Project Proposal

Ibrahim Taher and Forrest Hooton

March 12, 2018

Abstract.

Convolutional neural networks (CNN) have demonstrated the capacity to excel at object recognition. However, it fails at generating sequences of meaningful words, such as captions. We aim to accomplish that by adding a second deep learning algorithm into the model, recurrent neural networks (RNN). We propose that the feature selection aspects of CNN can be used as an input of an RNN model to generate relevant captions.

Introduction.

Computer vision is a field that aims to develop algorithms to create context for digital images or video.[1] Context means identifiers and/or sequences of words. The explosion of deep learning led to a variety of resources to test these algorithms, such as image datasets tagged with annotations. Convolutional neural networks, which are based off the visual cortex in work done by Hubel and Wiesel, are groundbreaking algorithms in annotating images with simple identifiers.[2] A CNN consists of three overarching components: an input layer, hidden layer(s) and an output layer.

Convolutional neural network hidden layers are comprised various convolution, normalization, fully-connected and pooling layers. Convolution is used to compute the similarity between a feature and a *receptive field* via element-wise multiplication.[3] The output for the convolution step is then normalized using the ReLU function.[4] Afterwards, pooling occurs. In practice, pooling is taking the max value of a given window in our normalized convoluted data. The window then slides by a given window stride rate. This effort is done to reduce the dimensionality of the data. Finally, a fully connected layer is a flattened vector of our reduced dimensionality image which allows for each value of the vector to be given a "vote" as to the classification of the image.[5]

While convolutional neural networks excel at classifying images using simple identifiers, they lack the capacity to generate captions. The recurrent neural network is more adept in solving this problem. A recurrent neural network is an algorithm that has the ability to create sequential data given an input. The extra input in RNNs allow for the notion of memory. Recurrent neural networks use loops instead of a feed-forward network.[6,7]

During each iteration the RNN uses input, output, activation and forgetting gates to handle this notion of memory. These gates are developed via linear combinations of the input vector at epoch t and hidden/cell activation vectors at either epoch t or t-1 depending on the gate. These gates are then transformed using a non-linear function such as the logistic or tanh function, depending on the gate. They handle memory by weighting what outputs in previous epochs are still relevant.[8]

Image-caption generation is a tool that has various applications in today's society. Take for instance, those who are visually impaired. If they have the ability to read words at a closer distance, then using this model will allow them to have a better understanding of the context of an image. Also, caption generation is important for all forms of media. Articles that are located in newspapers or on their respective websites are often associated with images depicting events in the article. Image caption generation can help remove the need for human supervision and manual creation of captions.[9]

Proposed Project.

In this project we aim to combine the convolutional neural network and the recurrent neural network to generate captions for images. First, we will reduce the dimensionality of the data using the CNN, thereby extracting relevant features of the image. These reduced features will then be inputs to our RNN, which will learn the associations between relevant features and the captions associated with the images over a series of iterations. In terms of high level algorithms, we are attempting to implement a *Word From Sequence* model; given an image and a part of a sequence, the model will try to predict the next word in the sequence.[10]

The model will be tuned using a grid search, in which it will have to be trained on combinations of several parameters. For the convolutional neural network, the model will be tuned for the number of features used in convolution, size of windows used and the stride of those windows for max-pooling. For the recurrent neural network a we aim to tune is dropout probability (the probability by which certain neurons don't fire.) Learning rate, number of layers and number of neurons will also be tuned for both. Both algorithms employ a variant of gradient descent known as backpropogation. Since backpropogation for large image datasets, over large intervals is computationally expensive, we will be employing a truncated backpropogation algorithm. Essentially, instead of updated the weight vectors, Θ at every interval, Θ will be updated every p epochs.[11]

The data comes from the Microsoft COCO 2014 dataset, which includes images mapped to associated captions. The dataset already comes pre-split with training and validation images that have five associated captions, and it comes with a testing set. The images are in .jpg format and the annotations are given in JSON format, with mappings to its respective image through an id.[12]

We plan to test our model using the BLEU and ROGUE metrics. BLEU is a method to evaluate the precision of a predicted caption. It is as follows:

$$BLEU_{(a,b)} = \frac{\sum_{w_n \in a} min(c_a(w_n), max(c_{b_j}(w_n))}{\sum_{w_n \in a} c_a(w_n)} \; {}^{13}$$

The ROGUE metric is a method to evaluate the recall of a predicted caption. It is as follows:

$$ROGUE_{(a,b)} = \frac{\sum_{j=1}^{|b|} \sum_{w_n \in b} min(c_a(w_n), c_{b_j}(w_n)}{\sum_{j=1}^{|b|} \sum_{w_n \in b} c_{b_j}(w_n)} \; {}^{14}$$

For both, a is a predicted caption, b is a set of ground truth captions, w_n is an n-gram and $c_x(w_n)$ is the count of n-gram w_n in caption c_x.

Using these metrics, we aim to maximize values returned, thus leading to the best possible predicted captions compared to the ground truth labels. By correctly fitting our data, through training and a grid search on our hyperparameters, we believe this model can generate interesting captions for unseen images.

References.

[1] Huang, T. (1996-11-19). Vandoni, Carlo, E, ed. Computer Vision : Evolution And Promise. 19th CERN School of Computing. Geneva: CERN. pp. 2125. doi:10.5170/CERN-1996-008.21. ISBN 978-9290830955.

[2] Hubel, D. H., and T. N. Wiesel. Receptive fields and functional architecture of monkey striate

cortex. The Journal of Physiology, vol. 195, no. 1, Jan. 1968, pp. 215243., doi:10.1113/jphysiol.1968.sp008455.

[3] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Handwritten digit recognition with a back-propagation network, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 1989

[4] An Intuitive Explanation of Convolutional Neural Networks. (2017, May 29). Retrieved March 11, 2018, from https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

[5] Rohrer, Brandon. How Convolutional Neural Networks Work? Brandon Rohrer, 2017, brohrer.github.io/how_convolutional_neural_networks_work.html

[6] Karpathy, Andrey. ”The Unreasonable Effectiveness of Recurrent Neural Networks” Andrey Karpathy, 2015-11-05, https://karpathy.github.io/2015/05/21/rnn-effectiveness/

[7] Britz, Denny. ”Recurrent Neural Networks Tutorial, Part 1 - Introduction to RNNs” WildML, 2015-17-09, http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

[8] Graves, Alex, et al. Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, doi:10.1109/icassp.2013.6638947.

[9] Kiruthika, N P, et al. ”EXTRACTIVE AND ABSTRACTIVE CAPTION GENERATION MODEL FOR NEWS IMAGES.” International Journal of Innovative Research in Technology & Science(IJIRTS), ijirts.org/volume2issue2/IJIRTSV2I2060.pdf

[10] Brownlee, Jason. ”A Gentle Introduction to Deep Learning Caption Generation Models”, Machine Learning Mastery, 2017-17-11, https://machinelearningmastery.com/deep-learning-caption-generation-models/

[11] ”A Gentle Introduction to Backpropagation Through Time.” Machine Learning Mastery, 19 July 2017, machinelearningmastery.com/gentle-introduction-backpropagation-time/.

[12] COCO - Common Objects in Context, http://cocodataset.org/#download

[13] Papineni et. al., BLEU: A Method for Automatic Evaluation of Machine Translation, 200

[14] Lin et. al., ROUGE: A Package for Automatic Evaluation of Summaries, 200