# Transfer Learning on Image Trained CNNs for Textual Analysis

*Chih Chen, Nathaniel Nguyen*

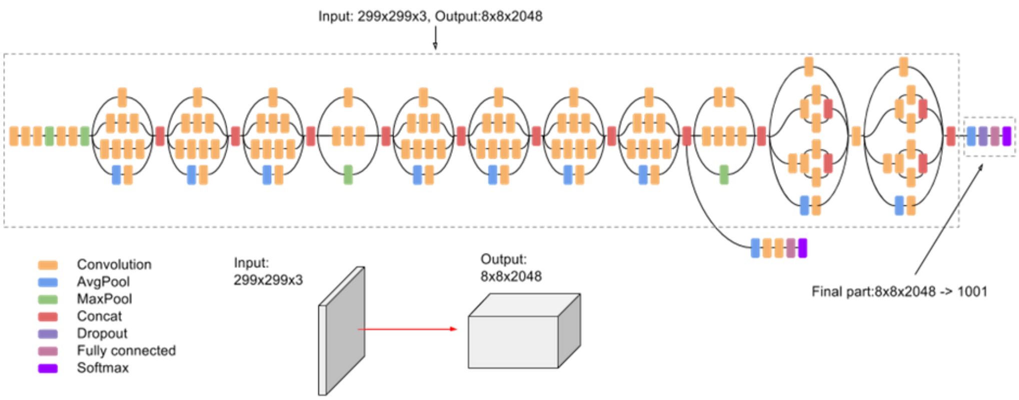
LIGN 167 Final Project

## Description

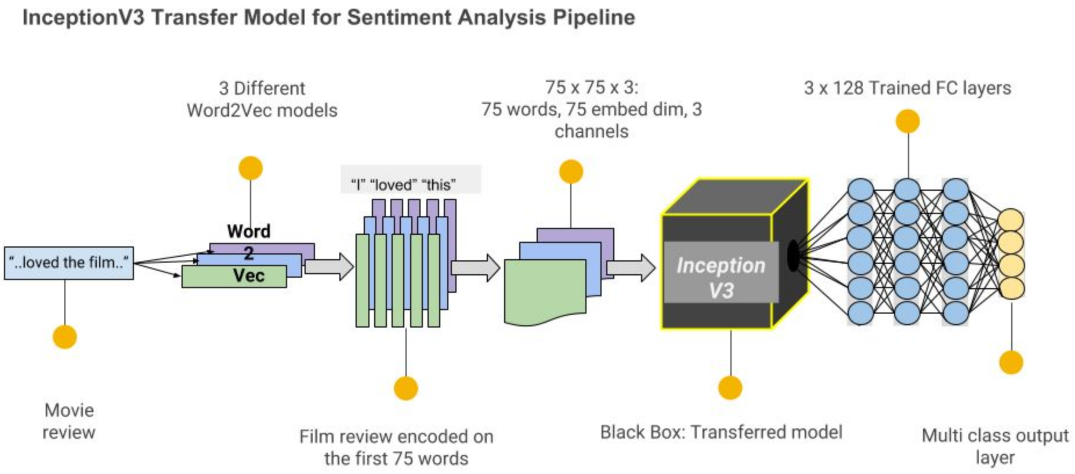
Accessibility to state of the art convolutional network image classifiers like AlexNet and the GoogleNet Inception architecture has spawned various transfer learning approaches using pre-trained models to solve tasks in different domains from which the adapted models were trained for. Stolar et al. [1] successfully adapted the AlexNet network to perform real time Speech Emotion Recognition tasks by providing probabilities of classification outcomes based on a range of 7 emotional states. This was made possible by processing speech patterns into spectrograms as samples for training the network. We saw similar potential in the word embeddings provided by Word2vec that provided a two dimensional framework to analyze sentiment. In this work, we explored the possibility of exploiting the abstract visual landscape provided by Word2vec for use in an existing image classification network to provide sentiment analysis on a body of text.

Our dataset consisted of 50,000 film reviews taken from the IMDB online database with each review ranked on a negative to positive scale of 1-10 with 10 indicating most positive. These reviews varied in range from a few sentences to several paragraphs with a highly diverse vocabulary between entries. There was still much to the data that needed manipulating before it could be used appropriately for image classification, but we believe that the potential word embeddings would provide our model with enough information to perform general binary sentiment analysis (positive or negative review favorability) on a given entry.

