Deep Convolutional Networks for Modeling Image Virality

Abhimanyu Dubey EE2110061@IITD.AC.IN
Department of Computer Science and Engineering, Indian Institute of Technology Delhi, New Delhi 110016.

Sumeet Agarwal SUMEET@IITD.AC.IN
Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi 110016.

Abstract

Study of virality and information diffusion is a topic gaining traction rapidly in the computational social sciences. Computer vision and social network analysis research has also focused on understanding the impact of content and information diffusion in making content viral. We present a novel algorithm to model image virality on online networks using the increasingly popular deep convolutional neural network architectures. Our proposed model provides significant insights into the features that are responsible for promoting virality and surpass the existing state-of-the-art by a 10% relative improvement in prediction.

1. Introduction

The study of virality has been slowly gaining traction in the domain of computational social science research. Owing to the increasing prominence of online advertising, understanding and predicting what content becomes viral on the Internet is an important, with applications ranging from intelligent content organization on the Internet (Jain et al., 2014) to Twitter trend analysis (Petrovic et al., 2011).

Apart from online marketing, the impact of several other domains of active Internet participation depends on content virality. The reach of professionals, organizations, social causes and non-profits spitballs exponentially once viral content is associated with the same. Hence, as described previously in Deza and Parikh's (Deza & Parikh, 2015) novel introductory study of image virality, content virality has been studied extensively in the domain of marketing research (Berger & Schwartz, 2011; Berger, 2013).

The computer vision community has seen a surge in the usage of deep learning for end-to-end learning for computer vision, from image classification (Krizhevsky et al., 2012), semantic segmentation (Chen et al., 2014) and even image captioning (Karpathy & Fei-Fei, 2015), and there has also been work on abstract ontological tasks, such as prediction of attributes (Parikh & Grauman, 2011), humor (Chandrasekaran et al., 2015), image memorability (Isola et al., 2011) and street image safety (Naik et al., 2014).

Deza and Parikh's work is an important stepping stone to understanding the nature of content virality. (Lakkaraju et al., 2013) describe the temporal relationships of image virality in mode detail, along with several other streams of research (Jain et al., 2014; Goel et al., 2015) discussing the nature of the underlying structure of diffusion present in viral content.

This posits the obvious question of the relative importance of the content matter of a viral image, and if it is content alone that can govern the extent of virality an image gains online. Deza and Parikh perform an extensive study of the same, using handcrafted computer vision techniques - identifying that it is possible, with a certain degree of accuracy, to predict the virality of an image based on the image content alone. We aim to, with this study, bridge two streams of research from computer vision (attribute learning and deep learning) and computational social science by constructing an end-to-end learning system for predicting image virality.

2. Virality Prediction

2.1. Quantifying Virality

The first question encountered in the study of attribute learning is the quantification of attributes. Previous

Appearing in Proceedings of the 2^{nd} Indian Workshop on Machine Learning, IIT Kanpur, India, 2016.

studies on attributes (Parikh & Grauman, 2011; Staiano et al., 2013) have observed that obtaining a relative label gives better prediction accuracy. Another study on rating data collection (Mojica Ruiz et al.) reveals a bimodal nature of ratings as well, where having a relative label instead of an absolute metric is less prone to label noise. In attribute learning through vision, we find a similar prediction pipeline (Parikh & Grauman, 2011), where pairwise comparisons are available and an ordinal ranking is constructed from the comparisons.

Image Virality Dataset We preserve the exact dataset provided by (Deza & Parikh, 2015), which was originally sampled from (Lakkaraju et al., 2013). However, in addition to the training pairs proposed by (Deza & Parikh, 2015), we add additional data by constructing random pairs of viral and non-viral images, distinct from the existing pairs in the test set. We randomly select 10M of the potential 25M pairs (compared to the 10,078 images in the original study). There are three pairwise splits - complete data prediction, random splits and Top/Bottom 250. For more details, we refer the reader to (Deza & Parikh, 2015).

