Image captioning with region attention and fine-tuning with Reinforcement Learning

January 21st, 2020

Ariane Alix
ariane.alix@ens-paris-saclay.fr

Sacha Bozou
sacha.bozou@ens-paris-saclay.fr

École Normale Supérieure Paris-Saclay, Departement of Mathematics

Abstract

In the computer vision field, deep neural networks were first mostly used for classification tasks. Now they are rapidly expanding to more challenging applications, in particular in computer vision. We will describe a system focusing on image caption generation: it seeks to automatically describe pictures in human terms. In this paper we report on experimental results with some approaches of these systems, applying it to the Flickr30k dataset.

1. Introduction

We focus on a specific type of image caption generation system: those with attention. These systems are typically based on 3 main parts: a convolutional neural network, an attention mechanism, and a recurrent neural network. The structure is summarized in Figure 1. of the Appendix.

1.1. Attention models

Visual attention mechanisms aim to select the most relevant elements in a picture given a context, and are called that way by analogy to the biological phenomenon of focusing attention on a fraction of the scene to compute adequate responses [3]. Concretely, the method weights the spatial features extracted by the CNN according to their perceived importance for the generation of the next word [4].

1.2. Recurrent Neural Networks

RNNs are typically used to predict sequences with patterns. They are similar to classical neural networks, except that the output obtained at each step is used as input at the next step, along with other features, to provide context.

Since RNNs are computationally expensive, we prefer to limit their memory to a few elements. Thus visual attention mechanisms help by selecting only the most important elements that the RNN needs to know [2] at each step.

2. Review of Anderson's and Rennie's approaches

The difference between 'top-down' and 'bottom-up' attention is that bottom-up signals are associated with unexpected and salient stimuli whereas top-down signals are determined by the current task : they 'look' for something.

2.1. Anderson's bottom-up model

One of the most interesting thing brought by Anderson in his image captioning model is the definition and extraction of the features. This is also the main difference with previous approaches like Xu's (see [7]) which was the state of the art at the moment of its publication in 2015.

The bottom-up approach of Anderson replaces the traditionnal CNN-based features extraction. Using a Faster-RCNN model, it identifies instances of objects belonging to certain classes and localize them with bounding boxes [1]. In conjunction with the Faster R-CNN, a ResNet-101 is used to extract the features from the chosen spatial regions. Figure 1. illustrates the difference in choice of spatial regions: we applied Anderson's Faster R-CNN to a picture from the Flickr30k dataset.

