# An exploration of how deep learning classifier design affects accuracy

Samsara Counts
George Washington University, countss@gwu.edu

***Abstract*—In this project, I study the effects of tweaking different parameters on deep learning classifier accuracy. To do this, I formulate a toy classification problem with images of humans in three broad occupational categories: Scientist, Artist, and Fashion (model). I use the toy dataset I created to train a deep learning classifier with 2 or 3 output classes, respectively, and strategically vary their training parameters, such as number of epochs and learning rate. I also experiment with using pre-trained and random ResNet-18 models as my base network. The highest performance I witnessed was 97% on a 2-way classifier trained on Scientist and Artist images and 83% for the 3-way classifier. Finally, I review my results along with observations about discrepancies in performance and the effects of different classifier design choices.**

***Keywords*—Machine Learning, Supervised Learning, Convolutional Neural Networks, Deep Learning**

## I. Introduction

Due to the explosion of interest in deep learning [4], convolutional neural networks evaluate images millions of times every day for a wide spectrum of applications. The popularity of machine learning with deep neural networks exploded, starting around 2012.

Despite the promise and widespread use of neural networks, they have some problematic limitations, the most significant being that they are difficult to interpret [8]. Even though a network may be accurate at solving a given problem, it is very challenging [12] to understand why a network is classifying an image into a given category. Though researchers have developed and refined methods of shedding light on why classifications are made, the field of neural network interpretability largely remains an open problem [12].

Despite the difficulty of understanding neural networks, machine learning engineers have developed a few standard practices to train accurate neural networks. One of them is using neural networks pre-trained on different widely-used benchmark datasets, such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [10]. The process of using is called transfer learning [9].

In this project, I study the effects of tweaking different training parameters on the accuracy of deep learning image classifiers. To do this, I formulate a very simple (toy) classification problem with images of humans in three broad occupational categories: Scientist, Artist, and Fashion (model). I use the toy dataset I created to train 30 deep learning classifiers with 2 or 3 output classes and strategically vary their training parameters, such as number of epochs and learning rate. I report the performance of the top 10 classifiers and the parameters they were trained on.

All my code for this project is available on Github at `https://github.com/samsaranc/ml_fp`.

## II. Related Work

The neural network architecture I used was ResNet [5], a state-of-the-art modeled developed by He et al. in 2016. Following common practice, I used ResNet models pretrained on the 2015 ILSVRC [10] dataset.

Lipton [8] examines both the challenges surrounding interpretability in deep learning and the motivations for underlying interest in interpretability, finding them to be diverse and occasionally discordant. Then, he addresses model properties and techniques thought to confer interpretability, identifying transparency to humans and post-hoc explanations as competing notions.

I developed my dataset from the datasets of two papers that focus on understanding bias and fairness in datasets, both conveniently working on a similar classification problem.

The first paper by Celis et al. [2] combined volume-based diversity (DPPs) with so-called combinatorial diversity– a measure of entropy over a single discrete low-cardinality attribute. For their experiments, they created a Scientist and Artist dataset that I use in creating our dataset.

The second paper by Kay et al. [6] characterize the gender bias present in image search results for a variety of occupations through several studies. They then experimentally evaluate the effects of bias in image search results on the images people choose to represent those careers and on people's perceptions of the prevalence of men and women in each occupation, finding evidence for both stereotype exaggeration and systematic underrepresentation of women in search results. For their studies, they create a dataset of google image searches per occupation, which I use in the creation of our dataset.

## III. Dataset

The final dataset for this setting is broken down into the categories Scientist, Artist, and Fashion (see Table I for details).

After reviewing the categories of the [10] (see Figure 1) and noticing that none of them contained images of humans, I decided to create a dataset of categories of images of humans. I was already familiar with the datasets of Kay et al. [6] and Celis et al. [2], so I started with the categories of Artist and Scientist, then added the Fashion category from [3].

I created a dataset for the task of classifying an image into the categories above by combining four sources:

| Class | Total Images |
|---|---|
| Scientist | 5223 |
| Artist | 5223 |
| Fashion | 5223 |
| **Total** | 15669 |

TABLE I: The datasets I created for my project and their respective sizes


Fig. 1: A subset of the 1000 image categories from the 2015 ILSVRC [10]

1) all three datasets from [2]
2) selected categories from [6] (specifics are below)
3) Not-pro-ED images with the hashtag #fashion from [3]
4) downloaded images from Google search queries (specifics are below)

I now discuss the curation techniques of those papers, followed by the steps of my supplemental data collection.

