Modeling Image Virality with Pairwise Spatial Transformer Networks

Abhimanyu Dubey
Massachusetts Institute of Technology
dubeya@mit.edu

Sumeet Agarwal
Indian Institute of Technology Delhi
sumeet@iitd.ac.in

ABSTRACT

The study of virality and information diffusion is a topic gaining traction rapidly in the computational social sciences. Computer vision and social network analysis research have also focused on understanding the impact of content and information diffusion in making content viral, with prior approaches not performing significantly well as other traditional classification tasks. In this paper, we present a novel pairwise reformulation of the virality prediction problem as an attribute prediction task and develop a novel algorithm to model image virality on online media using a pairwise neural network. Our model provides significant insights into the features that are responsible for promoting virality and surpasses the existing state-of-the-art by a 12% average improvement in prediction. We also investigate the effect of external category supervision on relative attribute prediction and observe an increase in prediction accuracy for the same across several attribute learning datasets.

CCS CONCEPTS

• **Human-centered computing** → **Social content sharing**; **Social media**; • **Computing methodologies** → **Computer vision tasks**;

KEYWORDS

deep learning for the web, convolutional neural networks, image virality, image attributes

1 INTRODUCTION

Online advertising has changed form persistently since the dawn of easily available Internet connectivity, beginning from blunt header and footer advertisements to carefully curated and unidentifiable advertisements fused closely with content. A call-and-response mechanism has started to emerge in the propagation of content on the Internet - content aggregator websites such as BuzzFeed [14], Gawker [22] and Funny or Die [40] immediately capture the *viral* content emerging on websites such as Reddit [44] and a subsequent marketing cascade follows that builds on the *virality* of content.

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 [19] novel introductory study of image virality, content virality has been studied extensively in the domain of marketing research [7–11].

It is important to note the subtle difference between viral content and content that is otherwise generally popular on the Internet. Earlier work by Khosla et. al. [32] aims at understanding the visual cues that make an image *popular* on the Internet. However, it is understandable that content that is popular is not necessarily viral, and dissemination networks are significantly different for the two classes of content. Deza and Parikh's work is an important stepping stone to understanding the nature of content virality, and Lakkaraju et. al. [36] describe the temporal relationships of image virality in mode detail, along with several other streams of research [24, 29] that discuss 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.

Across the multimedia and Internet media research communities, we have seen a surge in the usage of deep learning for end-to-end learning of complicated tasks. However, all these tasks have utilized computational models that concrete ontological categories, such as image classification [35], region proposals and semantic segmentation [23], image captioning [37] and visual question answering [4]. Extending traditional computational models to tasks that involve amorphous, abstract ontologies - such as image virality and popularity (that have large inter-user disagreements in label data), we require a rethinking of the problem approach, as done by prior work in image memorability [27, 33, 34], interestingness [25, 50] and urban perception [21, 38].

One of our most important contributions through this work is to demonstrate the effectiveness of a pairwise attribute approach in tasks that involve an amorphous ontology (for example, image popularity, beauty, memorability). We demonstrate a hierarchical novel approach, that utilizes transfer learning from traditional concrete domains, and provide substantial empirical improvements in prediction. Our approach is generalizable across abstract categories, and also outperforms prior work in attribute learning as well.

MM'17, October 23–27, 2017, Mountain View, CA, USA.

ISBN 978-1-4503-4906-2/17/10…$15.00
DOI: https://doi.org/10.1145/3123266.3123333

2 RELATED WORK

2.1 Predicting Content Popularity

A variety of different disciplines from computational social science to computer vision have approached the problem of predicting content popularity on the Internet. In the domains of machine learning and computer vision, most of the work currently present has been in the domain of Twitter or video - predicting the popularity of tweets [26, 42] and videos [39, 43, 46]. However, there is a growing body of work on predicting image popularity and virality [19, 32], with the work of Khosla et al [32], which focuses on introducing the problem of predicting image popularity, and then presenting a straightforward formulation with image features applied to a classifier. Deza and Parikh [19] introduce the different problem of predicting image virality with a novel metric for identifying virality, and demonstrate the effectiveness of several techniques for solving the problem using deep image attributes. Our work uses the datasets introduced by these studies and surpasses their proposed solutions by a large average improvement.