2.2. Anderson's top-down model

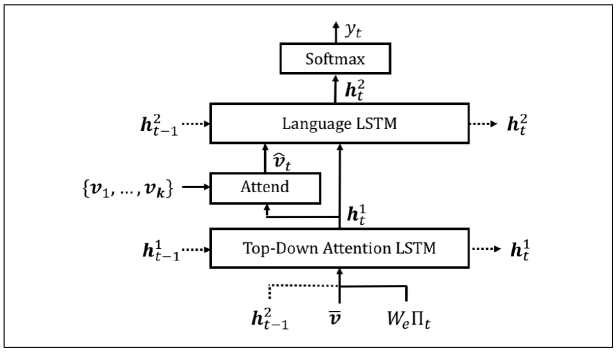

Figure 2. Attention LSTM and Language LSTM

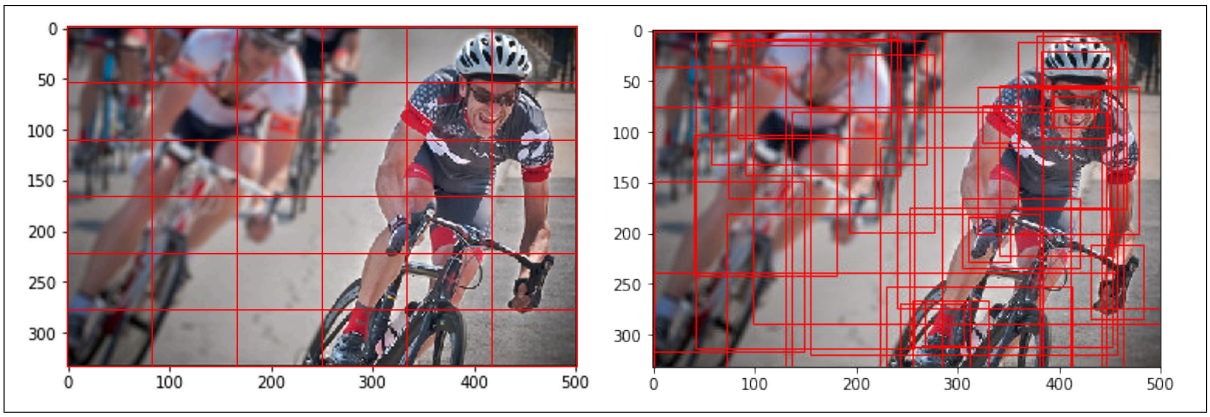

Figure 1. Left: 36 spatial regions used by Xu's method; Right: 36 spatial regions detected and used by Anderson's method

The top-down part in Anderson's method is actually the visual attention mechanism. As said before, 'top-down' means that the model looks for specific parts according to the situation: here it means focusing on the important previously detected regions, given a context.

It is an LSTM (Long Short Term Memory recurrent network), used in conjunction with a language LSTM, to iteratively select the zones of the picture to focus on and predict the next word of the sequence. At each timestep t, the attention LSTM receives as input the previous output h^2_{t-1} of the language LSTM, along with the mean-pooled image features $\overline{v} = \frac{1}{N}\sum_{i=1}^{N} v_i$ and an encoding x^1_t of the previous word.

It will then compute a convex combination $\hat{v}_t$ of the features to focus on, and pass it as an input to the language LSTM along with the hidden state h^1_t. Figure 2. summarizes the two-LSTM system.

2.3. Reinforcement Learning for Fine-tuning

We review here the Rennie's method extracted from [6]. The system is modelled as a Reinforcement Learning problem. We have :

- the agent : LSTM
- the environment : words and image features
- action : prediction of the next word
- reward : any metric on the caption generated (for instance the CIDEr score)
- states : cells and hidden states of the LSTM

The goal is to optimize image captioning systems with reingorcement learning algorithms. To this end, we use Self-critical sequence training (SCST). The principle is to adapt the classical policy gradient REINFORCE algorithm with the reward obtained by the current model under the inference algorithm used at test time as a baseline. Actually, the goal of the approach is to use a baseline that is conditionned under what has been generated by the training so far. This reduces the variance of the rewards. Basically, if we denote L the loss function, we can write its gradient as $\frac{\partial L(\theta)}{\partial s_t} = (r(w^s) - r(\hat{w}))(p_\theta(w_t \mid h_t) - 1_{w^s_t})$ where w^s is a sample from the model and $r(\hat{w})$ denotes the reward obtained by the current model under the inference used at test time. The SCST is described graphically in the Figure 3.

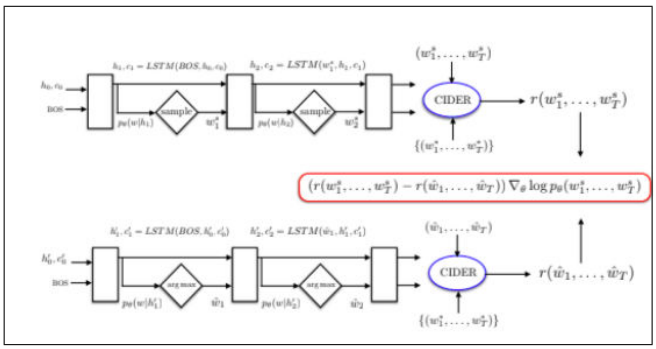

Figure 3. SCST structure

3. Experiments

3.1. Dataset

We originally planned to caption videos with the MVAD dataset. However, due to computational and algorithmic limitations, we restricted ourselves to the Flickr30k dataset of still images (31K pictures with 5 captions each).

