CCM Final Project Proposal - Charlotte Ji, Dian Zhang, Fan Yang, Yi Wen

Problem statement and stakeholders

In this study, we will evaluate the humans' perception against convolutional models' perceptions of typicality on different categories of images. The evaluation results would help us understand how machine learning models' rate typicality of images differently from humans. The results will further help guide the possibility of using machine learning models to simulate human typicality tests for psychologists and/or social science researchers in the future.

Methods

Behavioral experiment

We plan to collect human typicality ratings through crowdsourcing platforms such as Amazon Mechanical Turk if our resources allow. Otherwise, we will conduct a smaller scale survey on fellow NYU students, but taking into consideration that the human ratings collected in this way will not be representative of the population's rating in the United States. For each category, we will generate a final rating by averaging over all participants' ratings for that category.

Convolutional Networks

We will begin with testing the three different convnet architectures used by Lake et al., including OverFeat, AlexNet and GoogLeNet. Our baseline model will be the same non-convnet model built upon N one-vs-all linear SVMs, where N is the number of categories we are testing for. After this stage, we plan to extend our model selection to ensemble OverFeat models and other convnets that participated in the ImageNet challenge such as VGG and CaffeNet.

We will follow Lake et al. to define machine typicality rating using category scores, under the assumption that typicality is related to the strength of a model's classification response to a category. We will measure both the raw category score and normalized class probabilities, which are respectively better at capturing the ideas of raw typicality and contrast effects. For the baseline model, we will use SVM probability scores to measure typicality.

Image Datasets

We intend to follow the procedure demonstrated in the paper of Lake et al.. In order to introduce more variety of reliable insights for the models, we plan to choose images of categories that: 1) we expect not to vary much among humans in terms of typicality 2) we expect machines to yield different typicality ratings from humans. Overall, the general idea is to choose 8-10 categories that appeared in ImageNet, 10-20 images each category from Google search, or using LSUN/ SUN database. So far, we plan to select from the following categories 1) scene (e.g. beach, mountain) 2) architecture (e.g. castle, church) 3) human crafted objects (e.g. cars, airplanes).

Evaluation and Further Analysis

We will measure rank correlations of human and machine typicality ratings over each category of interest and compare the size of these correlations with Lake et al.'s findings. We will also try to

interpret any discrepancy between convnet and human judgements. Then, we will perform the same procedure as in Lake et al. on the hidden layer activations of comparable models in our selection to see how correlations between human and convnet typicality ratings develop across layers. We plan to further examine the activations qualitatively by visualizing them using a Deep Visualization Tool developed by Yosinski et al..