Automatic Advertisement Ratings

Brian Mora bmora@knights.ucf.edu
Dong Zhang dzhang@cs.ucf.edu
Amir Mazaheri amirmazaheri@knights.ucf.edu

1 ABSTRACT

The goal of this project is to explore a system that automatically generates objective scores, or rankings, for both video and image advertisements. For the image advertisements, the use of Convolutional Neural Networks (CNNs) is explored and compared to baseline methods such as, hand-crafted features and Support Vector Machines (SVMs). For the video advertisements, a different approach is taken. Annotated attributes and mid-level visual concept representations are employed to understand how a specific concept affects an advertisement. This is achieved through the use of Ranking SVM. Additionally, an approach similar to that of the image advertisements SVMs is applied.

2 INTRODUCTION

The advertising industry is massive, and surveying humans is very costly, and time consuming. According to Forbes, advertising is nearly $200 billion is spent on advertising in the U.S. alone. There are many stages to advertisement development. Advertising research, or ad testing, involves human surveying at various stages. These surveys and questionnaires are expensive, time consuming, and can be dependent on the structure of the survey, or questionnaire, being administered.

3 APPROACH

3.1 Images

3.1.1 Dataset

To begin, this project required the gathering of a dataset. The dataset first consisted of 500 image advertisements from 5 different product types: car, cosmetics, clothes, food, and technology. These image advertisements were obtained from Google Image search results. Annotations needed to be obtained for these image advertisements. For the annotations, five participants were surveyed. Each of these participants was asked to rate each advertisement from one (the lowest) to ten (the highest) on how much they liked the image advertisement. Each participant was given a break after every 100 image advertisements to avoid fatigue and boredom. The average of these five separate ratings became the annotations for the image advertisements.



Figure 1: Several of the image advertisements collected

Due to the small number of image advertisements, the training image advertisements were augmented. For this data augmentation the image advertisements were flipped and cropped, and these 400 training images were augmented to 958 training images.

3.1.2 Baseline Methods

The baseline method used to predict the annotation for these image was Support Vector Regression (SVR). The SVR was trained on several different extracted feature vectors from the collected image advertisements along with their corresponding annotations. Linear, Polynomial, and Radial Basis Function (RBF) kernels were used. C values of 0.1, 1, 10 and epsilon values of 0.01, 0.1, 1.0, 10 were used.

The feature vectors used as input for the SVR were: the highest level features of the pre-trained VGG-16 [1], the highest level features of pre-trained AlexNet [2], and Gist features.

The results of SVR are reported using Root Mean Square Error: $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{1,i}, x_{2,i})^2}{n}}$

Due to the high dimensionality of the feature vectors selected (1x4096, 1x4096, 1x960, respectively), and the small size of the training set (initially, 400 images were used for training), dimensionality reduction and data augmentation were employed. Principal Component Analysis (PCA) was used to reduce the dimensions of all the selected feature vectors to 1x20. The reduced dimension feature vectors of the augmented 958 training images were used for SVR.

Overfitting was occurring in the models, so the 500 image advertisement dataset was doubled to 100 image advertisements. The additional 500 image advertisements were obtained in the same manner outlined in the *Dataset* section above.

The same baseline methods were applied, but now instead of 400 training image advertisements, there were 800 image advertisements.

3.1.3 Experiments

Experiments could begin once the image advertisements, and corresponding annotations, were obtained. The first experiment was modifying, and fine-tuning, AlexNet. The last fully connected layer was changed to output a single value, and the Softmax loss layer was changed to a Euclidean Loss layer. With these two changes AlexNet was ready for regression. To fine-tune AlexNet, weights were initialized in Caffe [3] with the pre-trained weights, and using the image advertisements and Stochastic Gradient Descent, fine-tuned the last layer. Several learning rates were tried throughout this fine-tuning process.

3.2 Videos

3.2.1 Dataset

A video dataset was collected. This dataset consists of 276 commercials from 9 brands: Geico, Heineken, Mazda, McDonald's, T-Mobile, Toyota, Burger King, Budweiser, Coca-Cola. These commercials range in length from 30 seconds to 2 minutes. The videos were obtained from YouTube. Along with each video, YouTube statistics were also taken for each video. These statistics include: date of upload, views, likes, dislikes, number of comments. The last component of the dataset was annotated attributes. Two lists of attributes were annotated for each video which will be called List 1, and List 2, respectively. List 1 consists of 28 audio and visual attributes such as: pet, indoors, voice-over, day time, country music. List 2 consists of 31 visual attributes. This list contains attributes coming primarily from TRECVID [4], along with some refinements of attributes from List 1. Some attributes in List 2 are: Inside car, office, Slow Motion, animation.

3.2.2 Experiments

Once the video dataset was collected, the first method used was a Ranking SVM [5]. The ranking metrics used were: $\frac{likes}{view}$ and $\frac{view}{\log(1+n)}$. The latter being time normalized views. C values (training error penalization) of 0.00001, 0.0001, 0.001. 0.01, 0.1, 1, 10, 100, 1000, 10000 were used. The Ranking SVM learns weights, and these weights are used to sort the testing videos.