## Experiment



The GoogleNet Inception V3 module uses multiple convolution filters in combination with pooling layers to perform multiple feature extractions at once, concatenating the results to make a reinforced estimation of what it perceives the given image to be. Though many of the features which neural networks like GoogleNet were trained on are starkly different from our dataset, part of the aim of our work is to determine how translatable can some of these feature extraction methods work with embedded word vectors. Inception V3 was chosen as the prime candidate for our implementation because of it's notably lower computational costs when compared with some of its higher performing competitors [2] while the methodology in *how* images were classified were subsidiary.



Considering the limitations of our dataset size, computational power, and general novel approach of our work, we decided against a classification model consisting of all ten labels and instead opted for a binary classification model indicating positive or negative, removing neutral (labels 5 and 6) labels. To maintain consistent dimensions with each review, we captured only the first 75 words and built our Word2vec models on 75 embedding dimensions where the resultant 75x75 matrix would be used as our samples. Due to the abstract nature of our visual landscape, we believed using a 'jet colormap' like in the approach conducted by Stolar et al. would not provide salient enough features for extraction by the model. Instead we opted to construct three models based on different window sizes (k = {3, 5, 7}) for determining word relationships.

As a parallel control, we also implemented a two layer convolutional neural network built from scratch to measure the net transferability of the adapted classifier. CNNs have been shown to perform well in natural language processing tasks such as sentiment analysis and question classification[5] making it a dependable control model for comparison sake.

# Results and Analysis

***Table 1: Model validation accuracy***

| **Model** | **Validation accuracy** |
|---|---|
| Transferred model (InceptionV3) | 52% |
| Control model (CNN) | 70% |

Table 1 shows the validation accuracy of our models on 1000 never before seen samples. As noticed from the table, our transferred model achieved a validation accuracy close to that of chance with a score of 52%. This accuracy along with the observed reduction of loss over 10 epochs suggests our model was fitting the noise in the dataset rather than useful informative features.

We believe that the images produced from our Word2vec embeddings were perhaps too abstract for the pre-trained Inception V3 model to be able to extract enough meaningful features from its existing weights. As a result, the output from the Inception V3 layers seemed to just add noise to our input data which may explain why the validation accuracy remained unchanged despite an overall decrease in loss throughout training.

Although, the existing weights from the transferred model may not have been able to extract useful features, this does not suggest that a different approach where the Inception V3 weights may undergo a brief re-training period could not allow for improved results in the problem task of sentiment analysis. Such a claim can be supported by our 'vanilla' CNN architecture which was able to achieve higher accuracy using the same text to image data pipeline.

This built from scratch CNN achieved a 70% validation accuracy score trained on 10 epochs. This convolutional network was far simpler than the Inception V3 model yet it was able to produce relatively substantial results. We believe this suggests that despite our struggles in extracting useful features from existing image models, formatting reviews into an image input format and applying advanced, well experimented visual methods may still prove to be a viable approach in solving text based analytical problems.

## Conclusion

Our results did not provide an agreeable accuracy score and the implemented demo provided less than desirable results on a range of test data. Although our results indicated poor transferability for our Word2vec models to use for Inception V3, utilizing the same Word2vec embedding for a CNN built from scratch yielded significantly higher scores and reliability in demos. This leads us to believe that word embedding images may still prove to be a viable candidate to utilize state of the art image classification networks with limited data and training time.

## Future work

Seeing how abstract images constructed from Word2vec embeddings can potentially be candidates for an image classification task, we would like to explore alternative ways to characterize color channels for word vectors. Altering the window sizes may have affected the abstract 'contours' of our image too greatly and perhaps there are other qualitative features of word embeddings that would more succinctly capture the information provided by color bands. Additionally, we can use publicly available Word2vec vectors trained on a different but larger corpus.

If we can improve the accuracy score on a more modest neural network by tuning our Word2vec models, we can then proceed to explore ways to massage GoogleNet's trained weights to correlate with our data.

## References

[1] Stolar, M. N., Lech, M., Bolia, R. S., Skinner, M., & 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS). (December 01, 2017). Real time speech emotion recognition using RGB image classification and transfer learning. 1-8.

[2] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). *Rethinking the Inception Architecture for Computer Vision*. *Arxiv.org*. Retrieved 12 December 2018, from https://arxiv.org/abs/1512.00567

[3] Mass, A. L., Daly, R. E., Phan, P. T., Huang, D., Ng, A. Y., Potts, C. & 2011. *Learning Word Vectors for Sentiment Analysis.* Association for Computational Linguistics

[4] Karpathy, A. *Transfer Learning.* CS231n Convolutional Neural Networks for Visual Recognition. Retrieved 12 December 2018, from
http://cs231n.github.io/transfer-learning/

[5] Kim, Y. (2014). *Convolutional Neural Networks for Sentence Classification*. *Arxiv.org*. Retrieved 14 December 2018, from https://arxiv.org/abs/1408.5882