POLITECNICO
MILANO 1863

Artificial Neural Networks and Deep Learning
Homework 2 - Image Segmentation

Frantuma Elia - 10567359 - 945729,
Fucci Tiziano - 10524029 - 946638

A.Y. 2020/2021

Table of contents

Chapter 1

Introduction

1.1 Description of the task

The homework consists in solving a visual question answering (VQA) problem on the proposed dataset. The dataset is composed by synthetic scenes (see example below), in which people and objects interact, and by corresponding questions, which are about the content of the images. Given an image and a question, the goal is to provide the correct answer.

(1) **Q**: Is the man's shirt blue? **A**: Yes

1.2 Dataset

The dataset is composed by 29333 total images, 58832 training questions and 6372 test questions.

1.2.1 Images

The images' properties are:

- color space: RGB;
- image size: 400x700 pixels;
- file format: png;

1.2.2 Answers

The set of the possible answers is static and made of 58 possible answers belonging to 3 possible categories: 'yes/no' answers, 'counting' answers (from 0 to 5) and 'other' (e.g., colors, objects, ecc.). In the following the labels associated to each answer:

1.2.3 Data augmentation

We have not performed data augmentation. The dataset dimension was big enough for the complexity of our model and performing augmentation would have required a lot of time to adapt all the question/answer couples. Furthermore, we could do it just on the images and not on the answers set.

1.3 Validation set

No automatic validation set is provided. This means that a subset of the training set must be used to perform validation.

In our case, we parametrized the number of training images to be moved into the validation set, with a 10% probability.

1.4 Test set

The test set is provided as a set of 6372 (image-question) couples, without the attached answers. Participants are required to provide the answers for the test images by submitting the solution with the correct submission format.

1.5 Evaluation

Submissions are evaluated on Multiclass Accuracy, which is simply the average number of observations with the correct label.

Chapter 2

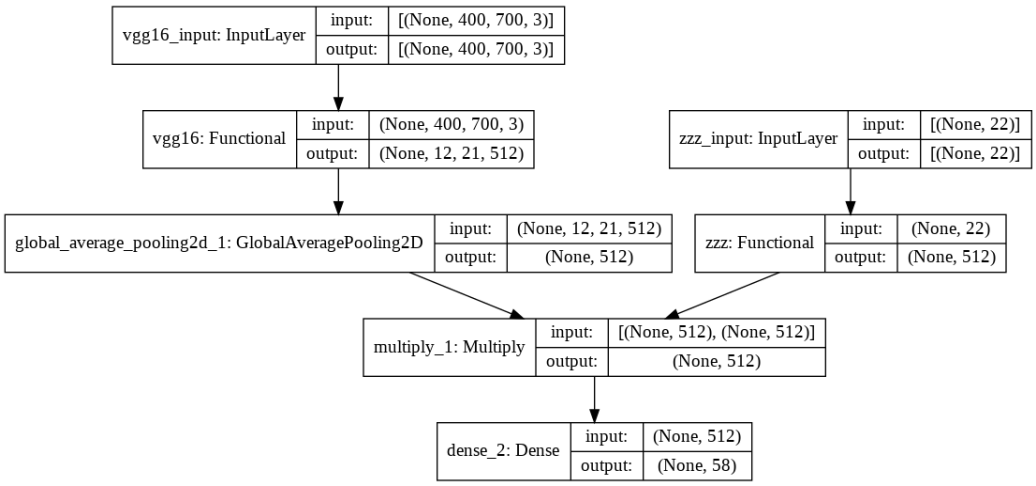
Neural network architecture

2.1 VGG + LSTM + FFNN

At first, we have thought about a network that analyzes in parallel the image and the question, using

- a convolutional network to extract the features from the images;
- an LSTM to process the question and extract the hidden state.

The two results are merged using a multiplication layer and then they are fed to the feed-forward neural network. A GAP layer is used after the CNN to adapt the output shape to the one of the LSTM.



(1) network architecture

2.1.1 Score

The best score of the network was 0.66211, obtained with the following settings:

- VGG standard preprocessing;
- `freeze_until` = 17;
- learning rate = 5e-4;
- 10% validation.

Chapter 3

Model tuning

Our first network had a bidirectional LSTM and one more hidden layer in the FFNN. We were fine-tuning VGG deeply, but we noticed that the model overfitted rapidly. We thought that adding a dropout layer in the FFNN could be a good idea to avoid that behaviour, but the model kept overfitting after few epochs. So we decided to reduce the model complexity, by removing the hidden layer in the FFNN, moving to a uni-directional LSTM and freezing more layers of VGG.

Then, we tried to replace VGG with InceptionResNetV2, as it provided better performance in the previous challenges, but in this case, it turned out to be slightly worst than VGG.

Since we obtained pretty good results straight from the first tries, we decided not to drastically change the architecture of the network.

Chapter 4

References

4.1 Links

- GitHub repository of the project: `https://github.com/tizianofucci/A2NDLVisualQuestionAnswering`
- Competition web page: `https://www.kaggle.com/c/anndl-2020-vqa`