Baré Alexandre (r0912072)

Visual Question Answering - Part I

H02C8b: Information Retrieval and Search Engines

Katholiek Universiteit Leuven
Faculty of Engineering Science
Academic Year 2021-2022

I. INTRODUCTION

In this work, we tackle the task of visual question answering for binary answers - Yes/No - with a subset of the COCO dataset (Lin et al. 2014) on sport activities. We mainly focus on the fusion scheme of the question and image representations.

II. RELATED WORK

Many different approaches to combine textual and visual representations exist. To only list a few, Malinowski, Rohrbach, and Fritz (2016) compared the most common options: features concatenation, element-wise addition and element-wise multiplication. The latter was reported as the best performing approach. Saito et al. (2016) applied the 2 last options and concatenated the resulting feature representations.

Given the promising results of element-wise multiplication, bilinear models that take the outer product of 2 vectors have drawn the attention of many researchers. As opposed to element-wise multiplication, each element of the text vector interacts with each element of the image vector via a multiplication. It however generates a bimodal representation of very high dimensions resulting in huge models with lots of trainable parameters. It thus poses a problem of scalability for high-dimensional representation vectors.

III. METHODS

The general architecture of the model is depicted in Figure 1. Each modal representation passes through a fully connected layer and a dropout layer for regularization.

After confirming the significant advantage of combining the features in a multiplicative manner, we implemented the fusion as a plain outer product between the image and question dense vectors. Given that the problem is cast as a binary classification problem, we did not suffer of the dimensional problems evoked above at this stage of the project.

Finally, the multi-modal representation passes through a last fully connected layer. The loss is a cross-entropy loss.

In terms of training, we make use of Adam (Kingma and Ba 2014) as optimizer with a weight decay for further regularization. We rely on the cosine-annealing warm restarts scheduler described in the work of Loshchilov and Hutter (2016) and for the last epochs, we switch to stochastic weight averaging (Izmailov et al. 2018) in order for the trained weights to better generalize.

IV. RESULTS

The learning rate, training and validation loss curves are reported in Figure 2. Firstly, we notice the effect of the first scheduler that loops over the following procedure: it progressively decreases the learning rate to converge to a minimum and suddenly jumps back to its higher initial value to escape from suboptimal local minima. Secondly, in the last epochs, we observe the effect of the second scheduler with an annealing phase where the learning rate increases and then stabilise to a plateau. At the last epoch, the learned weights are averaged over the different values computed during the plateau.

We report an accuracy over the test set of 66.11%.

V. DISCUSSION

Given the few training and validation samples, it seems our model suffer from some issues in capturing the semantics of an image and text and reliably joining them together. Attention mechanisms could be an option to help the network learn where to focus its efforts.

An end-to-end training of the network either by training our own image and text representations from scratch or by fine-tuning a backbone model could bring some further improvements.

VI. FUTURE WORK

To cope with the dimensionality issues evoked in bilinear models, Fukui et al. (2016) proposed multi-modal bilinear pooling to reduce the dimensions of the problem via a projection called the Count Sketch Projection. Kim et al. (2016) proposed to constraint the rank of the weight matrix to reduce the number of parameters in the bilinear pooling. And MUTAN (Ben-younes et al. 2017) generalized the 2 former approaches and operates with a Tucker decomposition of the weight matrix. In the second phase of the project, we will investigate these approaches and experiment over the use of attention mechanism to better focus the fusion on relevant parts of the images and questions.

