Fully Convolutional Networks for Semantic Segmentation

Nan Wei
UC San Diego
nwei@ucsd.edu

Renjie Shao
UC San Diego
reshao@ucsd.edu

Hongyi Ling
UC San Diego
holing@ucsd.edu

Abstract

1 Introduction

2 Method

2.1 LSTM

The baseline model use an encoder-decoder architecture. The encoder will take the image as input and encode it into a vector of feature values. The decoder will take this output from encoder as hidden state and starts to predict next words at each step.

For the encoder, we use a pretrained ResNet 50 and it outputs a 2048 dimensional feature map. Then, We use a Fully-connected layer to get 300 dimensional feature vector.
For the decoder, a LSTM is used to generate captions of images. In LSTM, we choose the demension of hidden states to be 512.
When we train LSTM, we use "Teacher Forcing" method. We use the teaching signal from the training dataset at the each time step.
For choice of loss function, we use CrossEntropyLoss. The training process is shown in Fig 3.

2.2 vanilla RNN

Model	Encoder	Embed Dim	Hidden Dim	Layer	Optimizer	Learning rate	L_2 Penalty
LSTM(baseline)	Resnet50	300	512	1	Adam	0.001	10^{-5}
RNN	Resnet50	300	512	2	Adam	0.001	10^{-5}
LSTM+GloVe	Resnet50	300	512	1	Adam	0.001	10^{-5}

Table 1: Hyper-parameters for different models

3 Experiments

In this section, we will show the performance of our networks and discuss the reason why it performs well. There is three subsections. The first is using deterministic approach to generate captions . The second is using stochastic approach to generate captions. The Third is using pre-trained word embeddings.

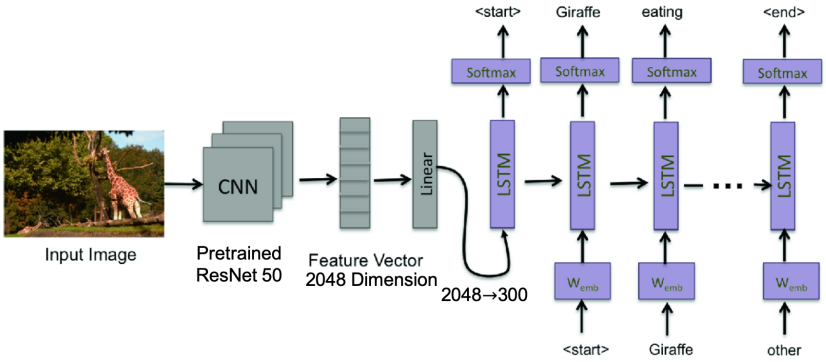

Figure 1: Model Framework [1]

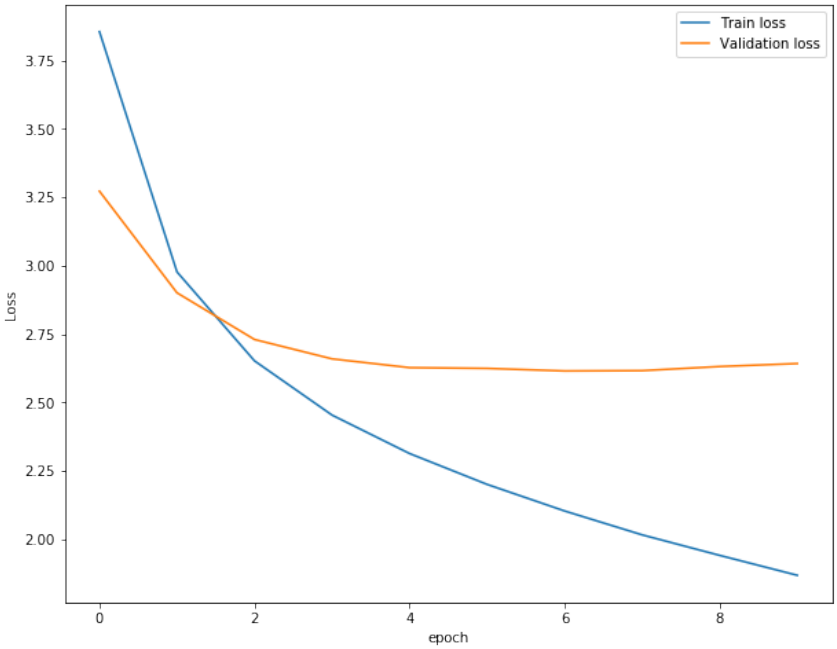

Figure 2: Training and validation loss of vanilla RNN

3.1 deterministic approach

3.2 stochastic approach

In this section, we show the BLEU-1 and BLEU-4 socres of our models.
The result of the baseline model is shown in Table.2. In the table, we can see that when Temperature = 0.1 the model performs the best. When Temperature is very little, stochastic approach is close to deterministic approach. The reason is that when Temperature is very low, the word with highest probability will equal to 1 which is quite same as the deterministic approach.

