VISUAL-TEX: Video Tagging using Frame Captions

Dylan Moore | Nick Troccoli | Kat Gregory

Stanford University | CS229 | December 13, 2016

Abstract

To utilize the wealth of information stored in our ever increasing volume of video content, we must develop ways of searching and analyzing videos. A simple yet effective benchmark in this task will be the auto-categorization of video clips by tags that describe the video content, which could be useful for indexing uncaptioned video on sites like YouTube as well as for video retrieval and video surveillance. To explore this problem domain, we developed a system to label video clips from Hollywood movies with 12 action tags. We generated captions for a sampling of still frames from each clip and then compared the performance of 49 combinations of feature extractors on these captions and classifiers on these features. We achieve over 35% accuracy.

Introduction

Although there is a great deal of research on text classification, text-based approaches to video classification have only focused on text that is viewable in the video or provided via transcripts. This is of limited use because viewable text suffers from the high error rate of OCR and transcript text concerns primarily dialog. Can we utilize visual information in text format by leveraging the power of NeuralTalk, which creates captions for still images?

Problem Statement

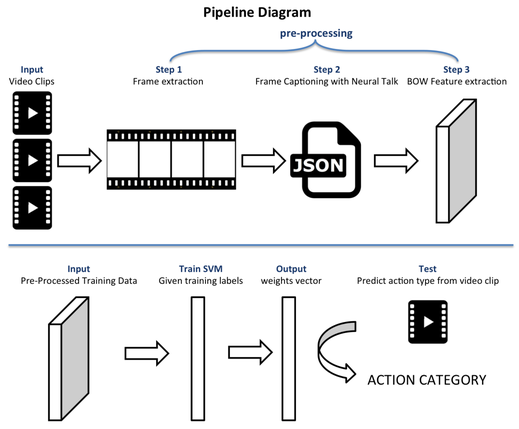
To label video clips with tags that describe their content by building a machine learning model that featurizes video clips as the NeuralTalk captions of sampled frames and distills these captions into a categorization tag for that video.

Datasest

We use the Hollywood2 Human Action Dataset developed by Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld, which contains 2516 clips from a total of 69 Hollywood movies, each clip labeled with one of twelve actions.

Approach

Given a video clip, we extract a sampling of 24 frames and generate a NeuralTalk caption for each. Then, we extract features from the captions for all sampled frames and use multinomial classification to select an action tag for the clip from twelve available actions. This approach is detailed in the diagram below.



System pipeline for a single feaure / classifier combination.

Processing and Captioning

We use Processing to sample 24 frames from a video clip. Each frame is then passed to NeuralTalk2, a pretrained recurrent neural net developed by Andrej Karpathy at Stanford, which describes the image with a sentence. The video clip is henceforth represented as the collection of these captions. This process is batch run via bash scripts.

Features

We then transform a clip's array of individual captions into a feature vector. We compared 6 NLP approaches to this task, including Bag of Words, Binarized Bag of Words, [2-4]-Grams and TF-IDF.

Classifiers

Similarly, we experiment with 9 classification algorithms, including Bayes, Decision Trees, Extra Trees, Gradient Descent, KNN, Logistic Regression, Multiclass, Random Forests, and SVM.

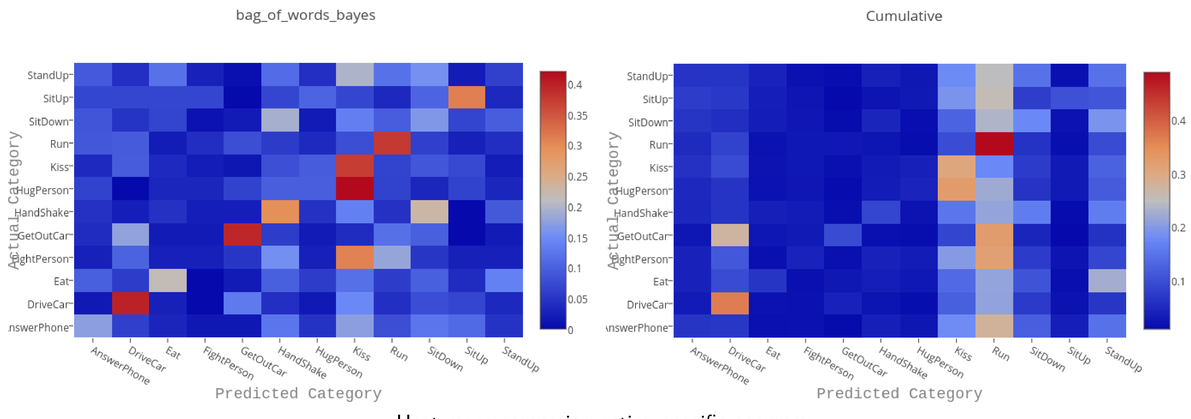
Comparison

	bow	2grams	3grams
bayes	29%	32%	33%
dtree	23%	23%	22%
extratrees	26%	31%	27%
gradient_de	27%	30%	24%
knn	25%	25%	25%
logistic_rgss	27%	26%	27%
multiclass	23%	23%	26%
rforest	35%	35%	34%
svm	32%	34%	35%
Overall	27%	29%	28%

	4grams	tfidf	Overall
bayes	32%	31%	31%
dtree	23%	25%	23%
extratrees	30%	29%	28%
gradient_de	27%	27%	27%
knn	25%	25%	26%
logistic_rgss	26%	34%	28%
multiclass	25%	30%	25%
rforest	32%	31%	33%
svm	33%	25%	32%
Overall	28%	28%	

Accuracy for each feature+classifier experiment. Random performance would be 8.3%. Binary bag of words not shown.

Analysis



Heat maps comparing action specific accuracy.

Of the models that we tested, SVM and random forest give the most accurate predictions, $\tilde{3}5\%$. The 2-grams and 3-grams featurizers are the most effective. As the heatmaps illustrate, the actions "Run," "Kiss," and "Drive Car," are most easily recognized. This intuitively makes sense when we consider that the entities recognized from these clips have very little overlap. e.g. any objects associated with outdoor landscape scenes appeared only in the running clips, which we suspect helped the models more effectively identify this category than categories that require temporal information, like "Stand Up," "Sit Down" or "Sit Up."

Conclusion

There are two significant implications of our results. The first is that phrase-based learning with captions, in our case an n-gram featurizer, produces promising results. The second is that we would expect a larger dataset to improve other classifiers; currently, however, random forest provides the best results as it does well with smaller datasets. With features that account for caption sequence, a larger training set, and other techniques such as entity extraction, we believe that our frame captioning approach would provide a promising way to categorize videos.

Future Work

Now that we have explored a classification task, we would like to try building a generative model that encodes temporal information - a NeuralTalk for video clips.