Social Media research has also looked at predicting content virality - focusing on the underlying structure of information dissemination in a social network and the cascading effect that aggregator websites have. They explain the nature of virality as a network phenomenon, and not entirely attributable to the inherent content itself [5, 29, 51]. Jain et al. [29] use an intricate system to identify viral videos in real-time. They accomplish this by analyzing a metric based on the number of re-shares a particular video has on Twitter and select the top-500 videos according to their metric. Goel et al. [24] model the spread of viral ideas as a diffusion process similar to the spread of an infectious disease [3, 6, 18], and introduce a formal metric for 'structural virality', which accounts for the spread of information through both large broadcasts (i.e. sources with a large audience), as well as multiple re-sharing with smaller individual audiences, and evaluate the metric on an extremely large dataset of a billion events on Twitter. Their findings are contradictory to [29], since they posit that structural virality is dependent primarily on the size of the largest individual broadcast (person with the most number of followers) and is low throughout.

2.2 Attribute Learning

Beginning with the seminal work of Parikh and Grauman [41] in 2011, predicting and understanding relative attributes in images is becoming an area of active research. In their paper, they introduce the importance of identifying 'relative' attributes in images, and propose a solution based on shallow image features and the RankSVM [31] formulation propsed by Joachims in 2002. There has been significant subsquent improvements to the original formulation, such as the work of Souri et al [49], who extend the original relative attribute learning formulation using a ranking network introduced by Cao et al [16].

Singh et al [48] propose a novel algorithm for relative attribute prediction using spatial transformer networks [28] in conjunction with a siamese architecture [17] to learn relative attribute prediction using the ranking loss proposed by [16]. They achieve the current state-of-the-art in realtive attribute prediction, surpassing the competing approaches of Xiao and Lee [52] which attempt to learn spatial extents of attributes using ensemble image representations. Our work extends research in this domain by demonstrating the applicability of these pairwise techniques in prediction tasks which are otherwise modelled as classification and regression, and the introduction of transfer learning into the domain of solving relative attribute prediction.

Our motivation, through this paper, is to provide a pipeline for understanding the nature of content virality. We understand that virality is a function of both content and the nature of the network, however, we tackle the problem of understanding the information - specifically, visual cues that can discriminate between viral and non-viral content. Our technical contribution is primarily to provide a general pairwise comparative deep architecture to learn a classification over the image set based on the strength of a visual attribute. We also aim to account for the disparity in the network topology present in the spread of viral information by creating a metric and dataset that balances the network effect. In the further sections, we present a more formal definition of the problem, and provide experimental results as well.

3 APPROACH

In [19], the authors introduce an image dataset for predicting image virality. This dataset has images along with a metric encapsulating the virality of the image, calculated by taking into account the number of reshares over a short amount of time, which is characteristic to viral images. We also experiment with the image dataset introduced by [32] to predict image popularity, details of which are explained in Section 3.3. We found that approaching these problems as regression tasks, where we aim to predict the virality/popularity metric for each image, was diffult to learn with considerable accuracy. This occured primarily because the labels themselves were varying in scale and were better understood in a relative sense.

3.1 Problem Formulation

Given the relative nature of the labels, we formulate the problem of virality prediction as a pairwise relative attribute prediction task. More specifically, we formulate a 2AFC (2 Alternative Forced Choice) problem, where the taask is to predict the more viral or more popular among a pair of images. We find this approach to be far more effective at predicting virality and popularity, since it eliminates variations in scale across samples. Our inference task can be explicitly stated as follows - for a pair of images (I_1, I_2) in our set of total images $\mathbf{I}$, and a relative label $y \in \{l, r\}$, we take a set of input triplets (I_1, I_2, y) during training time as input and predict the image with the stronger attribute (one of left or right) during test.