3.2. Extraction of the features

Unlike for the COCO dataset, there are no available properly extracted and formatted features for Flickr30k. Therefore, we used an adapted version of Anderson's bottom-up code to extract 36 salient features from each image in the dataset.

The network used for that part is a pre-trained caffe model combining a Faster R-CNN and a Resnet101. Since it is hard and expensive to finetune, we applied it directly to our dataset, which gave good results. An example of the detected features is given in Figure 1., and other examples are available in the Figure 2. of the Appendix.

3.3. Image captioning

Before we generated appropriate features from the Flickr30K dataset, we adapted the implementation from Pooja Hiranandani of Anderson's top-down captioning [5] to make it work with our versions of OS, Pytorch etc.; and we tested it for 2 epochs on the MSCOCO dataset. Since each of these epochs took 8 hours to complete, it motivated our switch to Flickr30k. We show here the results of some interesting experiments.

3.3.1 Metrics used

- Loss: combination of Cross Entropy, Multi-Label Margin Loss and other metrics (see later experiments),
- TOP-5 Accuracy: checks if the target label is one of our top 5 predictions,
- BLEU-1 to BLEU-4: modified precision metrics using n-grams: looks at predicted n-grams found in target captions with no repetition,
- ROUGE-L: looks for longest common subsequence taking into account sentence level structure,
- CIDEr: new automatic consensus metric of image description.

3.3.2 Results of the first experiment

After the 2 epochs on COCO, we ran the algorithm for 45 epochs on the Flickr30k with a loss function equal to $CrossEntropyLoss + 10 \times MultiLabelMarginLoss$, which gave the evolutions of metrics plotted in Figure 4.

The best results on the test set (bigger set than the validation) were obtained at epoch 33, with the values in Table 1. Figure4. of the Appendix shows some example captions.

BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE_L	CIDEr
66.5	49.0	35.5	25.6	47.8	55.0

Table 1. Metrics obtained at epoch 33

We also noticed that the training slows down around epoch 33, and that the TOP-5 accuracy looks like it is overfitting. To understand the source of this behavior, we plotted the evolution from epoch 33 to 47 of the Cross Entropy and Multi-Label Margin losses separately (see Figure 3. in Appendix). There we saw that the Margin Loss decreases 4 times faster than the Cross Entropy. We therefore decided to restart the training from epoch 33 with an increased importance of the Multi-Label Margin in the loss function: $CrossEntropyLoss + 20 \times MultiLabelMarginLoss$.

