Student: James Hahn

Research Advisor: Dr. Adriana Kovashka

Capstone: CS1950 Research

Topic: Measuring Effectiveness of Video Advertisements

Advertisements are unavoidable in modern society. Times Square is notorious for its incessant display of advertisements. Its popularity is worldwide and smaller cities possess miniature versions of the display, such as Pittsburgh and its digital works in Oakland on Forbes Avenue. Tokyo's Ginza district recently rose to popularity due to its upscale shops and constant onslaught of advertisements to pedestrians. Advertisements arise in other mediums as well. For example, they help popular streaming services, such as Spotify, Hulu, and Youtube TV gather significant streams of revenue to reduce the cost of monthly subscriptions for consumers. Ads provide an additional source of money for companies and entire industries to allocate resources toward alternative business motives or to increase profit for future competitive ventures. They are attractive to companies and nearly unavoidable for consumers. One challenge for advertisers is examining the advertisement's effectiveness or usefulness in conveying a message to their targeted demographics. If an advertisement is constructed, but an algorithm can predict with high accuracy whether it will be effective, and how effective in particular, businesses will save time and money. For example, a company creates an ad, but the algorithm predicts it will be ineffective, the business will have to spend more money to reimagine the ad, but potential revenue will increase since more consumers will enjoy the content after its recreation. In another scenario, a

business creates an advertisement and there exist quarrels among stakeholders whether a specific portion of the ad should be changed, but the algorithm says the advertisement will be highly effective. The company will save time since they know changing the ad will cost more money and time to reshape it, while either hurting or marginally increasing effectiveness. This challenge proves more significant in video advertisements. Rather than constructing a single, static image of content, a video advertisement possesses hundreds of frames of data with varying scenes, actors, objects, and complexity. Therefore, measuring effectiveness of video advertisements is important to impacting a billion-dollar industry across nearly any sector of industry.

Thankfully, Dr. Adriana Kovashka (Pitt) and several PhD students developed two datasets for advertisements for CVPR 2017 (IEEE Proceddings for Computer Vision and Pattern Recognition). The first dataset is 65k static image advertisements and the second dataset contains 3,477 video advertisements. The disparity between the two datasets increases the difficulty for the video effectiveness task. Today, machine learning and data science projects commonly utilize large datasets of hundreds of thousands or millions of samples. As such, advanced machine learning and computer vision concepts are required to extract as much data from the ads and increase the accuracy of predictions as possible. Not much consideration was given to the static advertisements, so the smaller video dataset was investigated further. The dataset is human-annotated across topic, sentiment, action/reason, whether it is funny, whether it is exciting, whether English is the primary language, and effectiveness. Funniness and excitement are binary labels. English-focus is a ternary label with survey options “Yes”,

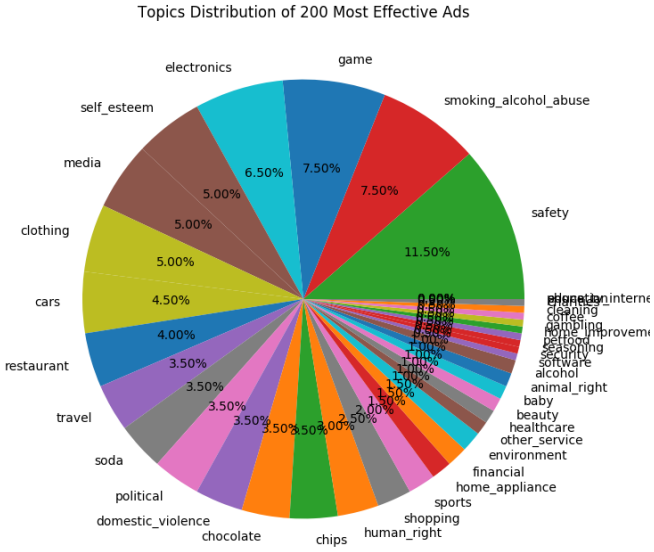
"No", and "Does not matter". On a more complex note, the action/reason feature combines the action of the consumer and reasoning behind the action. For example, in an car commercial, a sample action/reason statement is "I should buy this car because it is pet-friendly." The 38-dimensional topic label provides a general description of the ad's content, such as "cars and automobiles", "beauty", or "safety", while sentiments are 30-dimensional and examine evoked emotions, such as "cheerful", "amazed", or "angry". Finally, the output label, effectiveness, is a rating of 1, 2, 3, 4, or 5 and describes whether the given consumer is likely to purchase the product or whether the message was properly conveyed. Amazon Mechanical Turk assisted in finding enough human annotators to provide five labels for the features of each video. In addition, the duration of each video was extracted through use of the Youtube API.