Image Popularity Dataset This dataset is the data utilized by Khosla et al. (Khosla et al., 2014) for their popularity analysis. It consists of 2.3M images sampled from Flickr and labeled as 'popular' and 'not-popular' according to their upvote measure. The three sub-categories for construction of the dataset are 1-per-user,user-mix and user-specific. We refer the reader to (Khosla et al., 2014) for more information.

2.2. Problem Formulation

Based on the nature of the dataset, we can formulate the problem as a pairwise classification problem. At each instance in training, our model will receive two input images, and the model will have to learn to predict the image with the stronger attribute present. Having obtained the predictions of the network, we can construct an ordinal ranking of the images, and denote the top k as having the attribute present.

We have a set S of N images (obtained from Lakkaraju et al (Lakkaraju et al., 2013)), of which a subset S_v of N_v images are classified as viral, based on the virality metric defined by Deza et al. (Deza & Parikh, 2015). The model is fed a randomly generated ordered pair of images (I_1, I_2) from S - one from $S \setminus S_v$, the other from S_v. Hence, we can generate a total of $\mathcal{O}(N_v(N - N_v))$ distinct ordered image pairs of which we select d pairs to form set D, which is our dataset, which we split later into $D_{train}, D_{val}, D_{test}$ i.e. training, validation and test sets respectively, based on the existing splits of (Deza & Parikh, 2015). The output variable y in each image pair is the viral image index - $+1$ if I_1 is viral, and -1 if I_2 is viral.

2.3. Pseudo-Siamese Networks

We construct a convolutional neural network architecture to learn an attribute regressor by taking an input as a pair of images, and label as the winning image. The basic structure of our convolutional neural network involves two disjoint Siamese networks which share weights and are later combined to a fully-connected layer and trained discriminatively, following (Chopra et al., 2005). We take existing image classification architectures (**AlexNet** (Krizhevsky et al., 2012) and **VGG-Net-19** (Simonyan & Zisserman, 2014)), and discard the final decision boundary layer and fine-tune two such disjoint networks from their image classification weights. For newer layers, we randomly initialize weights following (Krizhevsky et al., 2012) and construct the final decision boundary.

2.4. Ranking Loss

Unlike (Chopra et al., 2005), we do not wish to learn a similarity metric, and wish to minimize our ranking loss. Hence, we formulate a loss function given by

$$E_p = \sum_{(I_1,I_2,y)\in D_{batch}} E_c + \lambda E_r \quad (1)$$

$$E_c = \max\left(0, y \cdot (g_r(I_2) - g_r(I_1))\right)^2 \quad (2)$$

$$E_r = \frac{1}{(f_r(I_2) - f_r(I_1))^2} \quad (3)$$

Here, the function g_r is simply the softmax of the outputs.

$$g_r(I_i) = \frac{e^{f_r(I_i)}}{e^{f_r(I_1)} + e^{f_r(I_2)}} \quad i \in \{1, 2\} \quad (4)$$

E_c minimizes the direct ranking error, and the softmax on the output neurons enforces the outputs of the network to be binary. The second term in the loss function can be thought of as a regularizer on the distribution of f_r learnt, and it enforces $(f_r(I_2) - f_r(I_1))^2$ to be as large as possible for each input pair. However, the weighing term λ must be kept small to prevent oscillations during training.

This architecture is referred to as the **PVCNN**, that is, the Pairwise Virality CNN in the experimental sections. For networks initialized with the AlexNet architecture, the results are indicated by **AlexNet-PVCNN**, and similarly for networks initialized with the VGGNet-19 architecture, the results are indicated by **VGGNet-19-PVCNN**. As mentioned in recent deep learning literature, we also employ standard

L2 regularization (weight-decay) and momentum for training our networks with stochastic gradient descent.

2.5. Feature Augmentation

To provide additional contextual information, we modify the **PVCNN** architecture introduced in the previous section with additional semantic information available from the dataset. We fine-tune an image classification network, with initial weights and architectures from (Simonyan & Zisserman, 2014; Krizhevsky et al., 2012) with class labels as the topic IDs of submitted images (this network is referred to as Topic CNN). Post training, we discard the decision boundary, and feed the penultimate layer weights as additional features to the fully-connected layer in **PVCNN**. This architecture is known as **TPVCNN** (Topic-PVCNN) henceforth (see Figure 1 for further details). We leverage topic features to supply additional relevant information.

