A Simple Loss Function for Improving the Convergence and Accuracy of Visual Question Answering Models

Ilija Ilievski
Integrative Sciences and Engineering
National University of Singapore
ilija.ilievski@u.nus.edu

Jiashi Feng
Electrical and Computer Engineering
National University of Singapore
elefjia@nus.edu.sg

Abstract

Visual question answering as recently proposed multimodal learning task has enjoyed wide attention from the deep learning community. Lately, the focus was on developing new representation fusion methods and attention mechanisms to achieve superior performance. On the other hand, very little focus has been put on the models' loss function, arguably one of the most important aspects of training deep learning models. The prevailing practice is to use cross entropy loss function that penalizes the probability given to all the answers in the vocabulary except the single most common answer for the particular question. However, the VQA evaluation function compares the predicted answer with all the ground-truth answers for the given question and if there is a matching, a partial point is given. This causes a discrepancy between the model's cross entropy loss and the model's accuracy as calculated by the VQA evaluation function. In this work, we propose a novel loss, termed as soft cross entropy, that considers all ground-truth answers and thus reduces the loss – accuracy discrepancy. The proposed loss leads to an improved training convergence of VQA models and an increase in accuracy as much as 1.6%.

1. Introduction

Visual question answering (VQA) requires an AI agent to answer questions about an image. As a challenging multimodal problem and a proxy task for visual reasoning, it has attracted a lot of attention from the deep learning community. Multiple models were introduced [6, 16, 10] and a new dataset with a specific focus on visual reasoning [9].

The currently largest VQA dataset, VQA v2.0 [3], contains 1.1 million questions for the 205 thousand MS COCO images [13]. Each question is paired with ten human-provided answers. The usual VQA model uses a pretrained ResNet [4] network to obtain an image representation and an LSTM [5] unit to learn a representation of the question words. The model then fuses the two representations into a single multimodal representation via element-wise multiplication or other more sophisticated methods. Finally, the most common answer out of the ten provided is used to train the model to classify the multimodal representation to a correct answer [20, 7, 19, 17, 18].

Recently, several representation fusion methods were developed [2, 11, 1] and some novel attention mechanisms were introduced [15, 14]. But, very little attention has been put on the VQA model loss function, which is an essential part of its training. The prevailing approach is to use the most common answer and a cross entropy loss function (Eq. (1)). However, a VQA model is evaluated by comparing the predicted answer with *all* the ground-truth answers for a given question and if there is a match, a partial point is given. This causes a discrepancy between the model's cross entropy loss and the model's accuracy as calculated by the VQA evaluation function, which in turn results in a delayed training convergence and reduced test accuracy.

In this work, we propose a new loss function, termed as *soft* cross entropy, that considers *all* ground-truth answers and thus solves the discrepancy problem. In contrast to the standard cross entropy loss, the soft cross entropy loss provides to the model a set of plausible answers for a given question and information about the question's ambiguity. As a consequence, the VQA models trained with the proposed loss have a stable training process, converge faster, and achieve on average 1.5% higher accuracy than models trained with the standard cross entropy loss function.

In summary, the contributions of this work are:

- We propose a novel loss function for VQA, that more closely reflects a VQA model's performance. The proposed loss is justified with error analysis and empirical evaluation.

- We provide an efficient code for reproducing the experimental results and to serve as a starter code to the VQA community.

Table 1. Best validation accuracy on the VQA v2.0 validation set.

	Loss Function	All	Y/N	Num	Other
AVG	Cross Entropy	46.8	55.8	29.8	42.4
	Soft Cross Entropy	**48.0**	**57.1**	**31.0**	**43.3**
POOL	Cross Entropy	58.8	70.1	37.5	53.1
	Soft Cross Entropy	**60.4**	**71.9**	**39.0**	**54.6**

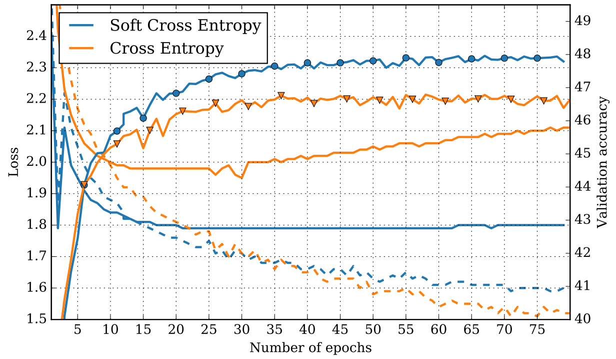

Figure 1. Training (dashed lines) and validation (solid lines) loss and validation accuracy (dotted lines) for the **AVG** VQA model.

