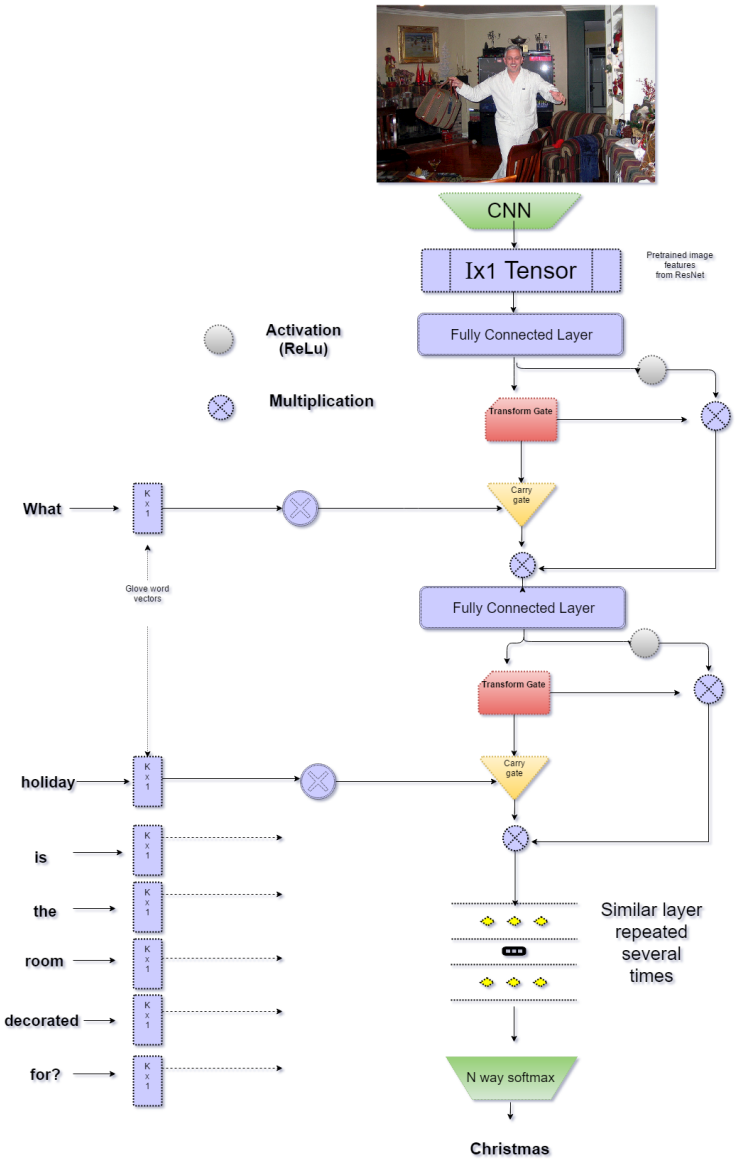

Figure 1. Basic architecture of highway networks for VQA task

[1] among published results. User jw2yang has higher accuracy than us on VQA Challenge leaderboard.

1.1. Highway Networks

For the given input x, and weight matrix W, a feed forward network tries to learn the following representation –

$$y = H(x, W) + \mathbf{b}$$

where, H is a non-linear transformation and $\mathbf{b}$ is bias. Highway Network as described by Srivastava et al[12], adds two gates to above equation, essentially transforming it to –

$$y = H(x, W_H) \cdot T(x, W_T) + x \cdot C(x, W_C)$$

where, T and C are *transform* gate and *carry* gate. It should be noted that x, y, H, T, and C should have same dimensionality.

1.2. Gating for Visual Question Answering

Since the original model was meant for a single learning task like image classification, the authors choose the carry gate $\boldsymbol{C = 1 - T}$, the same cannot be done for multimodal learning like visual question answering.

Our modification for the carry gate is to feed the word vectors of the words from the question, extracted from pre-trained models like Glove vectors[9] . $\boldsymbol{K \times 1}$ features from word embeddings are translated to dimensions of '$\boldsymbol{x}$' by point-wise multiplication. At each layer vector from only one word or average of few words are feed. This allows us to learn temporal relationship in the words in question. However, for this task, individual words in any order seem to perform equally well.