3.2 Architecture

Jaderberg et al [28] in 2015 introduce a powerful class of neural networks known as spatial transformer networks (STNs), which are aimed at learning an affine transformation of the input from the label provided. These architectures have been applied with success for virality prediction by [48]. We aim to select regions of the image corresponding to areas of interest for virality prediction. We can

formulate the task of STNs as learning an affine transformation $\mathbf{A}_\theta$ of the form

$$\mathbf{A}_\theta = \begin{bmatrix} s & 0 & t_x \\ 0 & s & t_y \end{bmatrix} \quad (1)$$

The parameters s, t_x, t_y can adjust the transformation by adjusting the scale and translation across x and y axes respectively. As described in detail by [28], these parameters are learnt via backpropagation.

Siamese Spatial Transformer Networks: The basic architecture is described as follows - the CNN **ViralNet** takes in inputs $(\mathbf{I}_1, \mathbf{I}_2)$ and label y. N is formed of two subnetworks N_1, N_2, which have identical weights and produce scalar output $v(\mathbf{I}_1)$ and $v(\mathbf{I}_2)$ each. Each of these subnetworks N_i is composed of two subnetworks each - a spatial transformer network S and a ranking network R.

The spatial transformer network S is a multilayer convolutional network which produces an affine transformation as output. This affine transformation is applied to the input image and fed to the ranking network R. The ranking network takes a region-of-interest of the original image as input and produces a ranking score for the attribute in question. As a baseline, we implement the architecture described in [48] and can be referred to for more details. This network is described in Figure 1.

Multiscale Region Pyramid: A primary issue in applying pairwise STNs to imagery on the Internet is that cues in virality prediction are present at different scales, in contrast to a singular scale in regular attribute prediction. We can include response at multiple scales by adding filter banks with larger filter sizes to capture larger regions of context. However, this would increase the model complexity. Instead, we can downsample input images to simulate a larger detector. Hence, in addition to the first spatial transformation that operates on the entire image, we learn two additional spatial transformations, which operate at 0.5× and 0.25× downsampled version of the original image. Given an input image $\mathbf{I}$, we train 3 spatial transformer networks S_1, S_2, S_3 to provide input to the ranking network:

$$S_1(\mathbf{I}) = \mathbf{A}_{\theta_1}\mathbf{I}_{\times 1} \quad (2)$$
$$S_2(\mathbf{I}) = \mathbf{A}_{\theta_2}\mathbf{I}_{\times 0.5} \quad (3)$$
$$S_3(\mathbf{I}) = \mathbf{A}_{\theta_3}\mathbf{I}_{\times 0.25} \quad (4)$$

Here, $\mathbf{I}_x$ represents image $\mathbf{I}$ downsampled by scale x. Contrary to [48], who provide image $\mathbf{I}$ and $S_1(\mathbf{I})$ as both inputs to the ranker network, we provide 4 images - $\{\mathbf{I}, S_1(\mathbf{I}_{\times 1}), S_2(\mathbf{I}_{\times 0.5}), S_3(\mathbf{I}_{\times 0.25})\}$ to the feature extractor (see Figure 2).

While the baseline network learns a ranking function $R(\mathbf{I}, S(\mathbf{I}))$, for each of the images in the input pair, we learn a ranking network $R(\mathbf{I}, S_1(\mathbf{I}_{\times 1}), S_2(\mathbf{I}_{\times 0.5}), S_3(\mathbf{I}_{\times 0.25}))$ to generate a ranking score for each image in the input pair. To avoid an increase in number of parameters, the ranking function is formulated as a feature extractor f_R followed by a ranker r_R. f_R extracts a vector feature representation on each of the input sub-images, and concatenates them to get vector t_R (see Equation 6. r_R then learns a linear ranking function on t_R to obtain the final scalar output v (see Figure 2 for a better understanding of the same). This network is known as **ViralNet-M**, and is described in Figure 2.