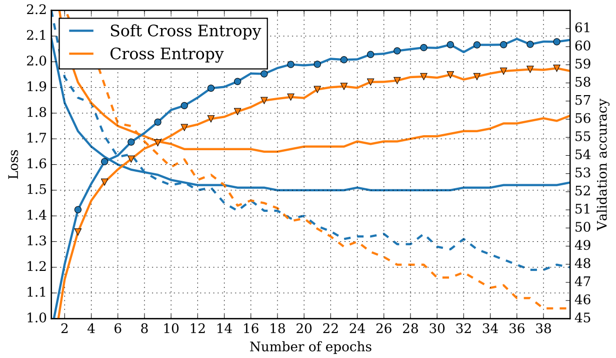

Figure 2. Training (dashed lines) and validation (solid lines) loss and validation accuracy (dotted lines) for the **POOL** VQA model.

2. Soft Cross Entropy Loss

The VQA problem can be reduced to a maximum likelihood estimation problem, where the model classifies a question-image pair to an answer from the training set. Generally, a deep learning model is trained on a classification problem using a cross entropy loss function:

$$\mathcal{L}(\boldsymbol{x}, c^*) = -\boldsymbol{x}_{c^*} + \log\big(\sum_{j=1}^{|\boldsymbol{x}|} \exp(\boldsymbol{x}_j)\big), \tag{1}$$

where $\boldsymbol{x}$ is a vector of network activations for each class and c^* is the index of the correct class.

However, contrary to conventional classification problems, the VQA evaluation metric considers a predicted answer as correct if the answer was given by at least three out of ten human annotators. The accuracy is then averaged over all $\binom{10}{9}$ subsets of ground-truth answers:

$$Acc(a) = \frac{1}{10}\sum_{k=1}^{10} \min\Big(\frac{\sum_{j=1, j\neq k}^{10} \mathbb{1}(a = a_j)}{3}, 1\Big).$$

As a result, the model's performance is not properly assessed with the cross entropy loss function during the training phase. The improper loss function has significant negative impact on the model's training and delays the convergence. Furthermore, it results in abnormal and counter-intuitive validation loss – accuracy relationship where both the loss and the accuracy increase (Fig. 1 and 2).

As a solution, we propose to use a loss function that considers *all* ground-truth answers. The proposed loss function, termed as *soft* cross entropy, is a simple weighted average of each unique ground-truth answer:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{w}) = \sum_{i=1}^{|\boldsymbol{c}|} w_i \Big(-\boldsymbol{x}_{c_i} + \log\big(\sum_{j=1}^{|\boldsymbol{x}|} \exp(\boldsymbol{x}_j)\big)\Big), \tag{2}$$

where $\boldsymbol{c}$ is a vector of unique ground-truth answers and $\boldsymbol{w}$ is a vector of answer weights computed as the number of times the unique answer appears in the ground-truth set divided by the total number of answers.

3. Experiments

We evaluate the proposed loss function on the recently released VQA v2.0 benchmark dataset [3]. To demonstrate the general applicability of the proposed loss we train a variant of the two most common VQA models [8, 11].

Both models use an LSTM to encode the question words to a vector representation. The **AVG** model is based on [8] and it utilizes the activations of the penultimate layer of pre-trained ResNet-152[4] as image representation and does not employ attention mechanism. The **POOL** model is based on [11] and it considers the tensor of activations of the last pooling layer of the same ResNet and employs attention mechanism over the regions to obtain an image representation in a vector form. Both models are trained with Adam [12] and a batch size of 64[1].

4. Discussion and Conclusion

From Table 1 we can observe that the proposed loss increases the overall accuracy by 1.2% in the simpler model and 1.6% increase in the pooling model. The accuracy is increased for both models and all answer types which proves the general applicability of the soft cross entropy loss.

In Figures 1 and 2 we can clearly observe the abnormal relationship between the validation loss and accuracy where they both start to increase near the half of the training process. Furthermore, we can observe how the cross entropy loss rapidly reduces on the training set without an increase in validation accuracy and a decrease in validation loss.

The evaluation results show that by modeling the VQA evaluation metric more faithfully than conventional classification loss functions, the proposed loss function is able to bring a consistent increase in accuracy for VQA models.

[1] Code available at `github.com/ilija139/vqa-soft`

References

[1] H. Ben-Younes, R. Cadène, N. Thome, and M. Cord. MUTAN: Multimodal Tucker fusion for visual question answering.

[2] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[3] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. 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.

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

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. *arXiv preprint arXiv:1704.05526*, 2017.

[7] I. Ilievski, S. Yan, and J. Feng. A focused dynamic attention model for visual question answering. *arXiv preprint arXiv:1604.01485*, 2016.

[8] 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.

[9] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint arXiv:1612.06890*, 2016.

[10] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Inferring and executing programs for visual reasoning. *arXiv preprint arXiv:1705.03633*, 2017.

[11] J.-H. Kim, K.-W. On, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016.

[12] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[14] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[15] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471*, 2016.

[16] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427*, 2017.

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

[18] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. *arXiv*, 1603, 2016.

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

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