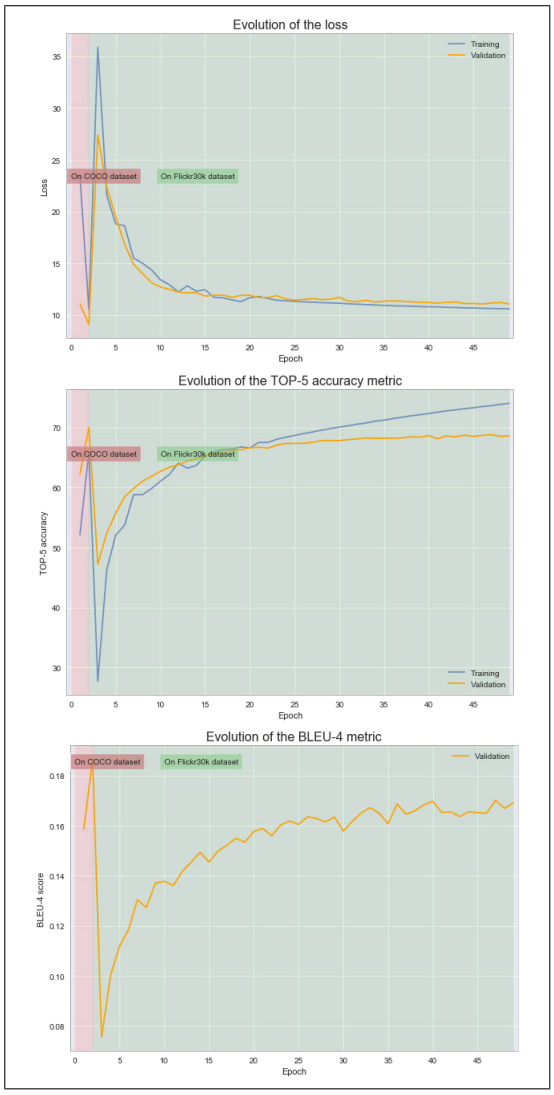

Figure 4. Evolution of metrics for 47 epochs

3.3.3 Other experiments

- Restarting from epoch 33 with loss $CELoss + 20 \times MLMLoss$. Results in Figure 5. of the Appendix.
- Restarting from epoch 33 with $CELoss + 15 \times MLMLoss - 10 \times BLEU4$. Results and captions are in Figure 6. and Figure 7. of the Appendix. Metrics on the test set obtained at the spike of BLEU-4:

BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE_L	CIDEr
67.6	49.6	35.4	25.1	48.2	53.2