1.3. Feature space

Most of the related work on visual question answering use some form of pre-trained image vector. All the models used are trained on ImageNet images, and the last layer before the softmax is extracted. It is common to use GoogLeNet [5] [16] [19] [13] or VGGNet [14] [15] [18] [11] [10] [1] [6][20] [8]. We use ResNet[4] for their ability to extract better features. Features obtained from VGGNet, GoogLeNet or ResNet are easily transferable to MS COCO Images and number of images in the VQA dataset are not big enough to train a independent image model. However, some of the models do fine-tuning[8], but with our model that is redundant because the *transfer gate* alters the weights of image features effectively doing fine-tuning constrained upon carry gate weights.

1.4. Attention

Most of the recent models [15] [16] [18] tackling the task of visual question answering have taken inspiration from success of soft attention models for image captioning[7] [17]. This has led to some improvement in VQA but comes at a cost of slower training process due to recurrent network.

Table 1. Results VQA Open-Ended and Multiple Choice (MC) on Testdev-2015 and Test-Standard *(first four MC scores are from Testdev)*

	Open Ended					MC
	Test-Dev				Test-Std	
Method	**All**	**Y/N**	**Other**	**Num**	**All**	**All**
Image only	28.1	64.0	3.8	0.4	-	30.5
Question only	48.1	75.7	27.1	36.7	-	53.6
Q+I	52.6	75.6	37.4	33.7	-	58.9
LSTM Q+I	53.7	78.9	36.4	35.2	54.1	57.1
CMV [5]	52.6	78.3	35.9	34.4	-	-
AMA [14]	55.7	79.2	40.1	36.1	56.0	-
iBOW [19]	55.7	76.5	42.6	35.0	55.9	61.9
DPPNet [8]	57.2	80.7	41.7	37.2	57.4	62.6
LCN [1]	57.9	80.5	43.1	37.4	58.0	-
AAA [16]	57.9	80.8	43.2	37.3	58.2	-
dLSTM+ [6]	57.7	80.5	43.0	36.7	58.1	63.0
SAN [18]	58.7	79.3	46.1	36.6	58.9	-
DMN+ [15]	60.3	80.5	48.3	36.8	60.4	-
OUR	**60.4**	**81.5**	**47.6**	**37.2**	**60.7**	**65.0**

While Noh et. al's [8] use of parameter prediction to obtain attention from hashing weights into smaller dimension and Xiong et. al's[15] use of episodic memory for attention avoids recurrence but are limited by size of learnable parameters. Our model achieves the effect of attention by learning the parameters of the gate. And thus it can be extended by adding more layers. *Carry gate* changes the network structure based on the improvement obtained from merging the weights from current word vector. In a control experiment where question words were replaced by random words, the network performance was diminished significantly.

1.5. Bi-directional Recurrence

As an experimental model we appended with input questions in reverse order. This improved the accuracy of the model only slightly but the computational cost was doubled and training took twice the time. Thus we believe there is not much gain in bi-directional recurrence as far as visual question answering is concerned. Similar results were reported for bi-directional LSTM by Ren et al [10].

2. Results

Currently our model is one of the top performers on the VQA challenge. We are still experimenting with the model and the hyper-parameters, and we expect to get better results. Table 1 shows comparison of results. Our best results are average of five models trained on different random seed.

We would like to thank NVIDIA for donating a GPU card for the research.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*, 2016.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[3] J. Ba and R. Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[5] A. Jiang, F. Wang, F. Porikli, and Y. Li. Compositional memory for visual question answering. *arXiv preprint arXiv:1511.05676*, 2015.

[6] D. B. Jiasen Lu, Xiao Lin and D. Parikh. Deeper lstm and normalized cnn visual question answering model. `https://github.com/VT-vision-lab/VQA_LSTM_CNN`, 2015.

[7] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[8] H. Noh, P. H. Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. *arXiv preprint arXiv:1511.05756*, 2015.

[9] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[10] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, pages 2935–2943, 2015.

[11] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. *arXiv preprint arXiv:1511.07394*, 2015.

[12] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[13] Q. Wu, C. Shen, A. v. d. Hengel, P. Wang, and A. Dick. Image captioning and visual question answering based on attributes and their related external knowledge. *arXiv preprint arXiv:1603.02814*, 2016.

[14] Q. Wu, P. Wang, C. Shen, A. v. d. Hengel, and A. Dick. Ask me anything: Free-form visual question answering based on knowledge from external sources. *arXiv preprint arXiv:1511.06973*, 2015.

[15] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. *arXiv preprint arXiv:1603.01417*, 2016.

[16] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *arXiv preprint arXiv:1511.05234*, 2015.

[17] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.

[18] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. *arXiv preprint arXiv:1511.02274*, 2015.

[19] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015.

[20] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. *arXiv preprint arXiv:1511.03416*, 2015.