Indian Institute of Technology Madras

Summer Research Fellowship Program

Experimental Analysis of Attention Models in Deep Learning

Author:
Gokul Karthik
CS18SFP3156

Guide:
Harish G. Ramswamy

18 June, 2018 to 10 August, 2018

Contents

1 Introduction

1.1 Overview

Attention based neural network models have produced more accurate results in recent years though there is no concrete theoretical proof on “Why attention works?”. Hence we experimented the models with and without attention under certain assumptions and tried to understand “How those models learn?”

We generated the proxy data for images with differentiated statistical properties for ’background’ and ’foreground’ regions. The trace of detecting this ’foreground’ region in the data by attention model is analyzed.

1.2 Artificial Neural Network

An artificial neural network is the network of computation models which is loosely inspired by the biological neural network. It generally consists of an input layer, hidden layer(s) and an output layer. The data starts from the input layer and transformed by the operations of an artificial neuron in the hidden layer(s) and the output layer to produce the result.

In a supervised learning approach, the parameters associated with the layers are adjusted in each iteration of passing data such that it reduces the loss between the original and computed values.

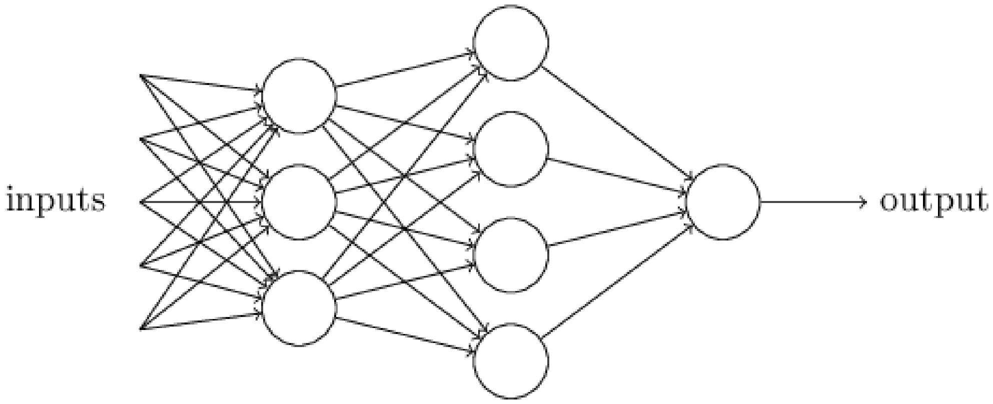
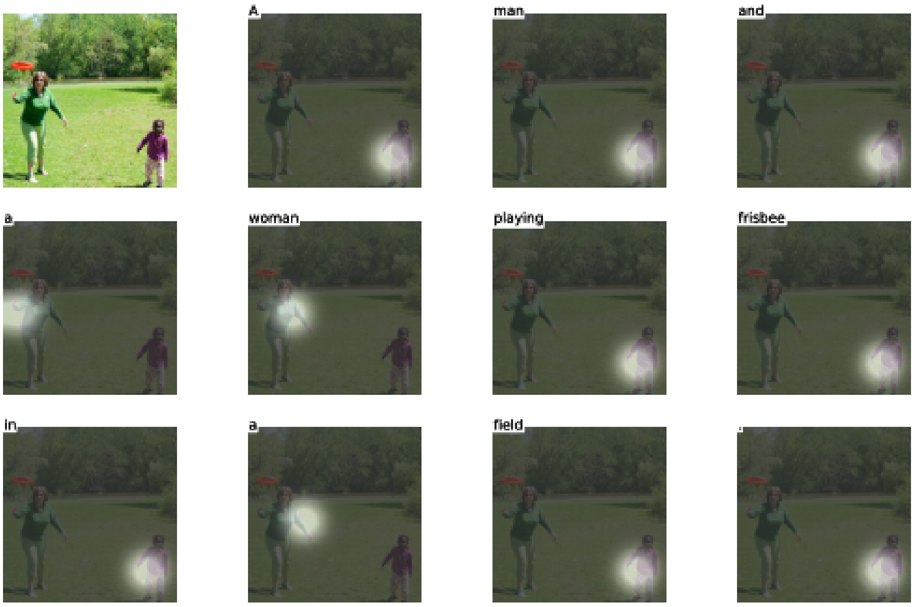


Figure 1: Artificial Neural Network (Source:http://neuralnetworksanddeeplearning.com)

1.3 Attention Models

Attention is the mechanism that was developed to identify the focus area in the data. Before the attention mechanism in machine translation, the text data were transformed into a fixed length vector and machine learning algorithms were applied over that vector. But this truncation into a fixed size vector resulted in the loss of information. Attention mechanisms partially solve this problem by looking over all the information in the area and focusing on the specific area based on the context.

