Nishant Velagapudi
W241 – Essay 2
Experimental Research Proposal

Image Classification Accuracy on Hand-Drawn Sketches

Question:

I propose an experiment design to identify whether drawing complexity affects classification accuracy by a simple ML algorithm. Image classification is ubiquitous: search engines do a phenomenal job of tagging photos with relevant features. I am curious to apply image classification to human drawings. With less time, artists will barely sketch a drawing (e.g. a stick figure). With more time, artists start to fill in detail. I am curious to know how much additional detail in a drawing causes an improvement in image classification.

Motivation:

The motivation is more to understand machine perception of drawing versus human perception of drawing. Games like "Taboo" are centered around a human version of this experiment: how quickly can a human mind accurately classify an image as it is constructed. By restricting the time allowed to complete each image, we can control for average "quality." We estimate that faster images will have less lines and features in general. Prediction accuracy across each type of image will help us identify what key features are important in classification for ML algorithms, which can lead to a better understanding of what could cause incorrect classifications. Operationally, this can give us a sense of where the underfit/overfit balance is for image classification as well: more detail is more computation and complexity. Ideally, algorithms aim for a balance of simplicity and accuracy for optimal performance and generalization.

I am curious to see if there are any images that are easy for a human to interpret, but difficult for the algorithm, or vice versa. Discrepancies between human perception and algorithm perception are potential spaces of learning: human cognition is considerably more complex and sophisticated than the simple patterns picked out by most machine learning algorithms. An interesting finding would be features within an image that are obvious clues to classification to the human eye but are missed by algorithms.

Treatments:

We aim to investigate how the time taken to draw a sketch affects algorithmic image classification accuracy. I propose using Mechanical Turk to incentivize random people to submit sketches. We define 10 possible images for the artist to sketch: these 10 images will be of a similar complexity. The artist will be given a predetermined length of time to sketch each of these 10 and instructed to represent the prompt image as well as possible. The artist will then be requested to label 10 images drawn by other artists. The images to sketch will be sampled uniformly, while the images for human labelling will be selected to ensure an even distribution of human labels across other drawings. The time allowed per sketch will vary linearly: we will give randomly assign artists 30 seconds, 45 seconds, 1 minute, and 2 minutes per drawing. An

example trial: artist receives prompt to draw ten sketches. The artist is given 45 seconds (variable under investigation) per image to sketch. In processing, we will train and evaluate the model separately within each of the treatment groups and evaluate accuracy as appropriate.

Treatment variation will be random. We could consider a stratified approach to treatment assignment: we could ask a question leading into the experiment that gauges how well the individual thinks they can draw. We could then stratify sampling across this skill level: each possible treatment would then have a similar distribution of self-evaluated skill. I think we can better gauge the effect of drawing simplicity on classification accuracy without this added element. Adding this leading question results in an element of uncertainty due to self-esteem variation: we can't be sure that the user indicated skill level is an accurate reflection of genuine skill level. In many cases, more skilled individuals will have a perception of lower skill ("Imposter syndrome" phenomenon). In any case, we cannot be sure that Mechanical Turk respondents would take the question of skill seriously even if posed.

Due to these constraints, it seems most appropriate to randomly assign artists to each treatment group and trust that individual level treatment effects will amortize out with sample size. In summary, we assume that randomization produces independence. We assume some relationship between artist skill (unobservable) and drawing complexity, but by randomizing assignment into various treatment groups, we assume that this will be negated when evaluating an average treatment effect.

Data processing will be held constant. Each of the images will be featurized (transformed into a set of machine-readable attributes) in an identical way. We will cross-validate the same algorithms and parameterizations in each of the treatment sets. Since we aim to capture variation in accuracy due to drawing time, the goal in this step is to reduce or eliminate variation in classification accuracy due to data or algorithm specific phenomena.

Outcomes:

The most robust way to quantitatively evaluate outcomes involves a single set of test data. We will train four models (each model will be trained and cross-validated exclusively on the data of a single treatment). We will evaluate accuracy on the test data-set and produce a confusion matrix of output versus actual for each treatment group. We care most about accuracy as well as common mistakes for each treatment group. The human labels attached to each sketch will also be used to increase robustness. For example, an incorrect classification by the algorithms that is labelled incorrectly by all artists has a different meaning than one that was labelled correctly by all humans. This is not exactly a covariate – but it is a guardrail metric that we will track throughout the experiment and use to contextualize our findings.

The goal is to establish a simple relationship between accuracy and time-to-draw. If we can do this and test the relationship for significance, we can infer how time to draw affects algorithmic accuracy. We can replicate this analysis for human labelling accuracy as well. Contrasting causality with respect to the ML accuracy with causality with respect to human labels should provide context and relativity.

I am neither clustering nor blocking in this experiment: we are randomly assigning treatments at the individual level. Mechanical Turk should allow us to reach individuals randomly and we thus avoid the constraints that would typically lead to a clustering approach. As discussed previously, asking the participants a leading question about art skill level could be used in a blocking approach. In that case, for every self-proclaimed high-skill artist in a treatment group, we would assign a low-skill artist.

Pilot & Risks:

We have an operating budget of $500 maximum. We will dedicate $100 to our pilot – offering $5 per task completion. We thus expect 20 responses (or 200 sketches) in our pilot. We will evaluate feasibility and ballpark effect size from these initial findings. Assuming that we offer $5 per task completion, we can have a maximum of 100 responses (1000 sketches and 990 human labels total). Using this maximum sample pool, we can understand the requisite effect size in prediction accuracy to establish significance. The pilot will establish if that is at all possible.

If there is any significant error with the conducted experiment, we hope to detect this during the pilot. This would reduce our effective sample size to 80 responses (800 sketches). While this would reduce the statistical significance of any findings gained, it serves as a layer of defense in case of any problems.

The primary risk is that the task will not be completed with effort. If artists given a longer period are not actually spending that time on the sketch, the effect will be predictably diminished. Covariance with artist skill is another risk that has been previously discussed.