Machine Learning - Final Project Proposal

Isaac Wasserman
Department of Computer Science
Haverford College
Haverford, PA 19041
iwasserman@haverford.edu

1 Proposal

For my final project, I will explore the use of generative adversarial networks (GANs) for data augmentation. Although GANs are most popularly used for domain transfer, the same basic architecture is also a good candidate for a data augmentation strategy called hallucination in which new (fake) training examples based on a small number of existing (real) examples. Though theoretically, the addition of these fake examples should not improve the performance of a discriminative model (since they can be no more representative of the true distribution than the examples they are based on), empirical studies have observed accuracy improvements on multi-class classification of up to 13% on benchmark low-resource datasets such as Omniglot **?**. Data hallucination has been shown to improve performance, even without GANs; for example, randomized stem hallucination can improve the accuracy of morphological inflection models on low-resource languages by up to 76% **?**. Surely, there is something to this strategy.

I will start by implementing a simple GAN for the hallucination of a simple image dataset like MNIST; this process will allow me to become familiar with the technique in a medium that I can easily interpret. Using a small subset of this dataset, I will test the effects of the addition of augmented data on a good out-of-the-box classifier (likely Adaboost) as well as a well known few-shot classification **?** architecture (likely a matching or twin network) **? ?**. I then plan to test GAN based hallucination on the morphological inflection model built by Anastasopoulos and Neubig, replacing their largely randomized method with one that better models the phonotactics of the language. The dataset they used was published as part of the SIGMORPHON 2018 shared task **?**, and it includes subsets for 58 language pairs[1] and 91 individual languages.

Acknowledging the possibility that this may be more than I can accomplish in the given timeframe, I may end up needing to forgo the application of GAN hallucination to the inflection model and instead apply it to a simpler image classification task. In this case, I would first test traditional data augmentation methods and compare the effects to those of GAN hallucination.

2 Related work

The performance gains realized by Antoniou et al. **?** are supported by a fairly extensive body of similar evidence-based studies on data augmentation in low-resource settings. These studies are, more often than not, concerned with medical imaging, in which segmentation is a more salient issue than classification.

Shin et al., 2018 **?** leveraged the popular Pix2Pix **?** architecture to augment a brain-tumor segmentation dataset. This task was considerably more complex as it involved the hallucination of image pairs. However, training their segmentation model on a combination of real and fake data, they observed performance improvements of up to 16% over unaugmented data (minimum improvement of 1%).

[1]The Anastasopoulos and Neubig model was built to be trained first on a high-resource language and then fine-tuned for a low-resource language with similar genealogy or morphological features

Preprint. Under review.

They also tested the effects of an entirely synthetic dataset; however, this resulted in a significant loss of performance compared to the baseline. Despite the improvements realized when compared to their baseline segmentor, even their best model was unable to outperform the best-in-class reference model **?**.

Sandfort et al., 2019 **?** carried out a similar study using the also popular CycleGAN architecture **?** to segment anomolous CT scans of kidneys, livers, and spleens. On average, this study showed no appreciable difference between the performance gains afforded by traditional augmentation and GAN-based augmentation. However, when the resulting models were tested on images that were out-of-distribution[2], they found that while the baseline model scored 0.101, the traditionally augmented and GAN augmented models received scores of 0.535 and 0.747 respectively. This increased flexibility is interesting as it suggests that the augmented examples constitute a wider domain than the examples they are based on. From a theoretical standpoint, this would have to be possible if GAN augmented datasets were to outperform others.

While all of the other related works cited have used GANs to create synthetic image data, Gupta, 2019 **?** applies GAN-based augmentation to language data for the purpose of sentiment analysis. Unlike the other applications, labeled language data for sentiment analysis is not scarce. The datasets used each contain between 2000 and 4000 real examples, and the classifier was pre-trained on a dataset of over 1.6 million examples. Real and fake examples were combined in a somewhat nontraditional way; instead of concatenating the real and synthetic datasets, separate classifiers were trained for each and their outputs were combined via bagging. In these experiments, accuracy was only improved by up to 1.2% when fake examples were added. Though these gains are less impressive than others cited above, these results support the idea that GAN-based augmentation can be applied to domains outside of image data.

3 Timeline

April 6th through April 15th:

- Learn more about GANs
- Learn more about data augmentation

April 16th through May 1st:

- Implement a simple GAN for fake MNIST example generation
- Collect baseline and random hallucination accuracies for Anastasopoulos and Neubig model

May 2nd through May 13th:

- Implement GAN hallucinator to replace random hallucinator
- Write report

[2]While the training set was based on CTs with contrast, these out-of-distribution images came from scans performed without contrast.