Baré Alexandre (r0912072)

Visual Question Answering - Part II

H02C8b: Information Retrieval and Search Engines

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

I. INTRODUCTION

In this work, I tackle the task of visual question answering for multiple choices with a subset of the COCO dataset (Lin et al. 2014) on sport activities.

II. RELATED WORK

ResNet (Residual Network) was developed by Microsoft research unit (He et al. 2015). Despite the fact that training error increases with the depth of a neural network, also known as the degradation problem. They made use of the concept of skip connection which significantly smooths the loss surface. A skip connection allows the information to shortcut a block of layers via a forked path where the second branch is an identity function. Both branches join back together thanks to a summation operation. This approach is extensively used in ResNets and allows to surpass VGGNet on ImageNet with a 8 times deeper stacking of layers while lowering the complexity.

BERT (Bidirectional Encoder Representations from Transformers) is a neural network architecture developed by Google AI Language unit (Devlin et al. 2018) for language representation. As its name indicates, BERT relies on bidirectional training of transformers such that it learns the context of words from their surroundings (both from the left and the right). To generalize to many NLP problems, BERT is pretrained on 2 main tasks: prediction of random masked words in a sentence and binarized next sentence prediction.

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.

III. METHODS

Image representations are computed from a pretrained ResNet50. Text representations for both the question and answers are extracted from a pretrained BERT.

Similarly to what we did in the first part of the assignment, we will rely on a bilinear model. As a reminder, bilinear models take the outer product of 2 vectors. 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. In our case, we significantly limited the dimensions of the text and image representation to a size of 64 by passing each through a dense layer.

To choose among multiple answers, being furthermore different from one question to another, a classification task with the cross entropy loss as used in part I is not appropriate. We leverage instead the triplet loss. It has been recently exploited specifically for image and text classification using neural networks in the work of Hoffer and Ailon (2014). Triplet loss allows for relative comparison between an anchor sample A and two other samples, a positive sample P that matches the anchor and a negative sample N that does not. $f(.)$ refers to the neural network used to compute the representation of the samples.

$$\mathcal{L}(A,P,N) = \max\left(\|\mathrm{f}(A) - \mathrm{f}(P)\|^2 - \|\mathrm{f}(A) - \mathrm{f}(N)\|^2 + \alpha, 0\right)$$

The inherent idea is to pose the task as a similarity problem. The purpose of the loss is to minimize the distance between similar samples and maximize the distance between dissimilar ones such that the difference between both distances is ideally not greater than the margin α.

In our visual question answering setting, we have a variable number of possible answers per question from which only one is correct. At each step, we therefore set the anchor to be the combined (image, question) representation, extract the only positive answer and sample 8 negative answers at random. This approach allows for a robust training where the network is shown different negative samples per question at each iteration. To be more precise, the anchor A is a tuple (image, question), the positive P is a tuple (image, correct answer) and the negative N is a tuple (image, wrong answer). The purpose here is that the representations of visual and text data in each tuple are fused with an outer product. We argue that repeating this approach for each tuple should help the network to better learn the interaction between image and text data.

The general architecture of the model is depicted on the next page.

In terms of training, we make use of Adam (Kingma and Ba 2014) as optimizer with a weight decay for regularization. For the same objective, we also make use of dropout layers. We rely on the cosine-annealing warm restarts scheduler described in the work of Loshchilov and Hutter (2016).

IV. RESULTS

Given that we are given the image features and question representations, we can handle a higher batch size of 512. The learning rate, training and validation loss and accuracy curves are reported in Figure 2, 3, 4. We see that the loss curves are relatively smooth but saturates quickly.

We report an accuracy over the test set of 34.23% and the right answer is on average ranked 2.82in the list of 18 answers when ordered by euclidean distance to the anchor.

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.