The results of the ranking SVM experiments are reporting using a pairwise error. For the ranking SVM generated sorted list, each pair is evaluated to either be correct or incorrect. If a pair is incorrect the error is incremented. This error is then divided by the number of pairs.

Several experiments were performed using this ranking SVM. The first being a single ranking SVM for all of the brands at once. Meaning that all of the training videos were used to train a single ranking SVM. This experiment was repeated using List 1, List 2 as the input feature vectors separately.

The next experiment was a separate ranking SVM for each brand, for a total of 9 ranking SVMs. Meaning that only the training

videos of a particular brand were used to train a particular SVM. This experiment was also repeated using List 1 and List 2 as the input feature vectors separately.

The final ranking SVM experiment was taking the learned weights from the single SVM for all the brands together, and sorting each brand's testing videos with these weights separately. This experiment was also repeated using List 1 and List 2 as the input feature vectors separately.

Once these experiments were performed on the attribute annotations as feature vectors, several other features were explored.

VGG on pre-trained weights (1x4096) high level features were extracted. Along with this, TRECVID concept detectors (1x60) were used. The ranking SVM experiments outlined above were repeated with these two features as the input feature vectors separately.

To obtain these features (VGG and TRECVID), frames were first extracted, and the features were calculated for every frame in each video. Before input into the ranking SVM, max pooling was done, so that the highest values within a video for a particular dimension, or concept, was used as input for the ranking SVM.

SVR experiments were also performed. These are very similar to the SVR experiments outlined in the *Images, Baseline Methods* section. The differences being, the SVR is predicting $\frac{likes}{view}$, and the feature vectors used as input were: List 1, List 2, VGG, and TRECVID. Similar to the ranking SVM experiments, a single SVR was first used for all the brands together. A separate SVR was also used for each brand.

4 RESULTS

Learning Rate	RMSE
0.0001	1.47066518527
0.0005	1.00233565894
0.0006	**0.995147059459**

Table 1: Reporting RMSE results of fine-tuning and regression-modified AlexNet.

	Original	Augment	PCA and Aug	PCA New
AlexNet	**0.8193**	0.82279	0.7490	**0.8378**
VGG	0.8193	0.8233	**0.7078**	0.8435
Gist	0.8220	**0.8208**	0.7430	0.8417

Table 2: Reporting RMSE results of SVR for several features (image advertisements). *Original* being set of 500 image advertisements; *Augment* being set of 500 image advertisements plus augmented; *PCA and Aug* being same images and augmented images from *Augment* except PCA is applied; *PCA New* being *Original* images plus 500 new image advertisements.

	List 1	List 2	VGG	TREC	Random
G	0.4929	0.4922	0.4	**0.3**	0.4929
H	0.4922	0.5303	**0.4**	0.7	0.4922
M	0.4984	0.5011	**0.3684**	0.4315	0.4984
Mc	0.5002	0.5007	0.3809	**0.3263**	0.5002
Tm	0.4964	0.5035	**0.4912**	0.6140	0.4964
T	0.5020	0.5051	0.6023	**0.4678**	0.5020
Bk	0.5050	0.5051	**0.4248**	0.5032	0.5050
B	0.4904	0.4962	**0.3296**	0.3956	0.4904
CC	0.4947	0.4942	**0.4**	0.4945	0.4947
Avg	0.4969	0.5023	**0.4219**	0.4703	0.4969

Table 3: Reporting error results of separate ranking SVM for each brand (video advertisements). Brands from top to bottom, Geico, Heineken, Mazda, McDonalds, T-Mobile, Toyota, Burger King, Coca Cola. The values in blue indicate the best value obtained for each brand.

	List 1	List 2	VGG	TREC	Random
Error	**0.4513**	0.5314	0.4795	0.5213	0.4980

Table 4: Reporting error results of single ranking SVM for all brands together (video advertisements).

	List 1	List 2	VGG	TREC
G	0.002048	0.253938	**0.001906**	0.002686
H	**0.002034**	0.002425	0.003082	0.003027
M	0.009040	0.232206	**0.001835**	0.003574
Mc	0.006410	0.099953	0.003330	**0.002765**
Tm	0.007739	0.291330	**0.002483**	0.002765
T	0.006449	0.219350	0.001983	**0.001633**
Bk	0.045220	0.192223	**0.007912**	0.014047
B	0.003924	0.290131	0.003636	**0.002534**
CC	0.005769	0.293265	0.006231	**0.003700**
Avg	0.009849	0.208314	**0.0036**	0.004082

Table 5: Reporting RMSE results of separate SVR for each brand (video advertisements). Brands from top to bottom, Geico, Heineken, Mazda, McDonalds, T-Mobile, Toyota, Burger King, Budweiser Coca-Cola. The values in blue indicate the best value obtained for each brand.

