# Embeddings are All You Need: Transfer Learning in Convolutional Neural Networks using Word Embeddings

Matt Baughman, Ian Foster (Advisor) & Kyle Chard (Advisor)

## Abstract

With recent advances in the use of efficient neural networks and in relational learning using word embeddings as prediction targets for image classification, the combination of these two concepts offers promise for efficient transfer learning. Given the properties of word embeddings to represent information-dense abstractions of language concepts in arbitrary vector spaces, the projection of an image into that same vector space has been shown to enable operations between images demonstrated in word embeddings. We describe how we extend this idea to show how training a neural network model under this regime can lead to transfer learning within the embeddings' vector space. This allows the model an advantage in predicting classes of images not previously encountered without any additional retraining. Additionally, we demonstrate this principle using a neural network architecture previously shown to be state-of-the-art for model efficiency and so further demonstrate the applications of these methods in light weight machine learning.

## Background

- Word embeddings map semantic features into vector representations which enables relational operations
  - E.g., “king”-”man”+”woman”=”queen”
- Typical image classification predicts a single, “most-likely” class given an image
- Past work has shown that, using transfer learning, you can map input images to the embedding spaces

## Methods

| | | | | |
|---|---|---|---|---|
| **cat** => | 1.2 | -0.1 | 4.3 | 3.2 |
| **mat** => | 0.4 | 2.5 | -0.9 | 0.5 |
| **on** => | 2.1 | 0.3 | 0.1 | 0.4 |

- Convert CIFAR-100 [3] labels to vectors using GloVe embeddings[4] and use these as training targets
- Using leave-one-out cross validation, train the model on 99 of the 100 classes and validate loss on the left-out class
- Analyze the difference in loss from the beginning of training to the end of training to determine if there is reduction in error, indicating the ability of the model to learn a general vector space

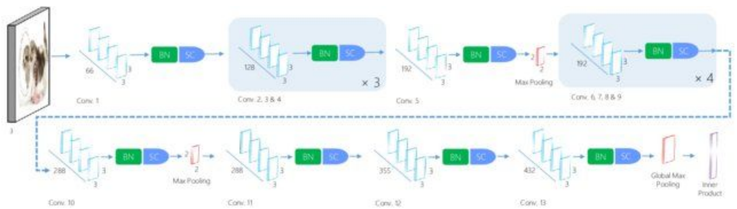


Fig. 1: Base computer vision architecture from [2].

## Evaluation and Results

- Fig. 2 illustrates the median and quartiles (classes #25 and #75 of 100) of epoch-wise loss (MSE: mean squared error)
- The median performer **decreased loss by 31.0%**
- Decreases plateaued after ~15 epochs, indicating a lack of further generalization ability using this training regime
- Loss continued to reduce for specific excluded classes, potentially demonstrating continued generalization for certain specific features present in the embedding space

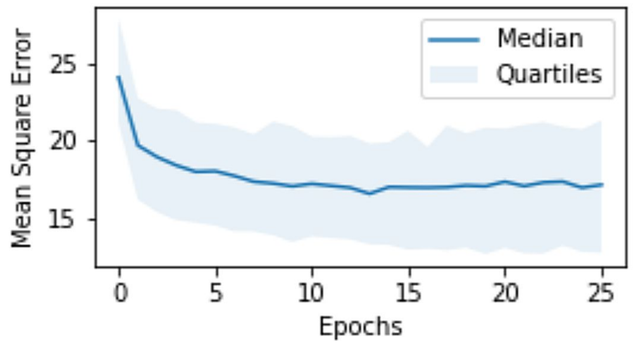


Fig. 2: Loss (MSE) vs. Epoch for the excluded classes.

- Fig. 3 shows the **majority of classes improved** with respect to their starting losses
  - 86 out of 100
- The best performer had its **MSE reduced by 65.5%**
  - The worst performer increased by 67.9%--a clear outlier
- Initial observations show a correlation between performance and presence of representative classes in the dataset
  - E.g., “willow” will do well given three other tree species

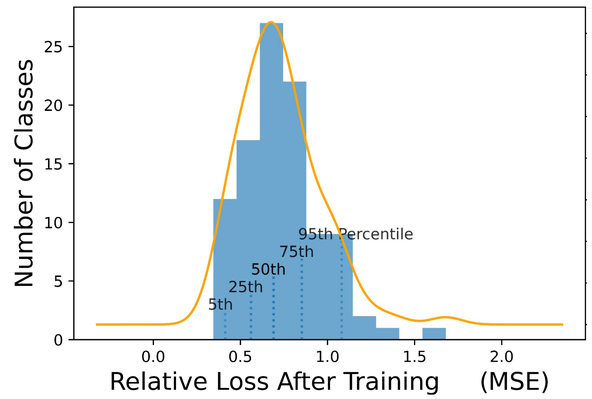


Fig. 3: Histogram of relative validation across excluded classes.

## Conclusion and Takeaways

- Illustrated the feasibility of transfer learning in convolutional neural networks using word embeddings as training targets
- Implemented a design that allows for this hybrid approach without using pretrained models
- Utilized a light weight computer vision model to demonstrate the flexibility of this method
- Demonstrated a ***~31% median decrease*** in validation loss for an unseen class over 25 epochs of training

## References

[1] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. 2013. “Devise: A deep visual-semantic embedding model.” (2013).

[2] Seyyed Hossein Hasanpour, Mohammad Rouhani, Mohsen Fayyaz, Mohammad Sabokrou, and Ehsan Adeli. 2018. “Towards principled design of deep convolutional networks: Introducing simpnet.” arXiv preprint. arXiv:1802.06205(2018).

[3] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. “Learning multiple layers of features from tiny images.” (2009).

[4] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. “GloVe: Global vectors for word representation.” *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).* 1532–1543.

[5] Mei-Chen Yeh and Yi-Nan Li. 2019. “Multilabel deep visual-semantic embedding.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 6 (2019), 1530–1536.

## Acknowledgements and Reproducibility

This research was supported in part by NSF grant 1816611.

All the code used in this work can be found at:
https://github.com/mattebaughman/embedding_net

Author contact: mbaughman@uchicago.edu