Machine Comprehension for SqUAD dataset

Vikas Bahirwani
Microsoft Corporation
vikasb@stanford.edu

Erika Menezes
Microsoft Corporation
emenezes@stanford.ed

Abstract

We focus on predicting the start and end indices of the answers. Our approach explores the effectiveness of RNNs, BiRNNs, LSTMs, BiLSTMs etc. We explore the use of Attention in addressing this problem as well. We also observe that our model is underfitting and our next steps would have been to develop a more complex model to overcome the same.

1 Introduction

Machine comprehension using Deep Learning models is a growing field in Natural Language Processing. While it possesses immense potential it also presents a lot of challenges - special the challenge to design a neural network to fit the downstream task at hand.

In this assignment, we address the reading comprehension task of generating answers from context paragraphs given the questions. Our approach starts with a simple baseline model using RNNs to capture question and context knowledge.

We then update our model to use LSTMs - which help us learn longer paragraphs and address the answer generation problem better than RNNs. We focus on predicting the start and end indices of the answers.

2 Background/Related Work

There have been many deep learning models proposed for machine comprehension. Wang et al.[1] work is based on the assumption that a span in a passage is more likely to be the correct answer if the context of this span is very similar to the question. The novelty in this paper is the Multi-Perspective Context Matching (MPCM) model that identifies the answer span by matching the context of each point in the passage with the question from multiple perspectives. There is also the work by Xiong et al.[2] that focuses on how to recover from local maxima from incorrect answers. The Dynamic Co-attention Network combines the question and the document in order to focus on relevant parts of both. Then a dynamic pointing decoder iterates over potential answer spans.

3 Approach

3.1 Data Analysis

To better understand the problem, we visualized the data by printing it from qa_answer.generate_answers(). Once the dataset has been read into a list of tuples of (context, question, question_uuid) we used this and the vocabulary passed as an argument to generate the answer for a given a_s and a_e. We are still working on encoding and decoding in order to generate model predictions.

3.2 Approach

Our first approach was to have a simple encoder decoder model as a baseline. This was implemented using an LSTM over the question, and another LSTM for the context by using the initial state as the final state from the question. We then used the final states from the question and paragraph through a 1 layer neural network to predict the start and end of the answer. The following steps capture the approach:

1. 1Question -> LSTM ->Q
2. Paragraph _> LSTM(initial = Q) -> P
3. KRep = [Q,P]
4. as = softmax(KRep * W1) + B1
5. ae = softmax(KRep * W1) + B1

We realise that the biggest drawback of this model is that it does not include attention and in order to fix this we come up with a slightly more complex model that involves the following steps:

1. Question -> LSTM -> Q
2. Paragraph -> LSTM -> P
3. A = softmax(P Q^T) //Compute context vector for Q->P
4. C_P = A Q // and mix with P
5. P = concat(C_P, P) W + b // Mix it with P (Krep)
6. as = softmax(KRep * W1) + B1
7. ae = softmax(KRep * W1) + B1

The last approach that we tried was to introduce non linearity in the decoder by using a ReLU activation function in the neural network .

1. Question -> LSTM -> Q
2. Paragraph -> LSTM -> P
3. A = softmax(P Q^T) //Compute context vector for Q->P
4. C_P = A Q // and mix with P
5. P = concat(C_P, P) W + b // Mix it with P (Krep)
6. as = softmax(KRep * W1) + B1
7. ae = softmax(KRep * W1) + B1

4 Experiments

We experimented (on all approaches) both locally and on GPU. Locally we used 1000 samples (from training) to train the model across 10 epochs (batch size 10). At each epoch, we calculated F1 on 50 validation dataset samples.

In addition, we also evaluated the F1 score of the overall model.

Here are the loss and F1 scores from the final approach.