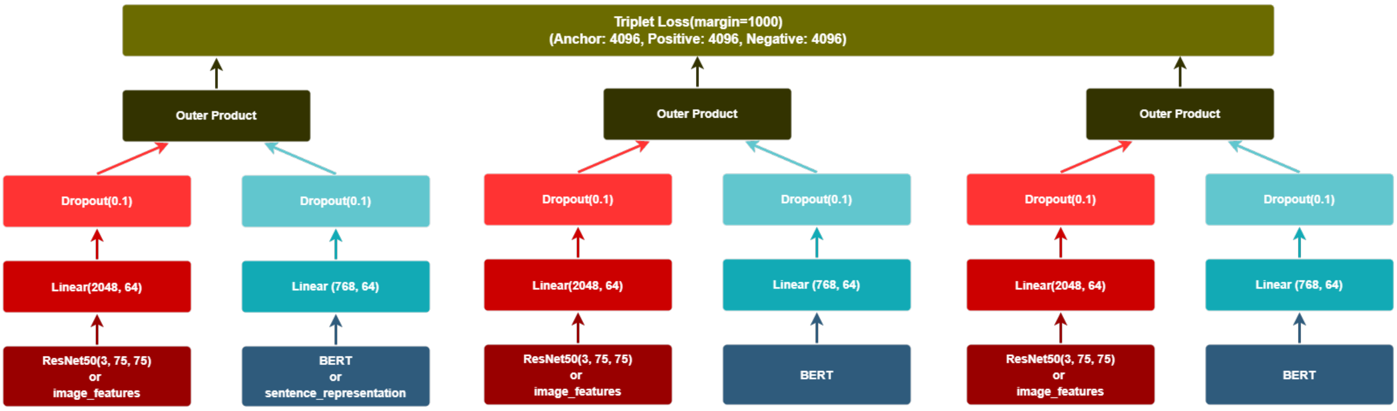


Figure 1: Model Architecture

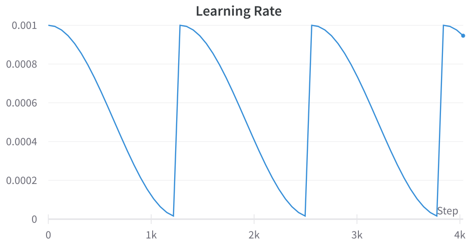
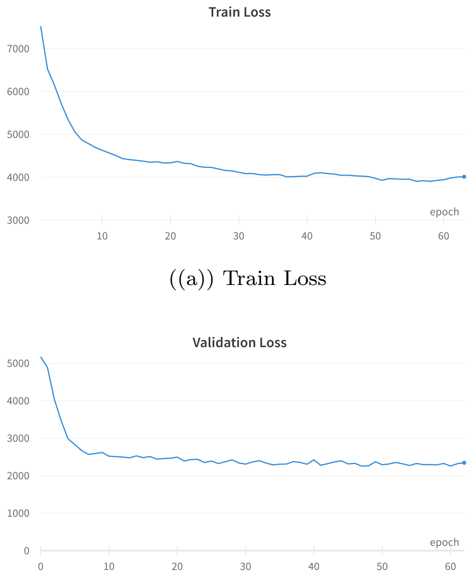


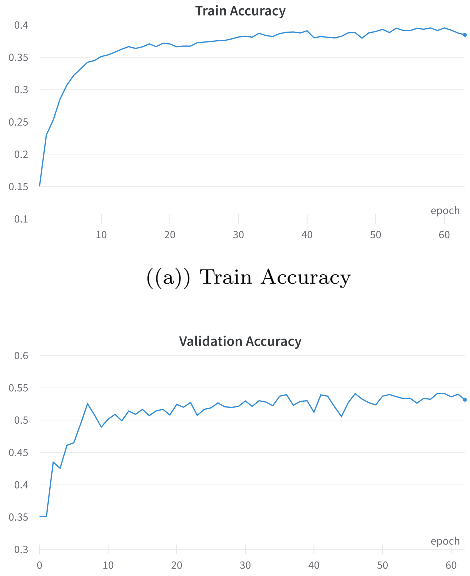
Figure 2: Learning Rate



((a)) Train Loss

((b)) Validation Loss

Figure 3: Learning Rate, Training and Validation Loss and Accuracy Curves



((a)) Train Accuracy

((b)) Validation Accuracy

Figure 4: Training and Validation Accuracy

REFERENCES

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: `1810.04805`. URL: `http://arxiv.org/abs/1810.04805`.

He, Kaiming et al. (2015). "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385. arXiv: `1512.03385`. URL: `http://arxiv.org/abs/1512.03385`.

Hoffer, Elad and Nir Ailon (2014). *Deep metric learning using Triplet network.* DOI: `10.48550/ARXIV.1412.6622`. URL: `https://arxiv.org/abs/1412.6622`.

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`.