Drawing: A New Way To Search
(Computer Vision)

Nguyet Minh Phu
Department of Computer Science
Stanford University
minhphu@stanford.edu

Connie Xiao
Department of Computer Science
Stanford University
coxiao@stanford.edu

Jervis Muindi
Department of Computer Science
Stanford University
jmuindi@stanford.edu

***Abstract*—This is a project proposal to explore ways in which machine learning can be leveraged to provide new ways of interacting with computers and communication in the realm of image search.**

1. Introduction

Using words can be limited when communicating across cultures and literacy level. An English learner may have challenges trying to come up with the right word for cup. A kindergartener may have trouble learning to write the word apple. Images are an alternate medium of meaning communication that can beneficially bridge those divides.

1.1. Problem definition

Our project aims to recognize the meaning of hand-drawn doodles, a critical first task in order to build any system that uses hand-drawn images for communication. If successful, this model can be applied for a variety of interesting tasks, including a new search interface where one can draw what they need and search for it, or an app where a language learner can draw an image and get the translation immediately.

More specifically, we first want to develop a system to recognize labels of hand-drawn images based on Google's QuickDraw dataset.

1.2. Dataset

Google's QuickDraw is the world's largest doodling dataset, consisting of hand-drawn images from over 15 million people all over the world. This dataset is highly applicable to our goal because doodles are what we will draw when we need to communicate something quickly via image. The dataset is available at https://github.com/googlecreativelab/quickdraw-dataset.

2. Methods and Intended Experiments

We want to benchmark several popular methods used for image classification to see how they apply to our problem of classifying doodles. We want to compare and benchmark several different approaches, including classical machine learning approaches and deep learning methods.

2.1. Baseline

For baseline, we will implement Logistic Regression with using our input dataset of raw image strokes over time.

2.2. Classical Machine Learning

For classical machine learning, we will use variants of a Support Vector Machine (SVM), comparing the effect of different image representations and kernels on SVMs performance. For representation, we plan to use Histograms of Pixel Intensity [Chapelle et al., 1999] and Edge Histogram Descriptor [Bhattacharya et al., 2006]. We notice from our literature review that the majority of image representation methods are applicable to colored images taken by camera, focusing on color distribution and texture [Tong and Chang, 2001]. These are not applicable to our black-and-white hand-drawn doodles. Thus, we also want to develop and explore other more compact yet effective ways of representation. For instance, a vector representing the location of the black ink used in the original QuickDraw dataset may be sufficient. For the kernel, we plan to use and compare the outputs of several, including Polynomial, Gaussian, Gaussian Radial Basis (RBF), Hyperbolic Tangent, and Sigmoid. Through experiments, we aim to find the most efficient kernel for our problem, then investigate further why such a kernel may be good for doodles.

2.3. Deep Learning

For deep learning, we will use Convolutional Neural Net (CNN) and explore also the applicability of transfer learning for our task. For transfer learning, there are two broad strategies [CS231n, 2018]. The first is to use the pre-trained model as a fixed feature extractor. New candidate inputs would be run through the pre-trained model, but will stop at the penultimate fully connected layer to get a feature vector. The other is with fine-tuning which allows

for back-propagation to occur and update the weights of the pre-trained model network. In the paper, "Do Better ImageNet models Transfer Better" [Kornblith et al., 2018], they explore how well the best performing ImageNet trained models do when applied to other image tasks and datasets. They find that the best performing ImageNet model do not act as the best fixed feature extract for other image tasks. The best fixed feature extractor was the ResNet network which only had relatively modest accuracy on ImageNet. However, they do find that the best image net models, when fine-tuned to the new image tasks do tend to fairly well on that new task. Thus, for our project, we would like to use ResNet as a fixed feature extractor and apply it to our classification task. Time-permitting, we would also want to take a top performing imagenet model (NasNet) and perform fine tuning of the models to see how well it can perform for our image task.

3. Evaluation

3.1. Accuracy

Generally, many image classification learning algorithms are evaluated with an error matrix [Lu and Weng, 2007], which works for data with clearly defined outputs. This is the case with our dataset. We also looked into other metrics that could reveal more information about the performance of our experiments. A number of image classification publications refer to evaluation metrics laid out in the ImageNet Large Scale Visual Recognition Challenge [Krizhevsky et al., 2012, Simonyan and Zisserman, 2014]. This includes AUC [Deng et al., 2009] as well as Top-5/Top-1 and average precision metrics [Russakovsky et al., 2015]. We can use average precision metrics as a basic accuracy evaluation of our model. AUC could prove useful since it is threshold independent and can explain how well our model separates the negative and positive classes. However, the Top-5/Top-1 metrics are less applicable because our input data only contains one object.

3.2. Efficiency

There are not as many evaluation metrics for efficiency. Some evaluation metrics include a combination of accuracy and duration [Goyal et al., 2017] or number of epochs [Lin et al.], but do not take into account other aspects of efficiency like computational power. We hope to find more nuanced evaluation metrics for efficiency in our project.

References

Prabir Bhattacharya, Mahmudur Rahman, and Bipin C Desai. Image representation and retrieval using support vector machine and fuzzy c-means clustering based semantical spaces. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 1162–1168. IEEE, 2006.

Olivier Chapelle, Patrick Haffner, and Vladimir N Vapnik. Support vector machines for histogram-based image classification. *IEEE transactions on Neural Networks*, 10(5): 1055–1064, 1999.

CS231n. Convolutional neural networks for visual recognition: Transfer learning. 2018. URL `http://cs231n.github.io/transfer-learning/`.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? *CoRR*, abs/1805.08974, 2018. URL `http://arxiv.org/abs/1805.08974`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Yuanqing Lin, Fengjun Lv, Shenghuo Zhu, Ming Yang, Timothee Cour, Kai Yu, Liangliang Cao, and Thomas Huang. Large-scale image classification: Fast feature extraction and svm training.

Dengsheng Lu and Qihao Weng. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870, 2007.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118. ACM, 2001.