3. Experiments

Our experiments were carried out in Caffe (Jia et al., 2014) with the Python framework. All experiments were run with NVIDIA TITAN X GPUs. The model weights mentioned for AlexNet and VGG-Net architectures were obtained from the online Caffe Model-Zoo. For training the neural networks, the ranking networks were trained with an initial learning rate of 0.001, and momentum was fixed as 0.9 and initial weights were sampled from a gaussian distribution with mean 0 and sigma 0.05 (following (Krizhevsky et al., 2012)). In the ranking loss, regularization (L2) was set with weight 0.05 and λ was set at 0.001 for the optimal results.The learning rate was decreased once the network began oscillating and the learning rate was decreased a total of 3 times before convergence was achieved.

4. Analysis and Conclusion

We see that our deep networks outperform the state-of-the-art (68.10% on Random Splits) comfortably on both the deeper networks on the Image Virality Dataset. The feature augmentation which leverages category information into the prediction also increases the prediction accuracy, which leads us to confirm the earlier hypothesis (Deza & Parikh, 2015) that some image categories are more likely to be viral than others. On the popularity dataset, our performance is competitive to the state-of-the-art, and we attribute this to the bimodal distribution of virality scores compared to a smoother distribution in the popularity dataset. A detailed examination of the deep features will be available in the full version of the paper.

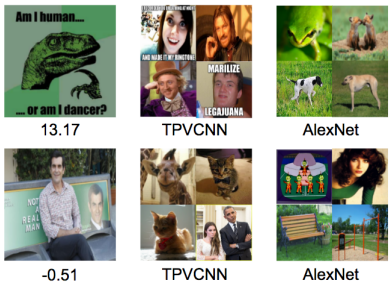

Figure 2. 4 Nearest Neighbours for two sample inputs in the space of pre-final layer activations. The first image is a sample with a high virality score (13.17) and the second image is a sample with a low virality score (-0.51).

We would like to conclude by summarizing our contributions - creating a novel pairwise architecture for abstract attribute prediction, which we aim to generalize to other abstract computer vision tasks as well, such as predicting relative attributes (Parikh & Grauman, 2011), memorability (Isola et al., 2011) and safety (Naik et al., 2014).

References

Berger, Jonah. *Contagious: Why things catch on*. Simon and Schuster, 2013.

Berger, Jonah and Schwartz, Eric M. What drives immediate and ongoing word of mouth? *Journal of Marketing Research*, 48(5):869–880, 2011.

Chandrasekaran, Arjun, Vijayakumar, Ashwin K., Antol, Stanislaw, Bansal, Mohit, Batra, Dhruv, Zitnick, C. Lawrence, and Parikh, Devi. We are humor beings: Understanding and predicting visual humor. *CoRR*, abs/1512.04407, 2015. URL http://arxiv.org/abs/1512.04407.

Chen, Liang-Chieh, Papandreou, George, Kokkinos, Iasonas, Murphy, Kevin, and Yuille, Alan L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014. URL http://arxiv.org/abs/1412.7062.

Chopra, Sumit, Hadsell, Raia, and LeCun, Yann. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pp. 539–546. IEEE, 2005.

Deza, Arturo and Parikh, Devi. Understanding image virality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1818–1826, 2015.

Goel, Sharad, Anderson, Ashton, Hofman, Jake, and Watts, Duncan J. The structural virality of online diffusion. *Management Science*, 62(1):180–196, 2015.

