Improved Generalization for Image Classification Through Memorization[*]

Roger Waleffe and Jason Mohoney
{waleffe, mohoney2}

February 24, 2021

1 Introduction

Machine learning models, especially deep neural networks, have shown great success across many disciplines, including computer vision [5, 6] and natural language processing [2]. While these models achieve high *average* accuracy on examples from the *same* distribution as their training data, their ability to generalize effectively remains in question. For example, the authors of [7] study the performance of object classifiers trained on ImageNet or CIFAR-10 by evaluating them on newly created test sets. By closely following the original dataset creation processes to generate the new data, the authors are able to study generalization in a benign environment. Even in this simple setting, they find ImageNet classification models perform 11-14% worse on the new test set when compared to the original test data. They conclude that "current classifiers still do not generalize reliably".

This problem is exacerbated when looking at the generalization performance on rare or atypical instances [3, 9]. Such examples are usually referred to collectively as the *long-tail* of the distribution. Many modern datasets for visual object recognition are known to be long-tailed—a few objects (or visibility patterns of a specific object) occur frequently, while many objects (or visibility patterns) occur infrequently, yet the total weight of these infrequent examples is a significant fraction of the dataset [3].

Naturally, learning to accurately model and predict atypical instances is a challenging problem, especially when current deep learning models primarily aim to minimize average error objectives. As a result, these networks often exhibit highly variable performance across examples from rare subpopulations [9]. When the number of such subpopulations is large however, as in long-tailed distributions, a model's performance on infrequent examples can even affect its average generalization performance. Thus, to improve the true accuracy of modern machine learning algorithms, it is necessary to consider generalization both in the context of common instances, and in the case of instances which appear only once, or a few times, in the training data.

Recent theoretical results [3] suggest that *memorization* of outlier labels is necessary for achieving close-to-optimal generalization error in the context of long-tailed distributions. Follow up work takes this claim a step further, arguing that "every sufficiently accurate training algorithm must encode essentially all the information about a large subset of its training examples" [1]. Intuitively, memorizing examples from subpopulations which appear only once (or a few times) makes sense—how else can a model generalize from one instance? Yet, this notion is in conflict with conventional beliefs which suggest memorization is detrimental to generalization. While overparameterized deep neural networks may have the capacity to implicitly memorize training examples, state-of-the-art object recognition models are not designed to explicitly memorize long-tailed distributions.

In this project, we aim to build on and test the recent theoretical work suggesting memorization is necessary for generalization by building a novel computer vision architecture which explicitly encodes this hypothesis. Specifically, the goal of our model is to concisely represent each training example and then use this information to classify future inputs. Our design takes inspiration from convolutional neural networks (known to extract useful image representations), k-nearest neighbor models (which explicitly memorize the training data), and transformer networks (useful for contextualizing a representation based on other information). We describe our intitial design in more detail in Section 3. We plan to evaluate our proposal by studying the generalization error and comparing against existing computer vision models (CNNs) and algorithms which naively memorize the training data (k-NN). Beyond average accuracy, we will study model and baseline performance on atypical subgroups and evaluate worst-group accuracy.

[*] https://rogerwaleffe.github.io/cs766/

2 Related Work

There are numerous works in the literature discussing machine learning algorithms and generalization. In addition to the theoretical works mentioned above, here, we focus on topics related to the hypothesis that memorization may improve generalization, especially in the case when the distribution is long-tailed.

Empirical Results on Memorization We are not aware of any computer vision works which empirically study models designed to explicitly memorize large fractions of the training images. However, overparameterized neural networks are known to have the capacity to memorize, even exhibiting the ability to fit completely random labels [10]. Further, recent trends seem to indicate that larger models generalize better. Additionally, some experiments have identified training images which appear to be memorized when using conventional models on standard benchmark datasets [4]. Removing these images decreases the accuracy of the model. Similarly, training algorithms which explicitly limit memorization, such as those use to achieve differential privacy, are known to perform worse than standard algorithms [3]. While these results indicate a possible correlation between memorization and generalization, here we propose a model which aims to encode this hypothesis explicitly.