Appendix

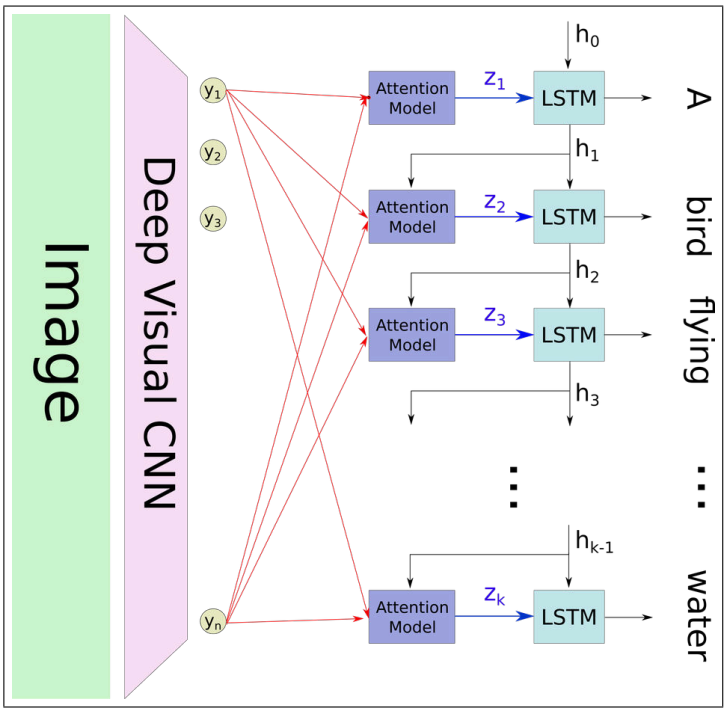

Figure 1. Global structure of image captioning models with attention

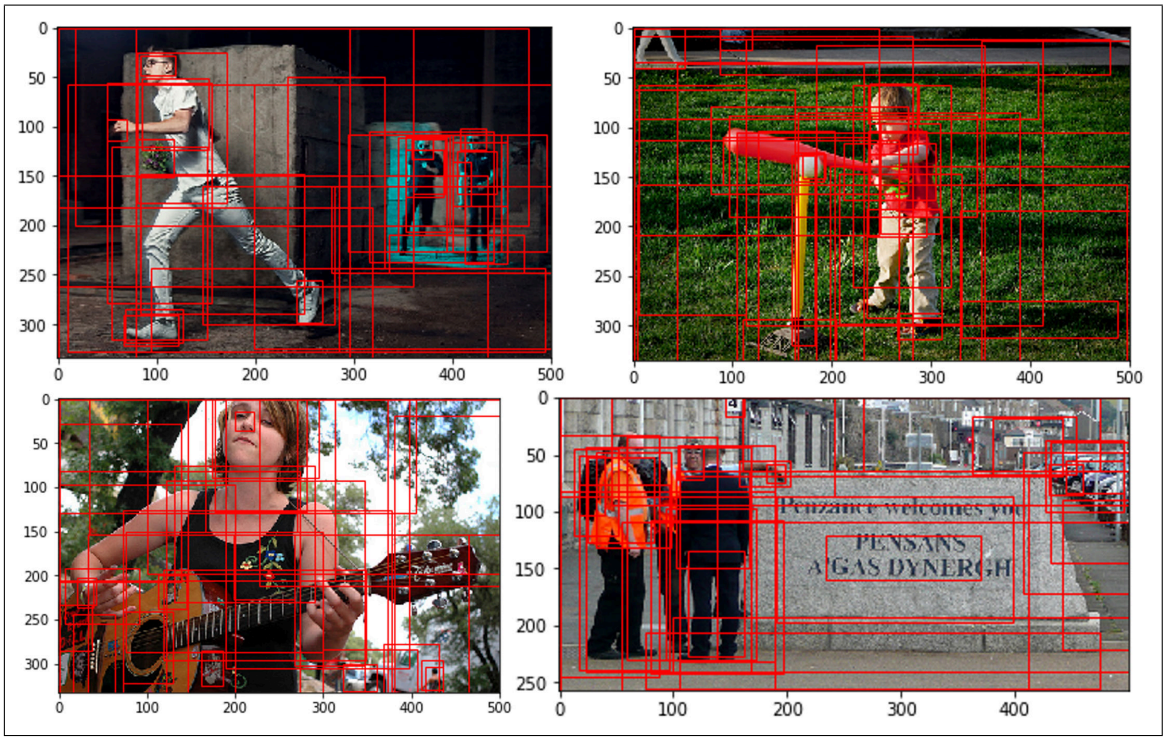

Figure 2. Examples of feature extractions

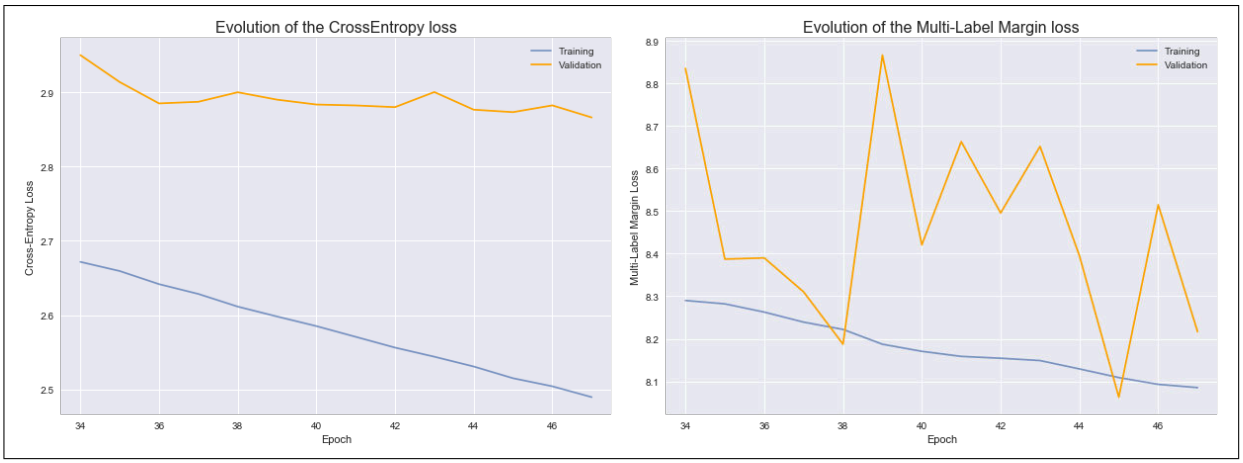

Figure 3. Cross Entropy and Multi-Label Margin losses after epoch 33

Figure 4. Predicted captions for some picture of the test set at epoch 33 and with the loss $CrossEntropyLoss + 10 \times MultiLabelMarginLoss$.

