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

Matt Baughman
mbaughman@uchicago.edu
University of Chicago - Department of Computer Science
Chicago, Illinois, USA

Ian Foster (advisor), Kyle Chard (advisor)
foster@anl.gov,chard@uchicago.edu
University of Chicago - Department of Computer Science;
Argonne Nat. Lab - Data Science and Learning Division
Chicago, Illinois, USA

ABSTRACT

Recent advances in efficient neural networks and relational learning using word embeddings as prediction targets for image classification indicate 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 similar relational operations between images that are possible with word embeddings. In this essay, 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.

CCS CONCEPTS

• **Computing methodologies** → **Logical and relational learning**; **Transfer learning**.

KEYWORDS

Relational learning, transfer learning, neural networks, computer vision, word embeddings

ACM Reference Format:
Matt Baughman and Ian Foster (advisor), Kyle Chard (advisor). 2021. Embeddings are All You Need: Transfer Learning in Convolutional Neural Networks using Word Embeddings. In *St. Louis '21: The International Conference for High Performance Computing, Networking, Storage, and Analysis, November 14–19, 2021, St. Louis, MO.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.****/*****.*******

St. Louis '21, November 14–19, 2021, St. Louis, MO

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.****/*****.*******

1 INTRODUCTION

As machine learning models become increasingly larger and classification problems become more complex, adding the contextual relations between image classes that have been widely demonstrated in word embeddings will allow for increased insights on previously unseen entities. Prior work has focused on transfer learning from previously trained computer vision models [1] or label contextualization [5], we have designed our methods to start from scratch and focus solely on refining the image-to-vector mappings. Additionally, we selected our convolutional architecture based on a model that has shown state-of-the-art performance for model efficiency to demonstrate the effectiveness of these methods at all scales.

This work is important as the increasing applications of artificial intelligence means an increased likelihood of encountering unforeseen environments or conditions. The ability to extract meaningful information from these situations will be critical to many applications. Additionally, the use of these types of transfer learning methods could provide a more concrete method to evaluate how models may generalize when increasing the number of classes.

2 METHODS

In this work, we recreate the SimpNet [2] architecture and use the CIFAR-100 dataset [3] as input data with the GloVe [4] Wikipedia 2014 + Gigaword 5 pretrained word embeddings as targets. Specifically, we use the SimpNet architecture and replace the classification block at the end with three fully-connected layers that map the output of the convolutional network to the 50-dimensional word embeddings. In this way, the network will produce a 50d vector which is evaluated against the corresponding vector for each class name (i.e., the best word representing the name of each class was converted to a vector representation; this is not always possible, for example in the case of "pickup truck" we simply had to use "truck").

To evaluate our hypothesis in a rigorous manner, we iteratively retrained the network to estimate performance under leave-one-out cross validation. In other words, we trained the model using 99 of the 100 classes for 25 epochs, then evaluated performance on the left-out class. We repeated that process for each class.

3 EVALUATION

To evaluate the effectiveness of our training process outlined above, we calculated the mean square error (MSE) between the output and target vector. In the course of training the network, we achieved an 80% test accuracy for the categories used for training. While Figure 1

shows the median and quartile range loss during the training regime across the 100 different classes, we examine the direct effects of training on loss by looking at the relative value of the end validation lose to the start.

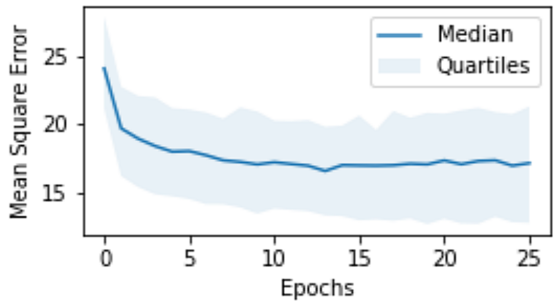

Figure 1: Loss (MSE) vs. epochs for the excluded class.

Figure 2 demonstrates the relative loss values for the left-out class following training. Specifically, we calculated the ratio of before to after loss values to demonstrate the increased performance on a specific class even when no members of that class were used in training. We found in the median case a reduction in loss by 31.0% following 25 epochs of training. In the extreme cases, loss was reduced by up to 65.5% or increased by 67.9%. We found that there is a correlation between performance gains and class feature representation in the training set (e.g., "willow" will likely perform well given the three other species of trees present in the dataset).

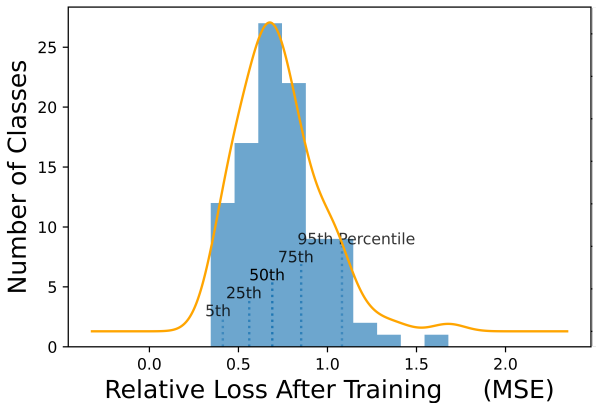

Figure 2: Relative loss values for the excluded classes.

4 CONCLUSION

In this work, we have demonstrated the potential for transfer learning via the use of word embeddings in novel environments. While computer vision and word embeddings have been previously combined to similar ends, we have demonstrated the ability of a model to learn the vectorization of never before seen classes without the need for previous training under a more traditional regime. This work motivates future demonstrations of transfer learning in this style as well as applications zero-shot computer vision and relational machine learning.

ACKNOWLEDGMENTS

This research was supported in part by NSF grant 1816611.

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.