Subgroup and Long-tail Accuracy A number of other works focus on improving the accuracy of underperforming subgroups. When subgroup labels are known, robust optimization techniques can improve worst-group accuracy [8]. More recent work [9] aims to improve performance when subpopulations are unlabeled—as is often the case for long-tailed distributions (e.g. rare visibility patterns). This technique makes use of clustering in image representation space to estimate subclass labels, and then uses this information for robust optimization. While these techniques improve generalization on tail examples by training to minimize the worst-group error, here we aim for the same goal, but by using an architecture which memorizes training instances.

3 Proposed Approach

In this section, we present our (initial) proposed architecture to incorporate explicit memorization of training examples into modern computer vision deep neural networks. Our design is depicted in Figure 1. While we wish to encode "essentially all the information about . . . training examples", raw image representations can be hard to work with. Thus, the first stage of our design will use an existing convolutional neural network architecture (e.g. ResNet) to extract image representations (embeddings e_i). All image representations for the training data will be explicitly memorized by our model. To compute the final output prediction, we plan to extend the basic idea of the k-nearest neighbor algorithm—classify an example based on the label of its neighbors—to include the full representation of each neighbor. Concretely, we plan to generate a contextual embedding (c_i) for an image based on the representation and label of each of its neighbors—a set which we can find efficiently and accurately using techniques such as locality sensitive hashing (LSH). The contextual embedding will then be used for the final classification. Motivated by recent work in natural language processing, we plan to use a transformer architecture to compute the contextual representation from an embedding and its neighbors.

A number of potential challenges and extensions to this base idea exist. We expect the main implementation difficulty to be training the full model end-to-end (i.e. differentiating through the neighborhood lookup). We believe this can be done. Alternatively, we could first train the CNN to produce meaningful representations (using self-supervised learning for example), and then train the transformer separately. It may be interesting to compare to this case even if the end-to-end training is successful. The modularity of the design also allows us to run interesting ablation studies. For example, we can study the affect of image representations by training our own CNN architectures or using massive pre-trained networks, and we can evaluate the affect of memorization and neighborhood aggregation by varying the transformer.

4 Timeline and Evaluation Plan

Our project and evaluation timeline are as follows:

- March 1: Implement the base CNN architecture.
- March 15: Implement the neighborhood lookup (LSH) and transformer architecture.
- March 29: Finish implementing the end-to-end training.
- April 12: Finish initial generalization experiments and comparisons to baselines.
- April 26: Study finer-grained accuracy (long-tail) and prepare presentation.
- May 5: Finish website.

As mentioned above, our main evaluation metrics will consist of the average generalization error and the performance on atypical subgroups. We plan to compare to standard deep learning models for computer vision, and simple models for explicit memorization.

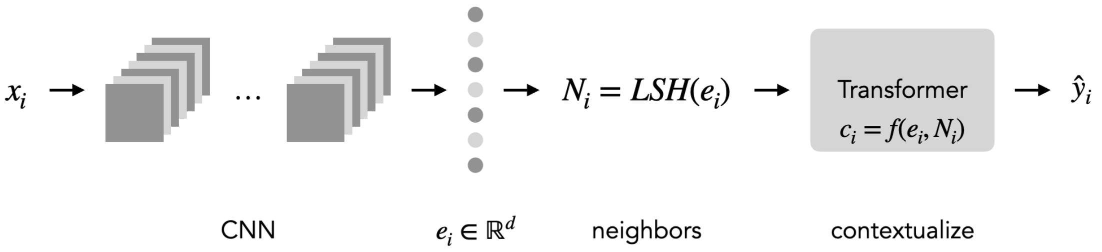

Figure 1: Graphical depiction of the proposed architecture to explicitly memorize training examples for image classification.

References

[1] G. Brown, M. Bun, V. Feldman, A. Smith, and K. Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? *arXiv preprint arXiv:2012.06421*, 2020.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] V. Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

[4] V. Feldman and C. Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *arXiv preprint arXiv:2008.03703*, 2020.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[7] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[8] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[9] N. S. Sohoni, J. A. Dunnmon, G. Angus, A. Gu, and C. Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *arXiv preprint arXiv:2011.12945*, 2020.

[10] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.