Image Recognition Training via Reddit

Computer Vision Project Paper

Kenneth Sanders

May 3, 2018

Introduction

Neural Networks (NNs) have been a popular concept for use in machine learning and computer vision in recent years. General uses of Neural Nets include classification, identification, prediction and recognition of real world data. However, a neural network is only as good as its training set, which must be accurately labeled yet vary enough to apply to all cases in the evaluation set. In the case of image recognition, a large set of categorized images must be provided, ideally with the categories of the training images matching the possible categories of image that the neural net will be used with. To fill this need, we look at image-based subreddits, which offer a vast dataset of images categorized into specific topics by Reddit users. The topic of an individual subreddit may be used to label each image in the dataset, which creates a dataset usable for machine-learning-based image classification. This project aims to show that the union of labeled subreddit datasets result in respectable accuracy when compared to traditional hand-labeled datasets used in machine learning, while offering an open-source, personally-available, and customizable dataset generator.

Reddit is a news and content aggregation site divided into topics by subreddits with many interaction similarities to online forums. Reddit users post images, text, or hyperlinks alongside a title to a relevant subreddit, and other users are subscribed to the subreddit may view it during their time browsing Reddit. Users also have the ability to vote on posts to either increase the post's karma score, which is then summed across all of a user's posts, giving a user an overall post karma score. In addition to the obvious potential dataset of subreddit categorized images, this democratic process of voting on submitted content in the forms of posts potentially creates a measure of how well each image represents the subreddit/category it was posted to. Subreddit URLs take the form of

https://www.reddit.com/r/{subreddit}

where {subreddit} is the name of the subreddit. This leads to the general method of referring to a subreddit, which is in the form *r/{subreddit}*, such as r/news or r/cats. Subreddits come in all types and sizes, ranging from the default Subreddits for new users like r/funny and r/pics to obscure subreddits frequented by only those familiar with their content, a few of my own personal examples being r/programmerhumor, r/bmx, and r/i3wm. In 2013, Reddit accepted over 40 million posts and served over 700 Million

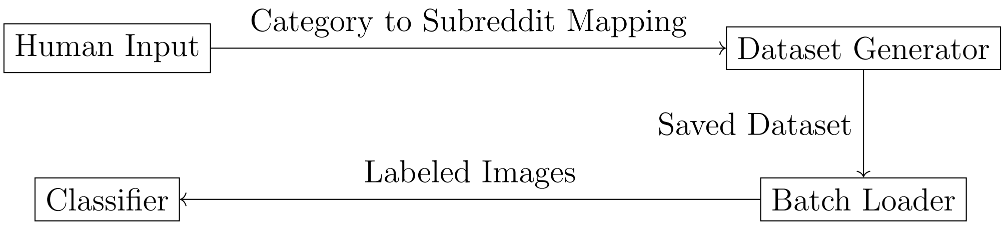

Fig. 1: Proposed process of generating a dataset and feeding it into a image recognition network.

visitors [Red]. While the fraction of those posts that are unique and viable as training data is unknown, the scale of the site is certainly promising. This project will aim to design, implement, and evaluate a convoluted neural network for image recognition which uses subreddit categorized data for training. For simplicity's sake, the initial neural net will be tested on the Kaggle competition "Dogs vs Cats"[Kag], since there are substantial supplies of cat and dog images on Reddit to be used in training. Once the theory is confirmed through initial evaluation, expansion of the design to a more generalized case will be examined, ideally culminating in a publicly available, open source program capable of recognizing the most relevant subreddit to an image given a set of subreddits. The program could also be easily adapted to do exactly what it was "trained" to do, and be applied to Reddit itself by making suggestions to users as to which subreddits would benefit most from their images, amongst other possibilities.

Approach

The approach is divided into two phases as part of a “ramp-up” design. The first phase consists of testing the hypothesis via training a convolutional network to differentiate between dogs and cats from a Reddit collected dataset, and comparing the results to the same network which was trained on the Kaggle Dogs Vs. Cats competition dataset [Kag]. The second phase will aim to expand on the first, and apply the hypothesis to larger and more varied dataset, namely CIFAR10, by collecting a similar dataset from Reddit, training two networks on each of the datasets, and comparing the resulting predictions made on both evaluation sets by both networks. By comparing the resulting predictions of both networks on both evaluation sets, we can obtain a better understanding of the quality of the trained networks and their ability to extrapolate results from never-before-seen images. The two phases of the project are discussed further in the following subsections.