1.4 Attention for Image Captioning



(a) A man and a woman playing frisbee in a field.

Figure 2: Learning of image captioning model with attention (Source: Show, attend and tell: Neural image caption generation with visual attention.)

The mechanism of describing image similar to humans is called as image captioning. In a typical image captioning process, the encoder-decoder model is used.

The image data are transformed into the vector space using convolutional neural networks in the encoder part. The recurrent neural network is then used over the encoded data to generate the caption. Attention mechanism, when added in between the encoder and decoder to change the focus over time, produced better results than the models without attention.

2 Data Generation

2.1 Model generating ”foreground” sub-image and ”background” subimages

Let the image i be segmented into k parts where each part can be represented using a d dimensional vector.

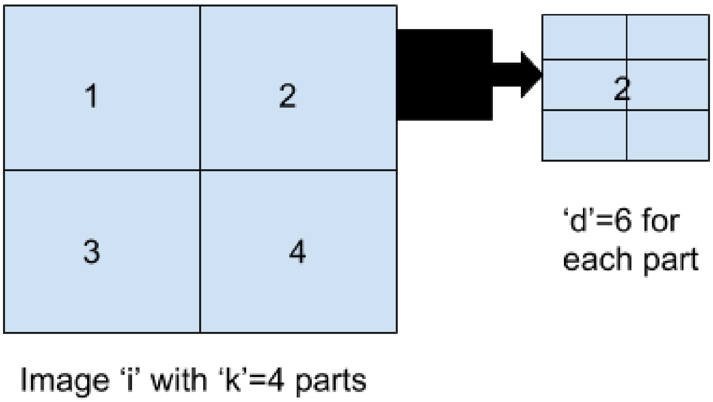


Figure 3: Image representation

We assume that the foreground object to be in only one of these k parts. Let that useful part be P_i and the other non useful parts be P_i. We generated representation data for each image by the following algorithm:

```
Foreground part = $i \in [1, k]$, selected uniformly at random ;
for each part in image $i$ do
    if the part is foreground then
        Generate $data \in R^d$ following the normal distribution with $mean \in R^d$ of 0s
          and $variance \in R^{d \times d}$ of random values;
    else
        Generate $data \in R^d$ following the normal distribution with $mean \in R^d$ of 2s
          and $variance \in R^{d \times d}$ dof random values;
    end
end
```

Algorithm 1: Generating 'foreground' and 'background' data

2.2 Model generating image labels from foreground sub-images

Each data point consists of $k \times d$ values, in which only the part P_i is foreground. Hence we used the $1 \times d$ values corresponding to the part P_i for classification. Let it be $x_{foreground}$.

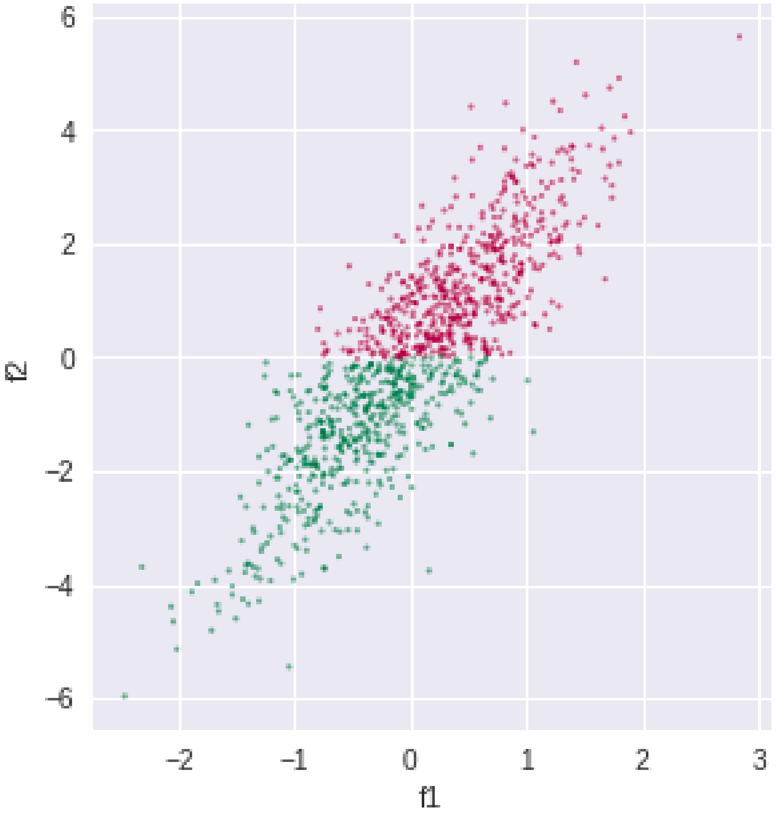


Figure 4: Classification of data for $k = 25$ and $d = 2$, where $f1$ is the the first element of foregroundpart P_i and $f2$ is the second element of foreground part P_i

