Project Name:	Attention Mechanism
Name:	Alec Mabhiza Chirawu
Chinese Name:	亚历克上
Student Number:	M202161029

Neural Networks and Deep Learning
Information And Engineering Department
USTB

Professor: 黄旗明(Roland)

22 April 2022

Summary

Attention is a strategy in neural networks that mimics cognitive attention. The effect amplifies specific sections of the input data while detracting from others, with the idea that the network should pay greater attention to that small but significant portion of the data. Gradient descent is used to learn which parts of the data are more significant than others, which is dependent on the context.

In the 1990s, attention-like processes such as multiplicative modules, sigma pi units, and hyper networks were introduced.Its adaptability stems from its function as "soft weights" that can fluctuate during runtime, as opposed to "hard weights" that must remain constant. Memory in neural turing machines, reasoning tasks in differentiable neural computers, language processing in transformers, and multi-sensory data processing are all examples of how attention is used.

TABLE OF CONTENTS

1 Introduction

In neural networks, attention is a technique that mimics cognitive attention. The effect enhances some parts of the input data while diminishing other parts. Uses of attention include memory in neural turing machines, reasoning tasks in differentiable neural computers, and multi-sensory data processing (sound, images, video, and text) in perceivers. The attention mechanism is most often used in sequence-to-sequence models, such as the prior probabilities based alignment and it has led to better results than HMM and CTC for text translation. In this work, we explore the use of the attention-based approach for the specific task of handwritten digit string recognition. We opted for an Encoder-Decoder paradigm, as in and proposed a recognition system built upon a CNN (Convolutional Neural Network) and two RNNs (Recurrent Neural Networks).

1.1 Methodology

The project is inspired from the sequence-to-sequence approach of Sutskever. System is able to transform a variable-length sequence of pixel columns, extracted from the handwritten digit string image, into a numerical string. Typical sequence-sequence architecativelyture consists of an encoder and a decoder RNN, our system is built with on three main components. Attention mechanism allows the Decoder to attend to different parts of the source sentence at each step of the output generation. Each Decoder output depends not just on the last decoder state, but on a weighted combination of all the input states.

1.2 Scope of the project

The attention mechanism is the part of the brain that decides where to focus its attention. Feature-based and Location-based mechanisms are not exclusive and can be well used.

2. Findings

2.1 Image Input

We normalized input images in height, width and value for channels. We then completed the images with 0 value to make them 256 pixels wide. We finally resorted to histogram stretching so that each image has pixel values between 0 and 255. All this is done for convenience purposes.

2.2 CNN for Feature Extraction

We used convolutional layers followed by max pooling layers, doubling the number of filters. The feature maps obtained after all these layers are flattened into one features map, then divided into a sequence of column features, ordered from "left to right". As in [6], we used a 2 × 2 pooling size for the first maximum pooling layer.

2.3 Encoder-Decoder

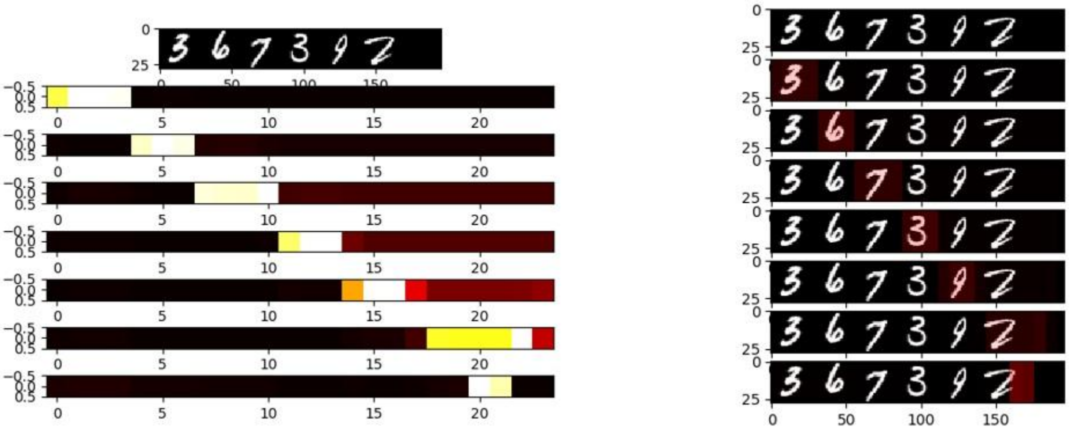
The Encoder and Decoder are models that perform step-by-step decoding of a sequence of data. The key difference between the two is the order of execution, which can only be done layer by layer. This model can handle images of any width as both the CNN and RNN do not have parameters based on the length of the sequence.