Dogs Vs. Cats

For this phase, we will compare image classification neural networks that were trained on the created Reddit Dogs Vs. Cats dataset to the same networks trained on the publicly available Kaggle Dogs Vs. Cats dataset. Since the Kaggle dataset is known to be an accurate and well curated dataset, a comparison between a network trained on any other Dogs Vs. Cats dataset and a network trained on the Kaggle dataset should give a good

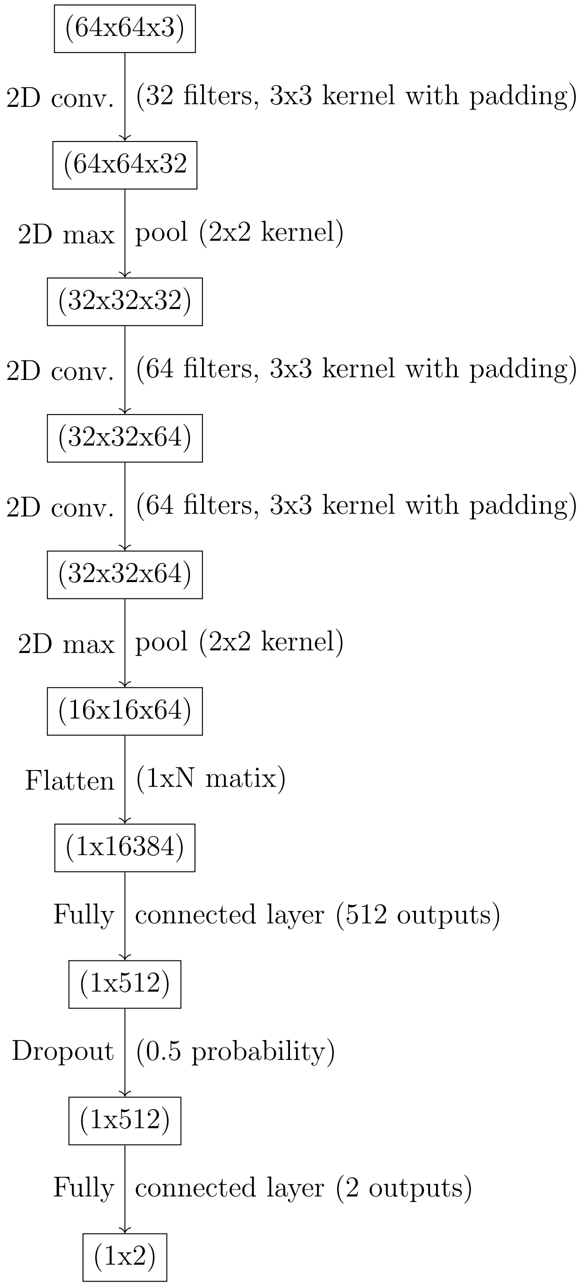

Fig. 2: Structure of the Simple CNN used for initial evaluation.

idea of the viability of the other dataset for training. Two different published neural network configurations will be used (with slight modifications) in order to avoid any bias or entropy from individual network configurations; Namely, the example CNN by Tensorflow built specifically for the CIFAR-10 dataset, as well as the Inception V3 model [**TFCNN**] [Sze+15].

Expansion

Once the dataset-generator and neural networks are written for the first phase, they should be easily modified to any set of image categories. From this, datasets will be generated to mirror the categories available in other publicly available and known-accurate datasets; for instance the CIFAR10/100 datasets. The network models used for the Dogs Vs. Cats phase may also be reused for this phase, with slight modifications to the input and output tensors. This code reuse allows us to spend more time perfecting the dataset generator and the classifiers than worrying about scaling the hypothesis to a more generalized form.

Dataset

The Datasets used are both official datasets, and datasets generated from Reddit. Specifically, the Kaggle Dogs Vs. Cats Training (Testing is not labeled, and so cannot be used for our purposes) dataset, along with CIFAR10 [Kag] [KH09]. The Kaggle Dogs Vs. Cats training set consists of a total of 25,000 labeled images of cats and dogs. A matching dataset of the same size is generated from Reddit, with 25,000 images of dogs and cats, labeled by the category of the subreddit that the image was found in. The expansion phase will use the 10-class CIFAR10 dataset and a Reddit-generated counterpart, which we refer to as Reddit10. The 10-class datasets were originally inended to contain 60,000 labeled iamges. The generated datasets were not modified whatsoever after collection (i.e. no hand-labeling, irrelevant image culling, etc.), so the datasets may show that the entire process of collection to training may be automated with human supervision as to acceptable dataset size and parameters to the recognition network. The datasets themselves are split into training and evaluation datasets with a ratio of 9:1 respectively. The division of test and training datasets allow us to evaluate the resulting neural networks with images that they have never been trained on, which in turn allows us to draw concrete conclusions on the abilities of the trained models in regards to class-prediction as opposed to class-recollection.