Celis et al. [2] curated their collection of images using Google image search as follows. Four search terms were used: (a) "Scientist Male", (b) "Scientist Female", (c) 'Painter Male", and (d) "Painter Female". They restricted the search to medium sized JPEG files that passed the strictest level of Safe Search filtering. They then collected the top 200 distinct images from each to create three datasets:

- Scientist: (a) and (b)
- Artist: (c) and (d)
- Scientist + Artist: (a), (b), (c), and (d).

Kay et al. [6] selected occupations from the Bureau of Labor Statistics, used the occupations as search terms, and downloaded the top 100 Google Image search results for each search term (from July 2629, 2013). The images were labeled by workers from Amazon's Mechanical Turk. For each image, Turkers were asked to indicate whether there were no people, one person, or more than one person in the image. They were also asked whether the people were women, men, children, or of unknown gender (and to check all that apply). Ultimately, three Turkers labeled each image.

Fig. 2: Example scientist images from the Scientist dataset [2]

For my dataset creation, when I read the list of occupations in [6], I noticed that several corresponded to the broader categories of Scientist and Artist. Hence, I used the pictures from those occupations in Kay's dataset for the one I was building. The occupations I chose from that list for those two categories are the following (totalling about 200 images per occupation):

- **Scientist**
  *doctor, computer programmer, chemist , biologist, engineer, lab tech, pharmacist, software developer, Web developer*
- **Artist**
  *architect, carpenter, painter, chef, designer, drafter, editor, photographer*

For my supplemental images in the Scientist and Artist categories, I used Google Image Download software to download all the images from a given Google image search. I searched the following queries on Google:

- **Scientist**
  *scientist, botanist, seismologist, scientist man black, scientist man asian, scientist man white, scientist woman white, scientist woman black, scientist woman asian, forensic scientist, aquatic scientist*
- **Artist**
  *artist asian woman, artist asian man, artist white woman, artist white man, artist woman hispanic, artist man hispanic, painter black man, painter black woman, painter man, painter woman, glassblower, potter man, potter woman, sidewalk artist*

After I downloaded the images, I manually processed them and discarded any images that did not contain a human or

did not fit the broader category, resulting in about 100 usable images per query.

The final dataset I obtained has the categories Scientist, Artist, and Fashion, with 5223 images per category. For the Scientist and Artist Categories,

### A. *Data Pre-processing*

To pre-process the data, I eliminated any images with overlaid text that obstructed any central figure(s) of the image. Second, I manually viewed each image and discarded any that were not photographs of people. Third, I programmatically eliminated any duplicate images with a quick average hashing function [1]. I also wrote Python scripts to uniquely name and identify each image with its original search query.

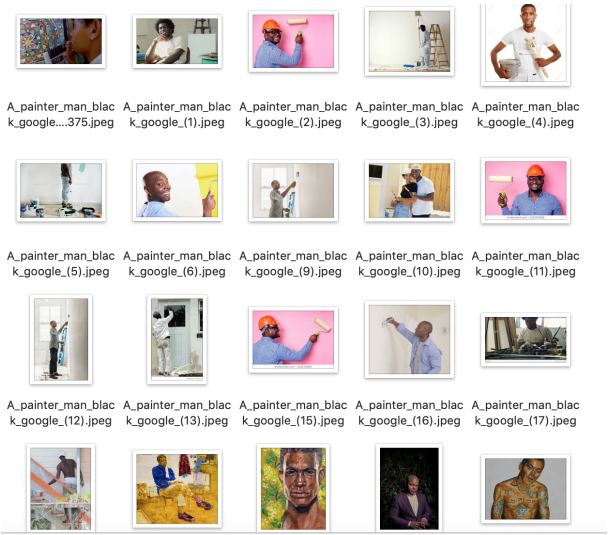


Fig. 3: Example Artist images from the Artist dataset

## IV. Approach

For each classifier I built, I trained an instance of the ResNet-18 [5] neural network architecture as my baseline classifier with an equal split on all of the data. To load and setup a model with the PyTorch system, I started with base code from the PyTorch tutorials.