$$t_R(\mathbf{I}) = [f_R(\mathbf{I}), f_R(S_1(\mathbf{I}_{\times 1})), f_R(S_2(\mathbf{I}_{\times 0.5})), f_R(S_3(\mathbf{I}_{\times 0.25}))] \quad (5)$$
$$v(\mathbf{I}) = \mathbf{w}_v^T t_R(\mathbf{I}) \quad (6)$$

Category Supplement Network: Apart from the label for the attribute strength, our target dataset contains labels for the category each submitted image belongs to. We utilize this additional supervision to improve prediction accuracy. For the training set, we train a neural network C on images with category values as labels. We append copies of trained C into our **ViralNet** and **ViralNet-M** architectures (one copy for each input image) and concatenate the features from C into the final fully-connected layer to produce scalar values v. These architectures, named **ViralNet-C** and **ViralNet-MC** are described in Figure 3.

3.3 Datasets

Viral Images Dataset: As described in [19], Lakkaraju et al. [36] crawled 132,000 entries submitted to Reddit over a period of four years, images of which form this dataset for virality prediction [19]. This dataset has 10,078 images, each with a relative score for virality. To create our train and validation sets, we sample images uniformly into training and test sets (80:20), and create image pairs within each of these. Our training data contains 10M image pairs, and validation data contains 1M image pairs, referred to as **Viral-Complete** in our results. For our test sets, we also use the two other datasets introduced by [19], the 'Viral and Non-Viral Images' dataset (referred to as **Viral-Pairs**) and 'Random Pairs' dataset (referred to as **Viral-RandomPairs**), which contain image pairs distinct from the training/validation sets. For more details, please refer to [19] . To train the category dataset, we use the 'Viral Categories Dataset' introduced by [19].

Image Popularity Dataset [32]: This dataset is the data utilized by Khosla et al. [32] 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 subcategories for construction of the dataset are-

One-per-user: Images sampled from the Visual Sentiment Ontology dataset [12] consisting of approximately 930K images from 400K users - collected by search Flickr for 3,244 adjective-noun-pairs. It represents the setting where different images correspond to different users, which is often the case in search results.

User-mix: This setting involves randomly selecting 100 users from the previous subset with between 10K and 20K public photos and accumulating all the corresponding images, resulting in approximately 1.4M images.

User-specific: The user-mix dataset is split into 100 different training and evaluation sets and results are averaged.

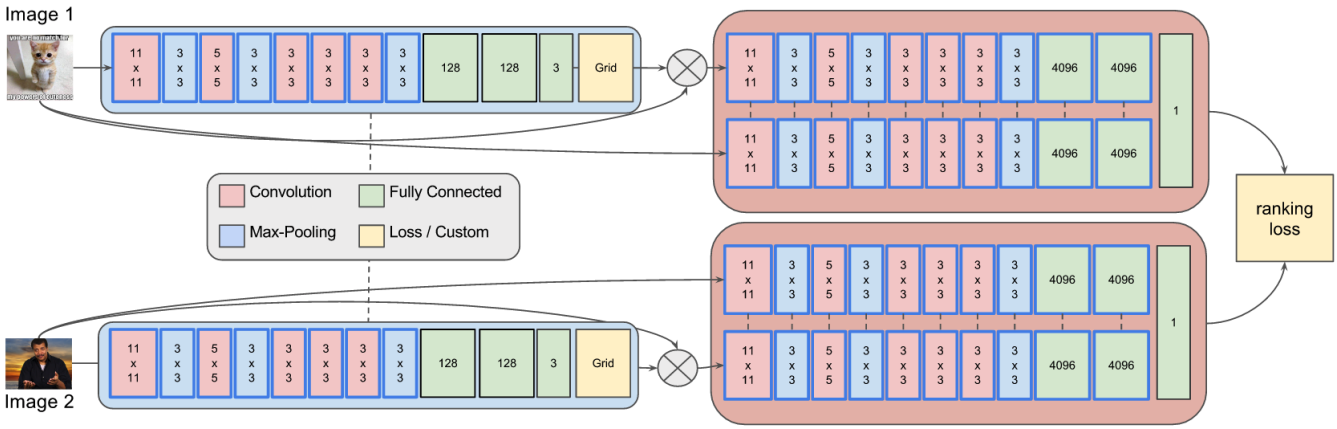

Figure 1: Full architecture for ViralNet. The first network, shaded in blue, represents the Spatial Transformer Network, whereas the consequent network, shaded in red, represents the Ranking Network. Note that both Spatial Transformer Network S and Ranking Network R share weights for both input images. Layers outlined in blue have been fine-tuned from AlexNet [35] weights.