2.4 Attetion Mechanism

At each step of attention, the mechanism receives the output sequence of the Encoder, with the attention vector at the preceding time step, and at the output of the Decoder at the previous time step. It produces a context vector using the following equations to work out how much energy is required to hold attention.

$$e_{T,L} = \omega \times \tanh(W \times C_{T-1}^{D} + V \times E_L + U \times A_{t-1} + b)$$

where w, b, W, V are trainable network parameters, C_{T-1}^{D} is the output of the Decoder, E_L is the element in the sequence outputted by the Encoder at position L, A_{t-1} is the attention vector at time $t-1$, and b is a bias.



(a) Value of attention vectors

(b) Vector and image overlay

2.4 ***Experiment***

Mnist database is composed of two datasets A and B and contains handwritten digit strings extracted from bank checks. All images have noisy background, and numbers with 2 to 8 digits. Some images have non numeral writing like lines, dots or dashes. The training set is made up of 2009 images for the training set and 3784 for the testing set.

The Handwritten Digit String is a dataset of 26 different numbers written by 300 people. The train set is composed of 1262 images with 10 different numbers written, and the test set is made up of 6698 images. All images have a white background and the color of ink used can vary. To evaluate the system performance on this dataset we had to:

- Separate the test dataset in two sets: the numbers present in the training set and the rest.
- Shuffle the training and the test set with the same number of images as before, but with all 26 numbers present in the training set.

3. Results

We used a computer having a CPU Intel® Xeon® X5675 @3.07GHz, with GPU Radeon RX 570 Series and Keras 2.7 with Tensorflow 2.7 back end. Categorical cross entropy was used as loss function during training. We created one unique recurrent cell combining the attention mechanism and LSTM cell.

We used a hard and soft metric to evaluate the performance of our system. The hard metric is validating only if all labels in strings are truly recognized. The soft metric is more reliable, to compare the results, than the hard one which is not linear. An improvement of 5% in the soft metric can result in an improvement of 0 to 20%. During training, the attention mechanism revealed to be hard to train from scratch and on harder database. A lot needs to be done to have great results.

4. Conclusion

In this work, we postulate that handwritten digit string recognition can be done using Attention Mechanism which doesn't require any segmentation. A lot need to be done since the results of this experiment are not precise due to time to train the model and testing. But Attention Mechanism have great impact and does work a lot.

5. Reference List

Chen, Y., Liu, L., Phonevilay, V. et al. Image super-resolution reconstruction based on feature map attention mechanism. Appl Intell **51,** 4367–4380 (2021). https://doi.org/10.1007/s10489-020-02116-1

Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. 2019. Fine-tune BERT with Sparse Self-Attention Mechanism. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3548–3553, Hong Kong, China. Association for Computational Linguistics.

Gang Liu, Jiabao Guo,Bidirectional LSTM with attention mechanism and convolutional layer for textclassification,Neurocomputing,Volume 337, 2019,Pages 325-338,ISSN 0925-2312,https://doi.org/10.1016/j.neucom.2019.01.078. (https://www.sciencedirect.com/science/article/pii/S0925231219301067)