I used an 20/80 test/training split, reserving 10% of my training data for the validation set. I wrote Python scripts to shuffle and randomize the image data I had for each split (see the `utils` folder in the Github repository). I encountered some difficult with corrupted images, so I also verified that each image was readable by https://www.pythonware.com/products/pil/ with another script. When loading the images for the neural network with PIL, I cropped them all to 256 pixels and used a series of random transforms and shifts to reduce bias in the dataset.

Following [7] and [11], I used a multitude of different techniques to prevent the classifier from putting too much weight into irrelevant features such as image coloring (tint), cropping, and positioning. First, I resized images and randomly sampled from the range [256, 480] to scale them. Next, I randomly sampled a 224224 crop from each image or its

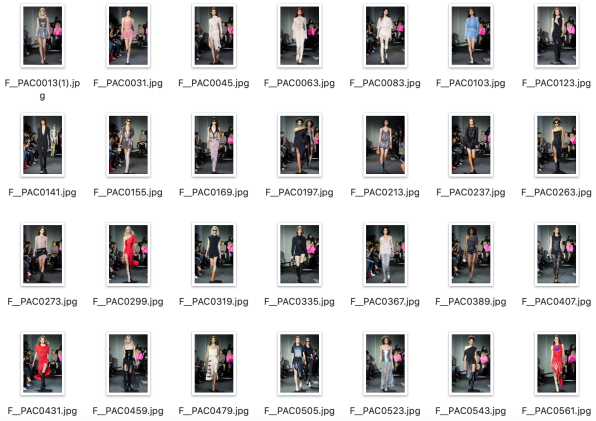


Fig. 4: Example Fashion images from the Fashion dataset

horizontal flip, with the per-pixel mean subtracted. I use the standard color augmentation as in [7] and batch normalization after each convolution before activation. For the network's optimizer, used Stochastic Gradient Descent with a batch size of 175. Finally, I started the learning rate at 0.01 and divided it by 10 when the network's error plateaus.

## V. Experiments

When training the classifiers, I fixed 10 different sets of parameters, then trained 3 versions of each combination of parameters on the learning rates of .01, .001, and .1, resulting in 30 total classifiers. I report the best-performing classifier for each of the categories.

I now describe the parameters I varied. There are three possible 2-way classifiers. Then using the binary characteristic of pre-trained vs. not-pre-trained (on Imagenet), I investigated six 2-way classifiers. I fixed the number of epochs to be 15 for the 2-way classifiers. There is only one possible 3-way classifier and I chose to vary the number of epochs between 15, 25, and 39, so that was three classifiers. I also tried a not-pretrained version of the 3-way with 15 epochs.

## VI. Results

I now describe the results of my 10 best-performing classifiers. Overall, the 2-way classifiers substantially outperformed the 3-way classifiers, with top accuracy rates in the $90^{th}$ percentile (see Figure 5), with the best 3-way classifier being approximate 80% accurate (see Figure 7).

The figures below illuminate different accuracy results on different datasplits as well as using versions of Resnet [5] that are either pre-trained on the ILSVRC dataset [10] or randomly initiated. Even though the ILSVRC [10] categories have nothing to do with the three categories I classify into, I observed that re-training ResNet models pre-trained on ILSVRC performed substantially better than models with random weights (Figure 5 and Figure 6). For the randomly-initialized 3-way classifiers, accuracy was in the $50^{th}$ percentile; for the random 2-way, accuracy was around the $60^{th}$ percentile.

For the 2-way classifiers, different learning rates did not seem to have much effect on each classifier's accuracy. For

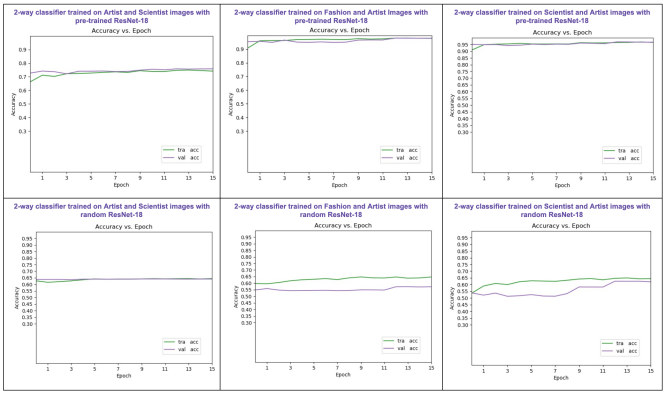


Fig. 5: Best 2-way classification results (with ResNet models either pre-trained on ILSVRC [10] or randomly initialized)