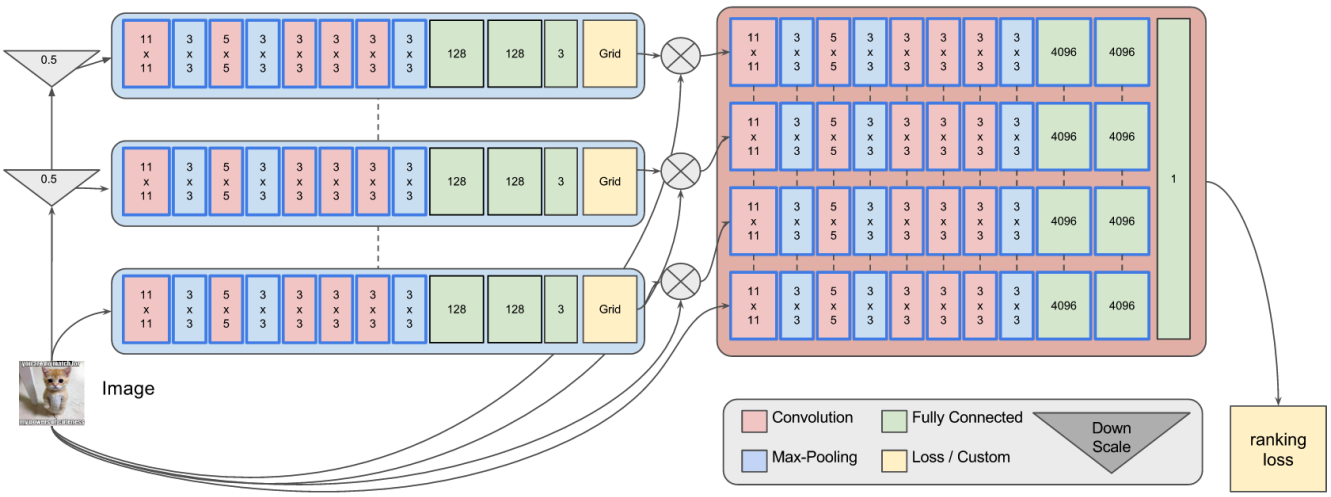

Figure 2: One half of the pyramidal multiscale ViralNet-M architecture. Here, instead of one spatial transformer network, we apply 3 spatial transformer networks at different scales and initialization settings for the image. The second half is ommitted due to space constraints but is identical in nature.

We use the same category model trained using the Virality data since there are no category labels present in this dataset.

Attribute Learning Datasets: As a comparative study, we also evaluate our model on the following relative attribute datasets-

LFW-10 (LFW): This dataset contains images of 10 face attributes, with 1000 training images, 1000 test images and 500 pairs per attribute for both training and testing. We follow the splits mentioned in [45]. For category information we use the publicly available OpenFace model to generate features [2].

UT-Zap50K-1 (Zappos): This dataset contains 4 shoe attributes, with 50,025 shoe images, and 1388 training, 300 testing pairs per attribute. We follow the splits mentioned in [53]. We do not train category models on this dataset owing to the lack of category labels.

OSR: This dataset consists of 6 outdoor scene attributes with 2,688 total images. We follow the splits mentioned in [48, 53]. For category information we use the publicly available Places205-AlexNet and Hybrid-CNN models [54].

3.4 Training

3.4.1 Learning to Rank using Gradient Descent. We train our **ViralNet** and **ViralNet-M** networks following the seminal work of RankNet [13].