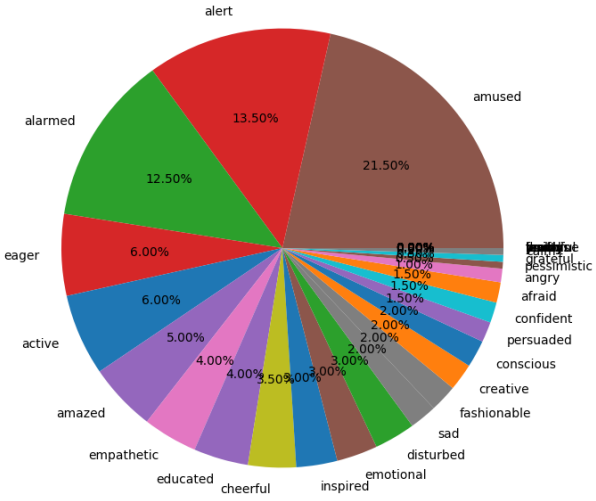
In addition to the features pre-packaged with the dataset, image processing proves handy in computer vision tasks with the ability to extract additional features from the video. For example, the average and median hue values of the middle 30% and 60% window of the video were computed in addition to optical flow vectors, number of shot boundaries, entropy of shot boundary, aggregated text statistics, and memorability. In general, extracting hues from the video details whether it utilizes more blues or reds or greys. A sad video typically utilizes blues and greys while an energetic, sports or entertainment video makes use of reds. The second aspect of the hue features is extracting data purely from the middle 30% or 60% of the video. This is useful because it creates an artificial gaze detection monitoring system. Users pay attention to the center of the video, but do bright colors in the exact center matter more than colors

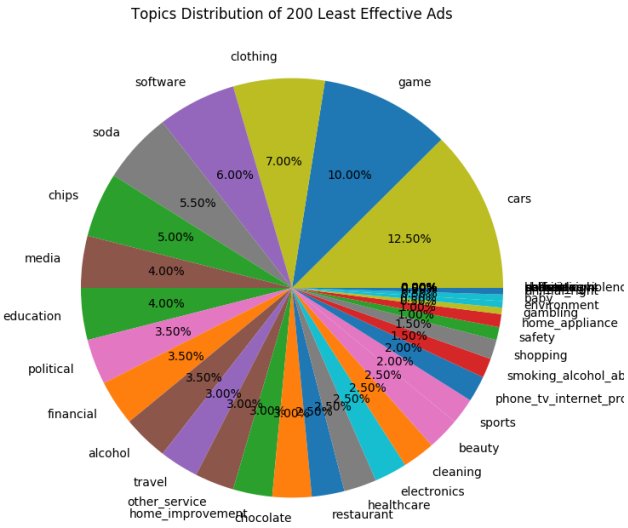
closer to the edge of the video? Optical flow measures where light travels in the video and how fast. For example, a car might drive 30mph to the right of the screen or it might drive 120mph to the top-left of the screen. Both direction and magnitude/quickness matter in this scenario as advertisers attempt to catch the attention of consumers. Calculating shot boundaries provides another metric for quickness of the video. It is easy to detect when a video changes scene or shot type, so calculating how many times this occurs throughout the video or how dense the changes occur by placing the counts in bins allow for easy analysis of quick progression of thoughts and ideas throughout the advertisement. A simple meta-metric of the shot boundary output is the entropy of the binarized data. This measures the overall intensity and volume of shot boundary changes. Next, optical character recognition, or ORC, assisted in delivering statistics about text in the video. Google Cloud Vision extracted the text from every half-second in the video. Then, data was aggregated across all frames. The number of meaningful word occurrences, total word count, average word length, and sentiment analysis provided future insights into the data contained within each advertisement. A meaningful word is classified as anything except for stop words (i.e. ‘the’, ‘a’, ‘at’, ‘on’, etc.) and non-vocabulary words (e.g. ‘foodalicious’). Before further analysis was performed, such as word counts, all extracted data were stemmed (i.e. ‘computing’, ‘computed’, and ‘computes’ become ‘compute’) to ensure proper data navigation; differences in verb tense are negligible. Finally, average memorability of the video was calculated on a frame-by-frame basis

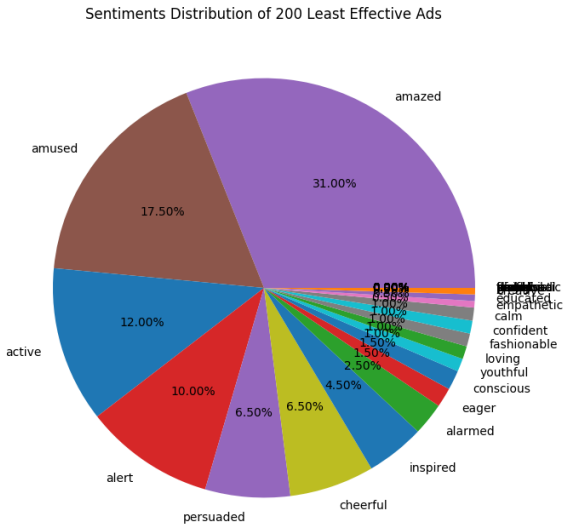
utilizing a classifier from CVPR 2011 by Isola, Xiao, Torralba, and Oliva: http://web.mit.edu/phillipi/www/publications/WhatMakesAnImageMemorable.pdf .