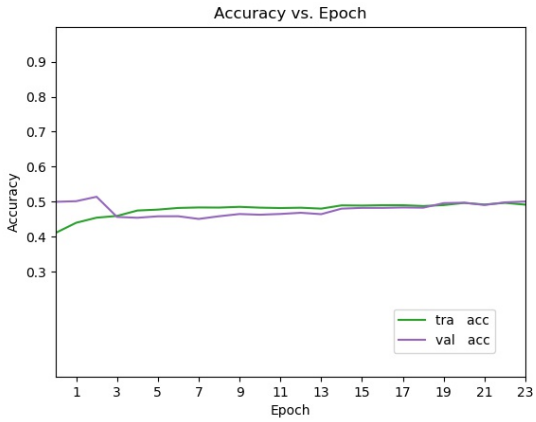


Fig. 6: Best 3-way classification results on randomly-initialized ResNet

the 3-way classifiers, longer epochs only seemed to alter the evolution of the error rate, as shown in Figure 8.

## VII. Discussion

Here I discuss some possible explanations for the outcomes I observed in my experiments, as well as some key limitations of my project.

The most important limitation to note is the bias embedded in my dataset caused by how I created it. Essentially, I handpicked the Google search queries I used for the different categories, and I handpicked the occupations from Kay et al. [6] that corresponded to the larger categories of Artist and Scientist. These kind of subjective choices are permissible for a class project, but not to create a dataset for research projects or to be deployed in the real world. Ideally, given more time and financial resources, I would create a dataset by taking occupation terms from the U.S. Bureau of Labor and Statistics (similar to the procedure used by [6]) and then formulate queries that consist of permutations of one occupation and one or two different "diverse" descriptors—genders and ethnicities/races. Then, I would scrape images from the web (Google, Bing, Twitter) using all of those

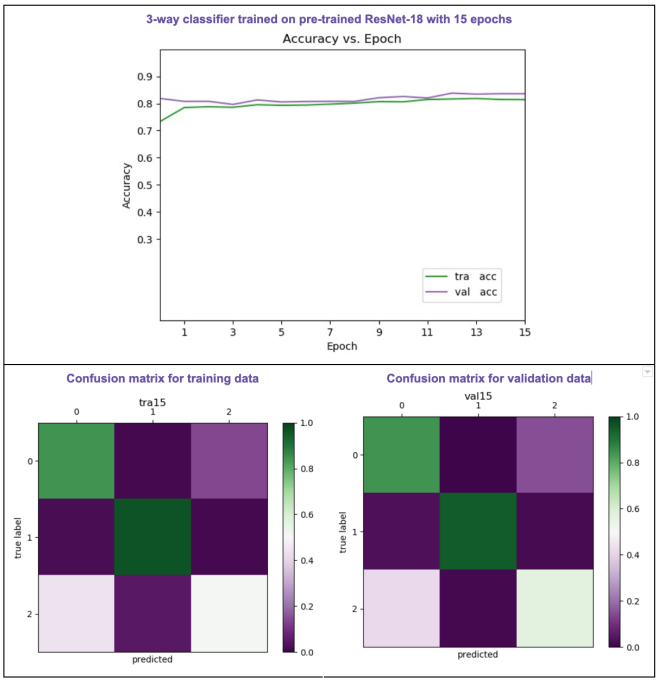


Fig. 7: Best 3-way classification results (ResNet pre-trained on ILSVRC [10])

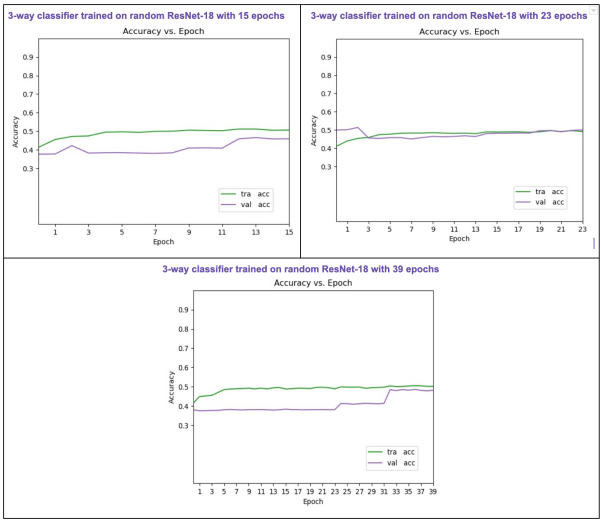


Fig. 8: Best 3-way classification results (ResNet pre-trained on ILSVRC [10]) with 15, 25, or 39 epochs

queries. The procedural aspect of how I would create the dataset is essential, because it ensures that my experiments are reproducible by other researchers.