Epoch 1 out of 10
train loss: 9.7127
Score 6.181432 , best_score so far 6.181432

Epoch 2 out of 10

train loss: 9.3238
Score 4.153968 , best_score so far 6.181432

Epoch 3 out of 10
train loss: 8.8529
Score 6.735965 , best_score so far 6.735965

Epoch 4 out of 10
train loss: 8.5121
Score 4.476740 , best_score so far 6.735965

Epoch 5 out of 10
train loss: 8.3757
Score 2.268926 , best_score so far 6.735965

Epoch 6 out of 10
train loss: 8.2868
Score 4.464495 , best_score so far 6.735965

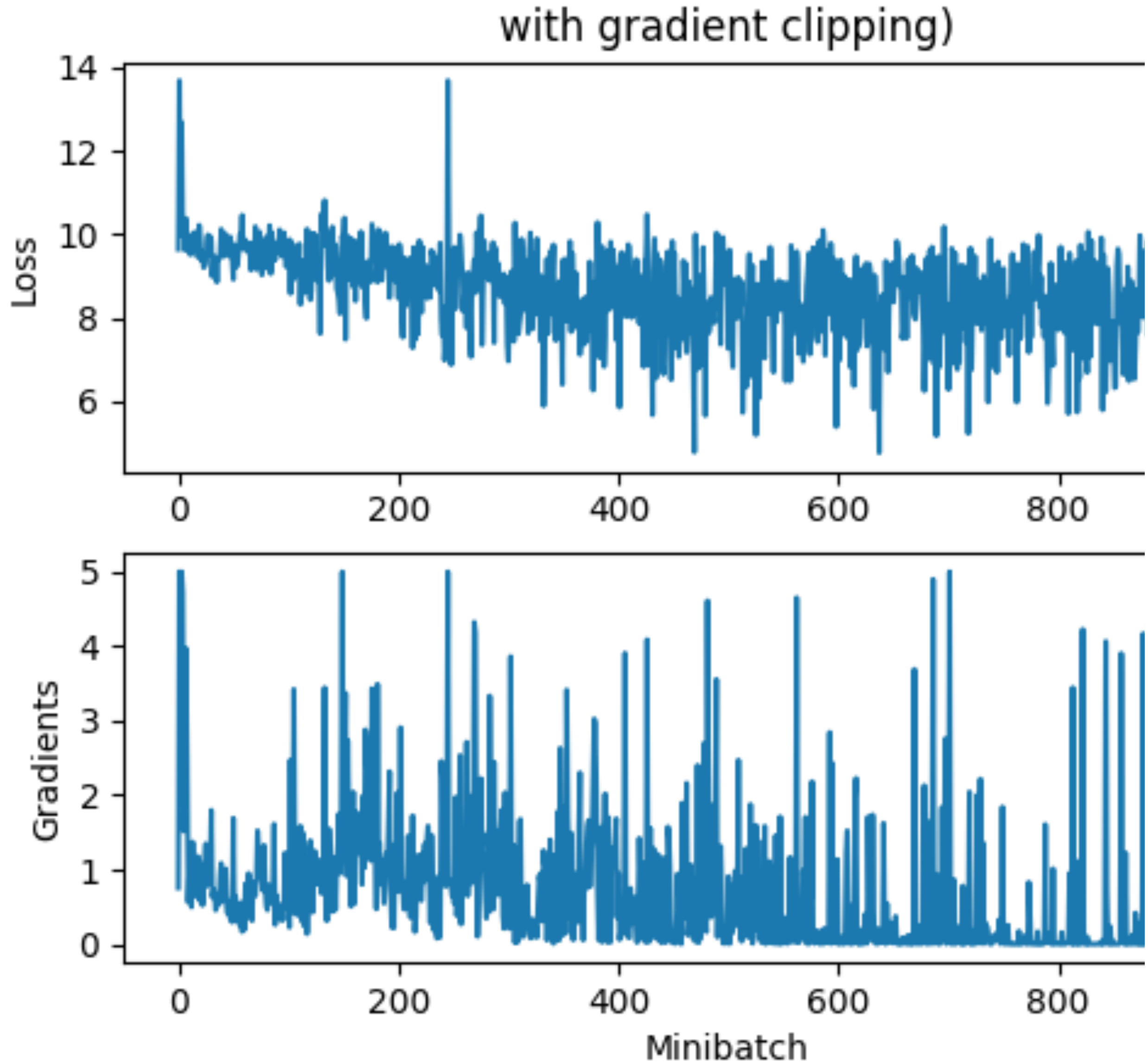
Epoch 7 out of 10
train loss: 8.2392
Score 3.814750 , best_score so far 6.735965

Epoch 8 out of 10
train loss: 8.2109
Score 4.473124 , best_score so far 6.735965

Epoch 9 out of 10
train loss: 8.2412
Score 4.029124 , best_score so far 6.735965

We noticed that we used our models were underfitting because our loss will not go down. (As opposed to overfitting where loss on train is almost 0 or F1 is very high but on validation set the performance degrades).

Furthermore, to confirm that we were not falling into the trap of gradient explosion, we generated the following plot. Essentially it was the same with and without clipping. (GradNorm 5)



5 Conclusion

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle that the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

References

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609-616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System*. New York: TELOS/Springer-Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory

recurrent synapses and cholinergic modulation in rat hiippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.