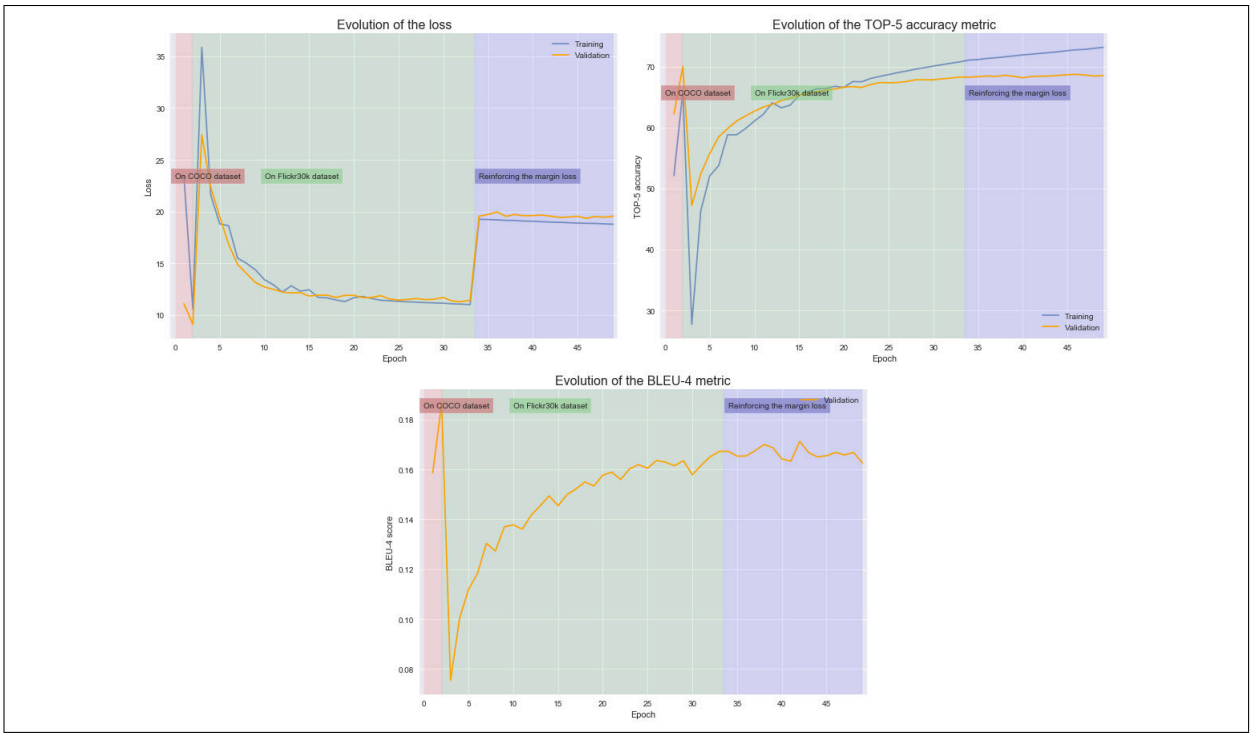

Figure 5. Evolution of metrics when restarting from epoch 33 and reinforcing Margin Loss importance