Another limitation of my project is the small amount of data I used. Though the size of my dataset is around ten to fifteen thousand images (for 2-way and 3-way classifiers, respectively), that is still pretty small for deep-learning-scale classifiers. If I were to use this dataset in practice, I would hope to get at least 40 thousand images in each category. With a small dataset, there is a higher likelihood that my classifiers

are overfitting on my data. In fact, I am fairly certain that the pre-trained 2-way classifiers are overfitting, because of their extremely high accuracy.

That the 2-way classifiers outperformed the 3-way classifiers in accuracy does not surprise me, because there are certain similarities between the images captured in each dataset, as I describe in the following paragraph, that I think might be harder to differentiate between. What did surprise me, however, is that the pre-trained classifiers performed much, much better than the randomly initiated classifiers. Perhaps I needed to train the randomly-initiated classifiers for longer to reach the optimal error rate, but, based on my observations and experiments with different learning rates and different epochs, I do not think that is the case. It is particularly strange because the ILSVRC categories are completely unrelated to the categories I was classifying into.

Personally, I was surprised that the highest performing classifier, depicted in Figure 5, was the 2-way Fashion and Artist images. In looking through the Fashion images (more examples are in Figure 4), I noticed that a large majority of them are either professionally taken (runways, catalogs, photo shoots) or selfies. To contrast, almost all of the images in the Scientist and Artist datasets are professionally taken. Another difference is that in both the Scientist and Artist datasets, it is common for an image's subject to be holding some relevant accessory (test tubes, a clipboard) in their pictures. To contrast, in the Fashion dataset most subjects do not have accessories that occupy their hands and/or affect their poses. For these reasons, I thought that it would be "easier" for the classifier to differentiate between images from the Fashion and Scientist and/or Artist and Scientist images, which did not happen.

## VIII. Conclusion

In this project, I trained instance of the ResNet deep learning architecture to recognize Artist, Scientist, and Fashion images, With the array of classifiers I trained, I demonstrate that, even with a toy problem, existing deep learning tools can lead to numerically high classification accuracy. Even though the ILSVRC [10] categories have nothing to do with the three categories I classify into, I observed that re-training ResNet models pre-trained on ILSVRC performed substantially better than models with random weights. Longer epochs did not seem to have much effect on accuracy nor did different learning rates.

In the future, to make a more rigorous version of this project, one could design robust tests to check if our classifiers are overfitting. One way to approach this would be gathering images from the Internet from different sources, such as personal blogs. Another direction to consider would be using an ensemble of classifiers at a time instead of just one. By training several neural network models, one could then use ensemble methods with the 2-way classifiers. Finally, one important direction could be investigating different types of randomly-initiation methods for network parameters and comparing how classifiers perform.

## References


[1] J. Buchner, "The imagehash python library," https://pypi.org/project/ImageHash/, accessed: 2018-04-30.

[2] L. E. Celis, A. Deshpande, T. Kathuria, and N. K. Vishnoi, "How to be fair and diverse?" *CoRR*, vol. abs/1610.07183, 2016. [Online]. Available: http://arxiv.org/abs/1610.07183

[3] S. N. Counts, M. Justine-Louise, and R. Pless, "Characterizing the visual social media environment of eating disorders," in *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, Oct 2018.

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[6] M. Kay, S. N. Patel, and J. A. Kientz, "Unequal representation and gender stereotypes in image search results for occupations," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15. New York, NY, USA: ACM, 2015, pp. 347–356. [Online]. Available: httpdoi.acm.org10.11452702123.2702603

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017. [Online]. Available: http://doi.acm.org/10.1145/3065386

[8] Z. C. Lipton, "The mythos of model interpretability," *CoRR*, vol. abs/1606.03490, 2016. [Online]. Available: http://arxiv.org/abs/1606.03490

[9] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct 2010.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[12] D. S. Weld and G. Bansal, "The challenge of crafting intelligible intelligence," *Commun. ACM*, vol. 62, no. 6, pp. 70–79, May 2019. [Online]. Available: http://doi.acm.org/10.1145/3282486