Evaluation

The previous sections were executed as described, and their results recorded. The following sections discuss those results.

Dogs Vs. Cats

The results for the Dogs Vs Cats phase were indeed promising, and may be noted in Figure 3. With a convolutional neural net structured as show in Figure 2, the network trained on the Reddit training set achieved slightly better results on the Reddit test set than the network trained on the kaggle training set, with 72.54% and 70.56% accuracy respectively. When these same nets were tested on the Kaggle Dogs Vs. Cats training set however, the network trained on the Kaggle training set did significantly ($\approx 9\%$) better than the network trained on the Reddit dataset, with accuracies at 84.86% and 75.78% respectively. From this, we may draw a reasonable conclusion that the dataset generated from Reddit is similar enough in label accuracy and image class relevancy to that of the Kaggle dataset, supporting the hypothesis.

Expansion

The next set of tests attempted to expand the hypothesis to the CIFAR10 dataset, which would help support the theory that this data-gathering method can be expanded to mimic other datasets with more diverse categories. Of course, obtaining a dataset only allows us to get to the model training phase, so the accuracy results when comparing our dataset to others depends just as much (if not moreso) upon the model being trained as the dataset being used for training. We used both the simple CNN shown in Figure 2, and the Inception v3 model described by Szegedy et al. in their paper *Rethinking the Inception Architecture for Computer Vision.* The results can be noted in Figure 4. The results of our expansion phase are less promising than the Dogs Vs. Cats phase, with the models trained on our Reddit10 dataset obtaining a notably lower accuracy than the same models trained on the CIFAR10 dataset. The most likely cause of this is the fact that the Reddit10 dataset did not obtain an equivalent number of images to the CIFAR10 dataset, since many subreddit-categories are not populated as well on reddit, resulting in a lacking number of images for frogs and deer. This resulted in less than 2,000 images for those classes in our Reddit10 dataset, compared to the 6,000 images for each category in the CIFAR10 dataset.

Model	Training Set	Test Set	Accuracy
Naive	Kaggle Training Set	Reddit Test Set	70.56%
Naive	Kaggle Training Set	Kaggle Test Set	84.86%
Naive	Reddit Training Set	Reddit Test Set	72.54%
Naive	Reddit Training Set	Kaggle Test Set	75.78%
Inception V3	Kaggle Training set	Reddit Test Set	80.92%
Inception V3	Kaggle Training set	Kaggle Test Set	85.38%
Inception V3	Reddit Training set	Reddit Test Set	82.64%
Inception V3	Reddit Training set	Kaggle Test Set	79.26%

Fig. 3: Dogs Vs. Cats phase results, comparing the effectiveness of the Reddit training set to the Kaggle Training set.

Model	Training Set	Test Set	Accuracy
Naive	CIFAR10 Training Set	Reddit10 Test Set	67.42%
Naive	CIFAR10 Training Set	CIFAR10 Test Set	69.03%
Naive	Reddit10 Training Set	Reddit10 Test Set	68.15%
Naive	Reddit10 Training Set	CIFAR10 Test Set	67.66%
Inception V3	CIFAR10 Training set	Reddit10 Test Set	73.32%
Inception V3	CIFAR10 Training set	CIFAR10 Test Set	78.45%
Inception V3	Reddit10 Training set	Reddit10 Test Set	74.43%
Inception V3	Reddit10 Training set	CIFAR10 Test Set	65.91%

Fig. 4: Expansion phase results, comparing the effectiveness of using our generated Reddit10 dataset Vs. the hand annotated CIFAR10 dataset.