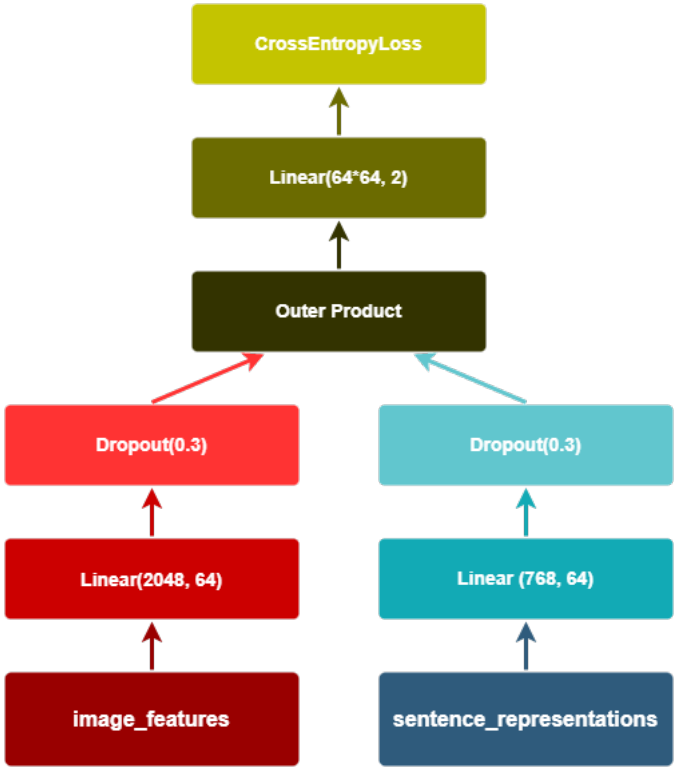
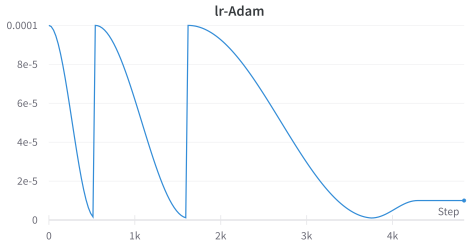
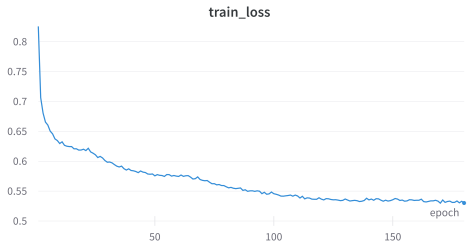


Figure 1: Model Architecture



((a)) Learning Rate



((b)) Train Loss

((c)) Validation Loss

Figure 2: Learning Rate, Training and Validation Loss Curves

REFERENCES

Ben-younes, Hedi et al. (2017). "MUTAN: Multimodal Tucker Fusion for Visual Question Answering". In: *CoRR* abs/1705.06676. arXiv: 1705.06676. URL: http://arxiv.org/abs/1705.06676.

Fukui, Akira et al. (2016). "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding". In: *CoRR* abs/1606.01847. arXiv: 1606.01847. URL: http://arxiv.org/abs/1606.01847.

Izmailov, Pavel et al. (2018). "Averaging Weights Leads to Wider Optima and Better Generalization". In: *CoRR* abs/1803.05407. arXiv: 1803.05407. URL: http://arxiv.org/abs/1803.05407.

Kim, Jin-Hwa et al. (2016). "Hadamard Product for Low-rank Bilinear Pooling". In: *CoRR* abs/1610.04325. arXiv: 1610.04325. URL: http://arxiv.org/abs/1610.04325.

Kingma, Diederik P. and Jimmy Ba (2014). *Adam: A Method for Stochastic Optimization*. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. URL: http://arxiv.org/abs/1412.6980.

Lin, Tsung-Yi et al. (2014). "Microsoft COCO: Common Objects in Context". In: *CoRR* abs/1405.0312. arXiv: 1405.0312. URL: http://arxiv.org/abs/1405.0312.

Loshchilov, Ilya and Frank Hutter (2016). "SGDR: Stochastic Gradient Descent with Restarts". In: *CoRR* abs/1608.03983. arXiv: 1608.03983. URL: http://arxiv.org/abs/1608.03983.

Malinowski, Mateusz, Marcus Rohrbach, and Mario Fritz (2016). "Ask Your Neurons: A Deep Learning Approach to Visual Question Answering". In: *CoRR* abs/1605.02697. arXiv: 1605.02697. URL: http://arxiv.org/abs/1605.02697.

Saito, Kuniaki et al. (2016). "DualNet: Domain-Invariant Network for Visual Question Answering". In: *CoRR* abs/1606.06108. arXiv: 1606.06108. URL: http://arxiv.org/abs/1606.06108.