Classification for each data point is done by the following algorithm.

```
$w \in R^{d\times 1}$ is generated uniformly at random with $low = -1$ and $high = 1$;
$y = x_{foreground} \times w$ ;
if $y < 0$ then
    $class = -ve(0)$;
else
    $class = +ve(1)$;
end
```

Algorithm 2: Classification of data points

3 Deep Learning Models

We built the following 3 types of models to classify the generated data. All those models can be generalized to have 3 hidden layers namely h_1, h_{mid} and h_2. Sigmoid is used as the activation function of hidden layer h_{mid} in attention models. Let x be the data point and w^i is the weight matrix and b_i is the bias corresponding to each layer.

3.1 Simple fully connected model

A simple fully connected model with 3 hidden layers is used to classify the data.

$$output = \sigma(w^4(\phi(w^3(\phi(w^2(\phi(w^1x + b_1)) + b_2)) + b_3)) + b_4)$$

3.2 Attention model with no weight tying for learning attention vector

An attention based model with softmax applied to one of the hidden layers is used to approximate the foreground identification. The softmax output, which maps to probabilities of foreground for each sub part, is multiplied with the input data to get the attentionised data. This data is further passed through a hidden layer for the classification task.

$$output = \sigma(w^4(\phi(w^3(x \times \psi(\phi(w^2(\phi(w^1x + b_1)) + b_2))) + b_3)) + b_4)$$

3.3 Standard attention model with weight tying

An attention based model with softmax applied to one of the hidden layers is used to approximate the foreground identification. The softmax output, which maps to probabilities of foreground for each sub part, is multiplied with the input data to get the attentionised data. Weight tying is additionally performed by adding all the sub parts of attentionised data. This data is further passed through a hidden layer for the classification task.

$$output = \sigma(w^4(\phi(w^3(\Sigma(x \times \psi(\phi(w^2(\phi(w^1x + b_1)) + b_2))), axis = 1) + b_3)) + b_4)$$

4 Results

Stochastic gradient descent is used as the optimizer and binary cross entropy is used as the classification loss measure. The number of epochs is fixed as 5000.

We modelled *attention loss* as the log loss between the $TRUTH\ vector \in \{0, 1\}^k$ denoting the usefulness of each part of a data point in classification and the result of softmax applied to the output of hidden layer h_{mid}.

	Model 1	**Model 2**	**Model 3**
Data 1	0.96	0.96	0.96
Data 2	0.63	0.92	0.92
Data 3	0.54	0.96	0.98

Table 1: Best Test Classification Accuracy

	Model 1	**Model 2**	**Model 3**
Data 1	-	0.88	0.89
Data 2	-	0.57	0.54
Data 3	-	0.58	0.99

Table 2: Test Attention Accuracy corresponding to Table 1 data

	Model 1		**Model 2**		**Model 3**	
	A	**B**	**A**	**B**	**A**	**B**
Data 1	2	4	2	4	3	4
Data 2	3	16	3	8	2	16
Data 3	2	2	2	4	1	16

Table 3: Model Specification corresponding to Table 1 data

Table Legend

- **Data 1:** Data generated with $k = 2$ and $d = 1$
- **Data 2:** Data generated with $k = 9$ and $d = 2$
- **Data 3:** Data generated with $k = 9$ and $d = 32$
- **Model 1:** Simple fully connected model
- **Model 2:** Attention model with no weight tying for learning at-tention vector
- **Model 3:** Standard attention model with weight tying
- **A:** Number of hidden layers
- **B:** Number of nodes in each hidden layer

5 Conclusion

We can see from *Table* 1 that a neural network model without attention is able to perform on par with attention based models on smaller dimensional data. The attention models are performing significantly better than non attention model for higher dimesional data and the difference between their classification accuracies increases with the increase in dimensions.

From *Table* 2, we see that the weight tying operationfor learning attention vector becomes very important for higher dimensional sub parts in an image.We can also see from the *Data* 3 rows of *Table* 1 and *Table* 2, that the significant difference in test attetion accuracy are not transforming to test classification accuracy.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In the Annual Conference on Neural Information Processing Systems (NIPS).

2. Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In the International Conference on Machine Learning (ICML).

3. Bahdanau, D., Cho, K., Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In the International Conference on Learning Theory (ICLR).

4. Ilse, M.,Tomczak, J., Welling, M. (2018). Attention-based Deep Multiple Instance Learning. In the International Conference on Machine Learning (ICML).

Project Link: https://github.com/GokulKarthik/Attention-Experiments-DL