	List 1	List 2	VGG	TREC
Error	**0.00322**	0.099973	0.004442	0.006017

Table 6: Reporting RMSE results of single SVR for all brands together (video advertisements).

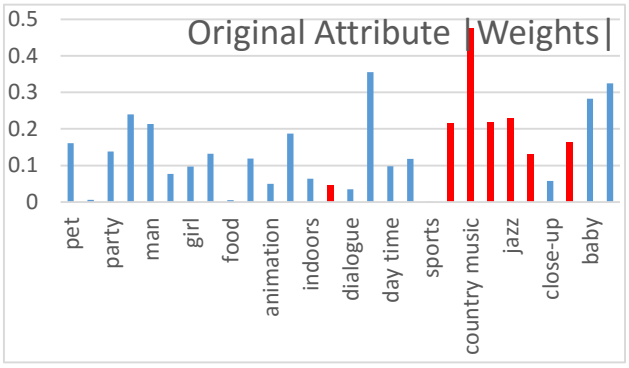


Figure 2: Absolute values of weights learned by ranking SVM for List 1.

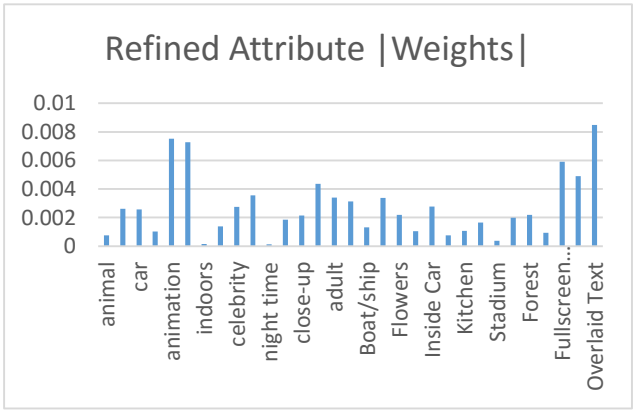


Figure 3: Absolute values of weights learned by ranking SVM for List 2.

5 CONCLUSION

The fine-tuned AlexNet results were promising. The RMSE was good, but when compared to the baseline methods, fine-tuned AlexNet performed worse. A problem became apparent: a good RMSE did not mean that the results were discriminative. None of the models were discriminative. The datasets themselves also posed problems. The 1000 image advertisements and the 276 video advertisements still proved too small for our models because significant overfitting was still occurring.

With the video advertisements ranking SVM experiments, it was observed that the weights learned for some of the audio attributes in List 1, such as *country music*, were several orders of magnitudes greater than the purely visual attributes from List 2. It can be inferred that audio features are as, if not more, important than visual features within a video advertisement.

The best reported RMSE for the ranking SVM experiments on the video advertisements showed a 15% improvement over randomly ranking the advertisements. This 15% improvement was obtained with the separate ranking SVM for each brand and was significantly higher than the results of a single SVM for all of the brands together. It can be inferred the video advertisements across brands is significantly different and that they cannot effectively be grouped across brands.

The most significant challenge with this problem was the amount data. The datasets being worked with are very small. Perhaps with more image and video advertisements, these methods would work better.

In future work we plan to expand the datasets significantly. In addition to this we plan to explore other mid-level features, even though raw features provided the best results. This is because we wish to be able to say *why* an advertisement is doing well, and this can be achieved through mid-level features like the TRECVID concepts. Finally, we plan to explore the audio aspects of the video advertisements.

6 REFERENCES

[1] Simonyan, Karen, and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition.” arXiv preprint arXiv:1409.1556 (2014).

[2] Krizhevsky, Alex, Ilya Sutskeyer, and Geoffrey E. Hinton. “Imagenet classification with deep convolutional neural networks.” In Advances in neural information processing systems, pp. 1097-11.05. 2012.

[3] Jia, Yangqing, and Shelhamer, Evan and Donahue, Jeff and Karaye, sergey and Long, Jonathan and Girschick, Ross and Guadarrama, Sergio and Darrell, Trevor. “Caffe: Convolutional Architecture for Fast Feature Embedding.” arXiv preprint arXiv:1408.5093 (2014)

[4] Paul Over, and George Awad and Martial Michel and Jonathan Fiscus and Wessel Kraaij and Alan F. Smeaton and Georges Queenot and Roeland Ordelman. “TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics.” In Proceedings of TRECVID 2015 (2015).

[5] Olivier Chappelle. “Training a Support Vector Machine in the Primal. In Neural Computation 19.5 (2007): 1155-1178.

Images used:

http://sellingwithinsight.com/wp-content/uploads/2016/04/car-magazine-advertisement-astonmartinadcopy.jpg

https://s-media-cache-ak0.pinimg.com/236x/07/72/ca/0772ca4044c8f30959009bf9bbeb5680.jpg

http://designmodo.com/wp-content/uploads/2011/01/Tech-Advertisements-21.jpg