After some simple pre-processing of the videos, analysis of the dataset is a necessary pre-requisite in order to ensure the dataset is balanced and unbiased before training a model to classify samples. The simple most important balancing procedure of any machine learning task takes place with the output label. Therefore, after analyzing the number of advertisements for each of the five effectiveness labels, a resounding number of videos came out as effectiveness 3. The smallest class, effectiveness 1, contained 150 samples. Therefore, in order to prevent overfitting of any classification algorithm, 150 random samples were extracted from each class. Therefore, on a five-way classification, the dataset was reduced from 3477 videos to 750 videos. The small dataset becomes even smaller. Furthermore, the distribution of topics and sentiments









were examined across the 200 most and least effective ads, as displayed in the charts above, to grasp general guidelines of which type of advertisements are typically more or less effective.

After analysis, the training of a model is required. In modern computer vision and machine learning, neural networks are a popular algorithm to classify samples of data. However, 21 features are extracted from each video and complicated features, such as topic or sentiment, require one-hot vector encodings in order to prevent bias in the learning algorithm, essentially converting the problem into a 124-dimensional machine learning task. The Python library Scikit-Learn allows for easy training of support vector machines (SVMs). A neural network requires additional work to construct the network's architecture, which is an unimportant waste of time for this task. Therefore, for simplicity, SVMs were utilized. The aforementioned dataset was split with 80% (600 samples) for training and 20% for testing (150 samples), leaving roughly 30 samples for each class to test the performance of the classification. On a five-way classification,

with baseline accuracy of 20%, training a single classifier on all features proved inefficient, achieving accuracy around 25%, which is a negligible improvement. Therefore, with the use of ensemble learning, an SVM was trained on each individual feature with a radial basis kernel and one-vs-one classification. Each individual classifer ranged from 16-30% accuracy. These weak classifiers voted on the 'correct' effectiveness label of each sample to produce an overall accuracy of 45% on the five-way classification task – a significant improvement over the baseline. Additionally, with analysis of the effectiveness class 3, the distribution of topics, sentiments, and other features is uniform and sporadic. The classification of these samples is difficult due to ambiguity of a video being in the middle of effective and non-effective. Therefore, a four-way classification task was performed without effectiveness 3, achieving 61% accuracy (baseline = 25%). Finally, to determine whether an advertisement will be effective or non-effective in general, a final binary classification task was undergone, achieving 78% accuracy (baseline = 50%). Results provide novel benchmarks when predicting effectiveness of video advertisements.

After achieving aforementioned results, the final task is developing a paper. This semester, work has undergone by me to write a paper to help Dr. Kovashka's submission to a top-tier computer vision journal, such as PAMI (Pattern Analysis and Machine Intelligence) or IJCV (International Journal of Computer Vision). This paper will go further into the technical details of the above procedures while providing citations of related work.

This project required a simple technology stack typical of standard machine learning projects. Python3 was the primary language of development to program scripts. OpenCV allowed for easy extraction of frames from videos, as well as gathering hue data of each frame. Google Cloud Vision, natural language toolkit (nltk), and the Youtube API provided necessary data extraction to further gain insights into the dataset. Development of scripts was rather rudimentary, lacking much organization and structure in the project, but many tasks were modularized to ensure easier code maintainability. Finally, development was quick as research was required to be handled on a weekly basis, setting up quick deadlines on a constant basis often more stressful than real development jobs. Results were more important than cleanliness of code, which was maintained to an extent throughout the project.

Through the course of this project, I learned a significant amount of invaluable information. I was always interested in computer vision, but Dr. Kovashka kindly provided me the opportunity to enter into the computer vision research community. With her help, I took her graduate computer vision course and helped co-organize a workshop on understanding advertisements at CVPR 2018. She is the kind of professor to place students on fully independent projects to assist in the development of their research abilities, rather than placing them in a minor lab assistant role performing menial work. As such, I performed literature reviews, data analysis, met with her on a weekly basis for research meetings, trained the machine learning algorithms, and will finally construct a conference-level paper for my first submission.

Most of the work was done throughout the school year, so non-technical problems consisted of an additional stressful workload on top of constant projects and exams. Dr. Kovashka was a pleasure to work with, providing any necessary guidance and feedback to help me improve as a student, researcher, presenter, and person in general; I want to be like her one day. She is possibly one of my top 3 professional mentors I have ever had. No portion of the project was terrible or boring in particular, but I reached mental roadblocks several times throughout the process. For example, I had an idea for cue exploration, but when analysis was performed, it was a terrible indicator of effectiveness. Or we would extract a bunch of features, train the algorithm, slightly improve accuracy with mediocre results, and then run out of ideas for future avenues just to end up at square one.

This project proved valuable to my development, academically, since I eventually decided to begin a Bachelor's thesis. Once again, Dr. Kovashka is my research advisor since she is the only computer vision professor in the department and she always knows what to say to help me improve. I am thankful to have the opportunity to work on this project despite the technical and non-technical challenges, ambiguity at points, and difficulty and fast-paced results-focused environment, differing significantly from the typical development environment I am used to from my internships.