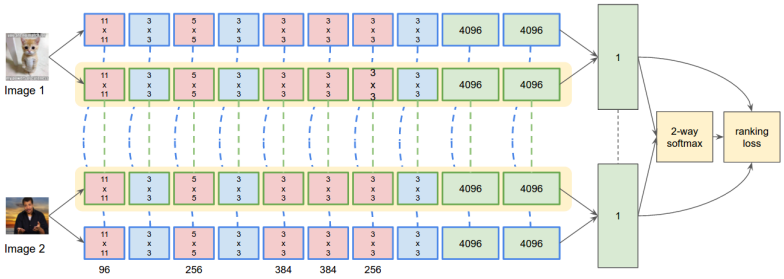

Figure 1. Architecture for **TPVCNN** using the AlexNet hierarchy.Yellow - Loss/Function, Red - Convolution, Blue - Max-Pooling, Green - Fully Connected. The layer blocks with a blue outline imply fine-tuning from AlexNet LSRVC weights, green outline imply weights from Topic-CNN. The dashed lines represent the layers which have identical weights. Yellow background shade represents fixed weights (no training). The different shades of dashed lines imply weight-sharing from two different networks. The **PVCNN** architecture does not have the two yellow fixed networks.

Image Virality Dataset (Deza & Parikh, 2015)			
Algorithm	**Complete Data**	**Top/Bottom Split**	**Random Split**
SVM + Image Features	53.40%	61.60%	58.49%
Human	-	71.76%	60.12%
SVM + Deep Attributes-5	-	-	68.10%
AlexNet - PVCNN	60.11%	69.97%	63.35%
AlexNet - TPVCNN	63.21%	72.48%	66.84%
VGGNet-19 - PVCNN	64.47%	72.25%	71.03%
VGGNet-19 - TPVCNN	**65.28%**	**75.88%**	**75.19%**
Popularity Dataset (Khosla et al., 2014)			
Algorithm	**1-per-user**	**User-mix**	**User-specific**
Deep Learning Features (DeCAF)	28%	33%	26%
Combined Features (GIST,Object,Color)	31%	36%	**40%**
AlexNet - PVCNN	27.78%	32.91%	29.55%
AlexNet - TPVCNN	30.56%	36.86%	33.75%
VGGNet-19 - PVCNN	29.92%	35.64%	34.81%
VGGNet-19 - TPVCNN	**31.57%**	**38.21%**	38.85%

Table 1. Table summarizing our empirical results on the Viral Images and Popularity Datasets. Scores reported are percentage accuracies, and all baselines have been reported from (Deza & Parikh, 2015) and (Khosla et al., 2014).

Isola, Phillip, Parikh, Devi, Torralba, Antonio, and Oliva, Aude. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*, pp. 2429–2437, 2011.

Jain, Puneet, Manweiler, Justin, Acharya, Arup, and Choudhury, Romit Roy. Scalable social analytics for live viral event prediction. 2014.

Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, Jonathan, Girshick, Ross, Guadarrama, Sergio, and Darrell, Trevor. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pp. 675–678. ACM, 2014.

Karpathy, Andrej and Fei-Fei, Li. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

Khosla, Aditya, Das Sarma, Atish, and Hamid, Raffay. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web*, pp. 867–876. ACM, 2014.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Lakkaraju, Himabindu, McAuley, Julian J, and Leskovec, Jure. What's in a name? understanding the interplay between titles, content, and communities in social media. *ICWSM*, 1(2):3, 2013.

Mojica Ruiz, I, Nagappan, Meiyappan, Adams, Bram, Berger, Thorsten, Dienst, Steffen, and Hassan, Ahmed. An examination of the current rating system used in mobile app stores.

Naik, Nikhil, Philipoom, Jade, Raskar, Ramesh, and Hidalgo, César. Streetscore–predicting the perceived safety of one million streetscapes. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pp. 793–799. IEEE, 2014.

Parikh, Devi and Grauman, Kristen. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 503–510. IEEE, 2011.

Petrovic, Sasa, Osborne, Miles, and Lavrenko, Victor. Rt to win! predicting message propagation in twitter. 2011.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Staiano, Jacopo, Albanese, Davide, et al. Exploring image virality in google plus. In *Social Computing (SocialCom), 2013 International Conference on*, pp. 671–678. IEEE, 2013.