Analysis

Overall, our results were as hypothesized. Obviously, every network model and dataset can not be tested for accuracy, but our experiments do successfully compare the resulting datasets from our generation method to hand annotated datasets via a few simple cases. The results were as expected, with models trained from our datasets having only slightly lower accuracy than that of hand-annotated datasets overall. For instance, in Figure 3, one will note that for both models, the model excelled at classifying images in the evaluation set that originated from the same source, while performing somewhat worse than the evaluation set from the other source. This suggests that there is consistency within the datasets, but the datasets as a whole may not be as consistent with each other. In Figure 4, this trend continues, with the CIFAR10 trained networks performing better on the CIFAR10 test sets than the Reddit trained networks, while performing slightly worse than the REDDIT10 trained networks on the REDDIT10 evaluation set. This does allow us to determine that our experiments are successful and crowdsourced datasets are a viable alternative to hand-labeled ones. However, these results are dependent upon many factors, such as Reddit itself, not the dataset generation algorithm or image recognition model alone. This means that the generated datasets could change over time, or more fundamentally, that two datasets generated from the same subreddits at different times will be different. For consistency, the dataset should be gathered and saved locally, which our method does normally carry out. Even still, the potential for change and lexical drift as Reddit itself changes means that these datasets are probably not the best for a consistent and optimally accurate classifier.

Conclusion

We conclude this paper by looking forward to the next steps in this field. Dataset generation allows scientists around the world to obtain massive datasets that would otherwise be unfeasible to collect. By showing that datasets generated can result in similar model accuracy after training, we can begin to rely on the conclusions drawn from the datasets. Unfortunately, more work must be done to analyze possible biases in the dataset generation model, such as the demographics of Reddit users and how that may affect the sterotypical image link posted to Reddit. However, as a whole our results do show consistency with standard datasets, so we may assume that for non-essential applications that any existing bias is fairly negligable, which is what we aimed to prove.

References

[Aba+16] Martin Abadi et al. “TensorFlow: A System for Large-scale Machine Learning”. In: *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*. OSDI’16. Savannah, GA, USA: USENIX Association, 2016, pp. 265–283. ISBN: 978-1-931971-33-1. URL: `http://dl.acm.org/citation.cfm?id=3026877.3026899`.

[Boe14] Bryce Boe. “PRAW: The Python Reddit API Wrapper”. 2014. URL: `https://github.com/praw-dev/praw`.

[Kag] *Kaggle, Dogs Vs. Cats Competition*. `https://www.kaggle.com/c/dogs-vs-cats`. Accessed: Feb. 12 2018.

[KH09] A Krizhevsky and G Hinton. “Learning multiple layers of features from tiny images”. In: 1 (Jan. 2009).

[KSH17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Commun. ACM* 60.6 (May 2017), pp. 84–90. ISSN: 0001-0782. DOI: `10.1145/3065386`. URL: `http://doi.acm.org/10.1145/3065386`.

[Le+12] Quoc Le et al. “Building high-level features using large scale unsupervised learning”. In: *International Conference in Machine Learning*. 2012.

[ML12] Julian McAuley and Jure Leskovec. “Image Labeling on a Network: Using Social-Network Metadata for Image Classification”. In: *Computer Vision – ECCV 2012*. Ed. by Andrew Fitzgibbon et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 828–841. ISBN: 978-3-642-33765-9. URL: `https://cs.stanford.edu/~jure/pubs/image-eccv12.pdf`.

[Niu+14] Z. Niu et al. “Semi-supervised Relational Topic Model for Weakly Annotated Image Recognition in Social Media”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 4233–4240. DOI: `10.1109/CVPR.2014.539`.

[Red] *Reddit blog*. `https://redditblog.com/2013/12/31/top-posts-of-2013-stats-and-snoo-years-resolutions/`. Accessed: Feb. 12 2018.

[SMC12] J. Schmidhuber, U. Meier, and D. Ciresan. “Multi-column deep neural networks for image classification”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. Vol. 00. June 2012, pp. 3642–3649. DOI: `10.1109/CVPR.2012.6248110`. URL: `doi.ieeecomputersociety.org/10.1109/CVPR.2012.6248110`.

[SZ14] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *CoRR* abs/1409.1556 (2014). arXiv: `1409.1556`. URL: `http://arxiv.org/abs/1409.1556`.

[Sze+15] Christian Szegedy et al. “Rethinking the Inception Architecture for Computer Vision”. In: *CoRR* abs/1512.00567 (2015). arXiv: `1512.00567`. URL: `http://arxiv.org/abs/1512.00567`.

[Wan+12] Sheng-Yuan Wang et al. "Learning by expansion: Exploiting social media for image classification with few training examples". In: *Neurocomputing* 95 (2012). Learning from Social Media Network, pp. 117 –125. ISSN: 0925-2312. DOI: `https://doi.org/10.1016/j.neucom.2011.05.043`.