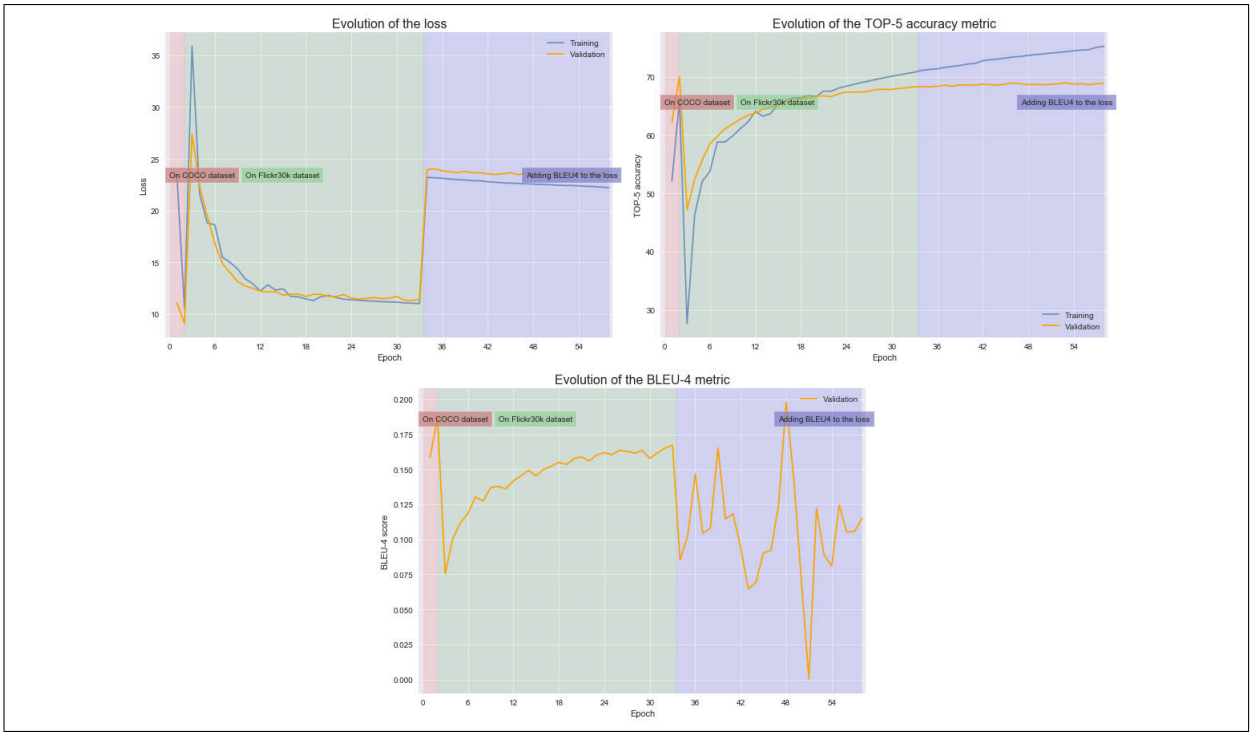

Figure 6. Evolution of metrics when restarting from epoch 33 and adding BLEU-4 to the loss

Figure 7. Predicted captions for some picture of the test set at epoch 33 and with the loss $CrossEntropyLoss + 15 \times MultiLabelMarginLoss - 10 \times BLEU4$.

References

[1] Peter Anderson et al. Bottom-up and top-down attention for image captioning and visual question answering. 2017.

[2] R. Blaine and D. Lawson. Image captioning with attention, 2016. http://cs231n.stanford.edu/reports/2016/pdfs/362_Report.pdf.

[3] L. Blier and C. Ollion. Attention mechanism, 2015. https://lab.heuritech.com/attention-mechanism.

[4] Maurizio Corbetta and Gordon L. Shulman. Control of goaldirected and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215, 2002.

[5] Pooja Hiranandani. Pytorch implementation of bottom-up and top-down attention for image captioning, 2019. https://github.com/poojahira/image-captioning-bottom-up-top-down.

[6] Steven J. Rennie et al. Self-critical sequence training for image captioning. 2017.

[7] Kelvin Xu et al. Show, attend and tell: Neural image caption generation with visual attention. *JMLR W&CP*, 37, 2015.