The given task of Visual Question Answering aims at answering natural language questions referred to an image. For this reason we decided to adopt an architecture that combines image features extracted by a CNN with the hidden state computed by a RNN as input to a classifier in order to predict the answer.

For what concerns the CNN used, we tried with different networks such as MobileNetV2 and different versions of EfficientNet(B0,B2...). What we find out is that using images with reduced dimensions and networks which extract a smaller number of features leads to worse performance. Hence we used EfficientNetB7 without any image preprocessing.

Our research has focused mostly on the development of the natural language processing network. We started using LSTM/GRU layers to build the RNN but considering the nature of the task assigned, it seemed necessary to implement an attention mechanism to highlight relevant features in the input questions. To do that we used Multi-Head Attention blocks [1] (Figure 1) which runs the input through an attention mechanism several times in parallel achieving the hoped results.
At the beginning we used embeddings learned from the data without any preprocessing of the text, but then we started to take into consideration the idea of finding a proper pre-trained word embedding for our model. We decided to use a popular embedding technique called Glove to have a lower dimensional vector space better encoding semantic relations. We examined different versions of Glove and the one which brought the best result has been Glove42B.300d.
A further step was to deal with missing tokens, so tokens of our dataset not present in the Glove embedding space. The majority of these missing tokens were due to the presence of a lot of saxon genitives in questions and we also found out a couple of grammatical mistakes. Hence we filtered them with a preprocessing function otherwise we would have had out of dictionary word vectors.
Then we went even deeper in the preprocessing of our text trying to manage another intrinsic issue of this task which is the word bias, a phenomenon that affects popular embedding methods such as word2vec and Glove which leads to undesirable word associations (for instance gender bias). This problem cannot be fixed but can be at least mitigated by trying to remove the tokens which potentially lead to misunderstandings such as words that occur very frequently in the language. So we applied a "blacklist" of tokens to be removed from our set of tokens which are essentially words coming from the following categories: punctuation, adposition, particle, conjunction, pronoun. As we hoped, removing a small part of the training corpus led to an improvement in the overall bias.

In order to develop the final model we tried several architectures with different combination of the image and question features (such as dot-product or concatenation), number of dense layers in different places of the network, hyperparameters for both the multi-head attention block and in general the whole network.
Finally for what concerns the training process we adopted a two-phase training thinking that training simultaneously the two parts of the network would have led to a less accurate update of the weights. So firstly we trained the model on our

dataset freezing the pre-trained layers i.e. the embedding layers and the whole CNN. Then the following training was instead performed freezing the embedding layers, the Multi-Head Attention blocks, and only six of the seven blocks of EfficientNet. This two-phase learning led to a significant improvement.

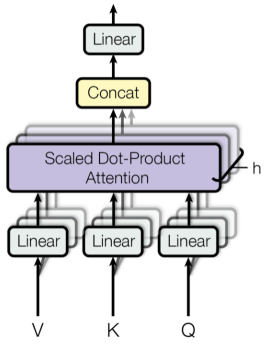

Figure 1: Multi-Head Attention Block

References

[1] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].