Temperature	0	0.1	0.2	0.7	1	1.5	2
BLEU-1	87.31	87.29	87.22	87.05	85.89	72.19	55.73
BLEU-4	19.02	17.3	17.13	15.41	13.16	5.66	3.48

Table 2: BLEU-1 and BLEU-4 scores of the baseline model with the stochastic approach

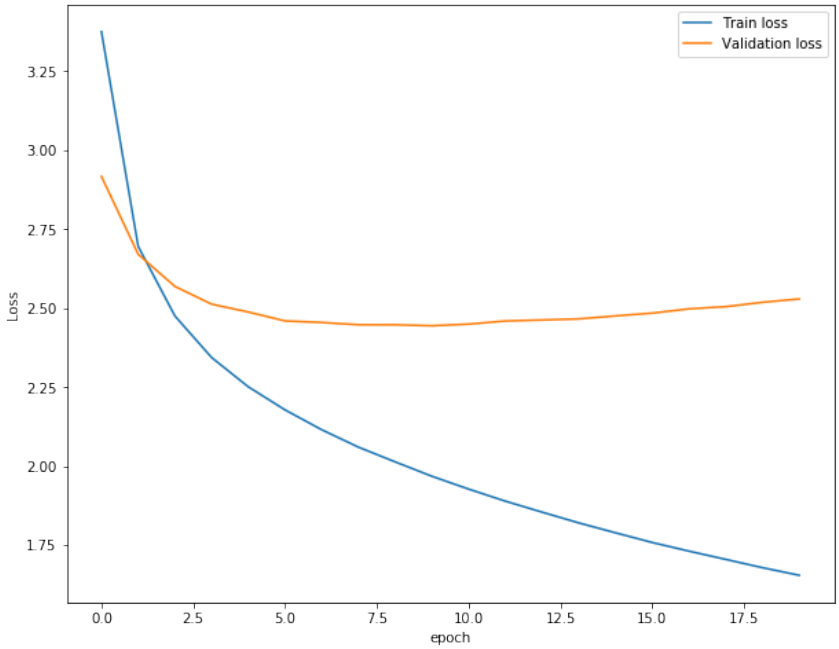

Figure 3: Training and validation loss of LSTM without pretrained embedding

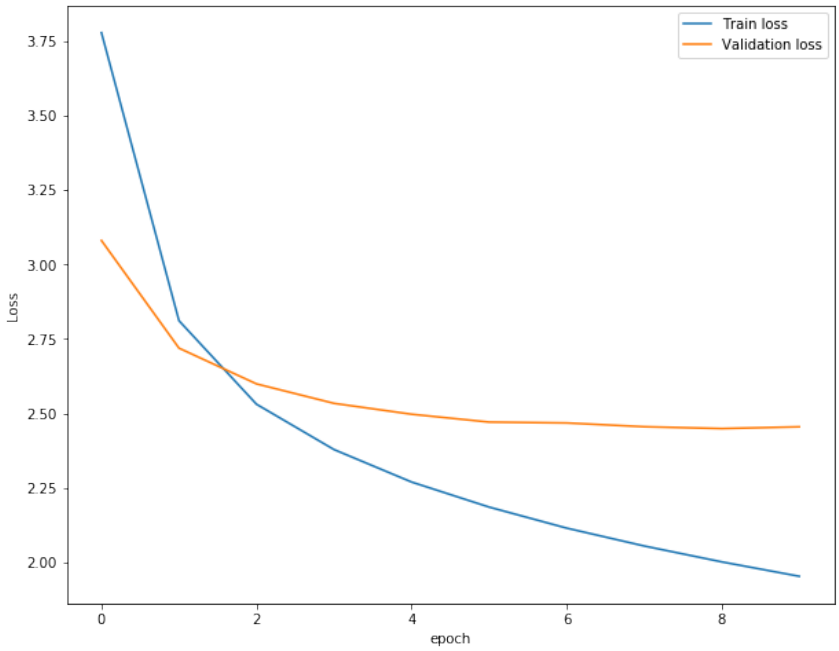

Figure 4: Training and validation loss of LSTM with pretrained embedding

3.3 pre-trained word embeddings

In this section, we show the performance of the network using pretrained word embeddings.

Model	Loss	Perplexity	BLEU-1	BLEU-4
LSTM	2.299	9.966	87.33	17.04
LSTM+GloVe	2.30	9.97	87.31	19.02
RNN	2.46	11.68	87.04	17.22

Table 3: Performance for different models

4 Individual Contribution

Nan Wei

I implement transfer learning with DeepLabv3 and do experiments on basic CNN and DeepLabv3. I also write the report.

Renjie Shao

I implement U-Net architecture and do experiments on U-Net. I also help implement IoU and visualization of segmentation and write some parts of the report.

Hongyi Ling

I implement the generation part of our model and do experiments on baseline model. In addition, I write parts of the report.

References

[1] Anurag Tripathi, Siddharth Srivastava, and Ravi Kothari. Deep neural network based image captioning. In Anirban Mondal, Himanshu Gupta, Jaideep Srivastava, P. Krishna Reddy, and D.V.L.N. Somayajulu, editors, *Big Data Analytics*, pages 335–347, Cham, 2018. Springer International Publishing.