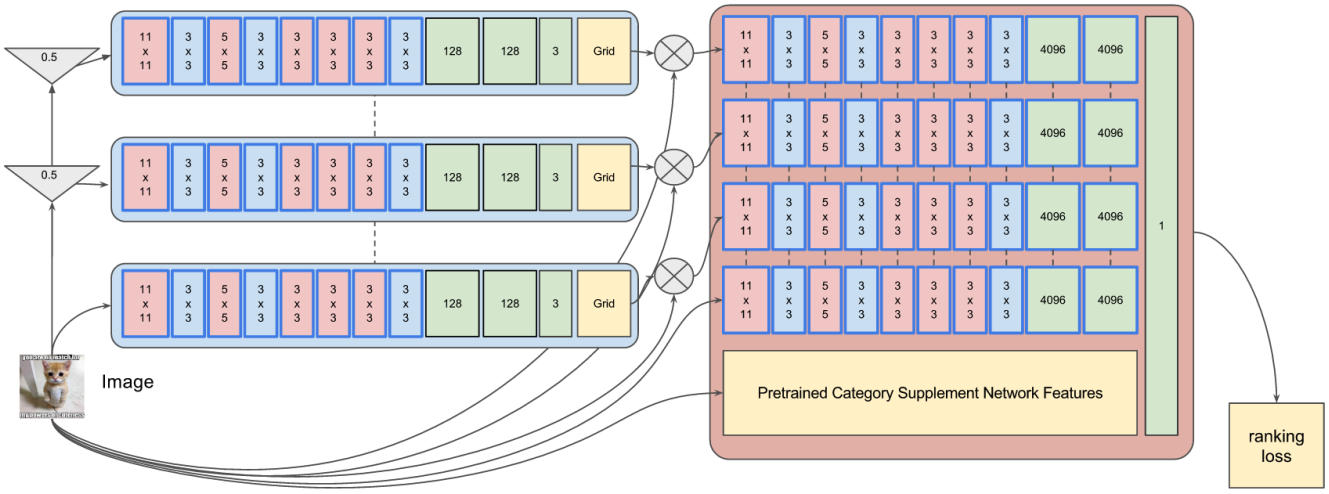

Figure 3: One half of the pyramidal multiscale ViralNet-MC architecture. This architecture retains all parameters from the ViralNet-M architecture except the final fully connected layer has an additional input of the category network. Similarly, for ViralNet-C architecture, the parameters are identical to the ViralNet architecture, but for the additional features from the category network as input to the final fully connected layer.

Specifically, the outputs $v(\mathbf{I}_1)$ and $v(\mathbf{I}_2)$ of the network are mapped to a probability P via a logistic function $P = \frac{e^{v(\mathbf{I}_1)-v(\mathbf{I}_2)}}{1+e^{v(\mathbf{I}_1)-v(\mathbf{I}_2)}}$ and then optimized using the cross entropy loss:

$$L_{rank}(\mathbf{I}_1, \mathbf{I}_2, y) = -y \cdot \log(P) - (1-y) \cdot \log(1-P) \tag{7}$$

Here, $y = 1$ if attribute strength is higher in I_1 compared to I_2 and vice-versa. For equality attributes, $y = 0.5$. The nature of the loss function, as described in [13] is such that it asymptotes to a linear function and is more robust to noise compared to a quadratic function.

3.4.2 Spatial Transformer Networks for Ranking. As mentioned in [48], it is possible that the spatial transformer networks predict regions outside the input image in order to minimize ranking error (by producing the same blank image input). To counter this, we also add the spatial transformation penalty as described by [48], which penalizes the spatial transformer networks if they produce regions outside the boundaries of the image. This spatial loss is given by -

$$L_{spatial} = (C_x - s \cdot t_x)^2 + (C_y - s \cdot t_y)^2 \tag{8}$$

Here, t_x and t_y are the translations from the affine in horizontal and vertical directions, and s is the isotropic scaling factor. C_x and C_y are the center coordinates of the input image. For **ViralNet** and **ViralNet-C**, the loss function is hence [48]

$$L = (1-\lambda_1)(1-\lambda_2)L_{rank} + \lambda_1 L^1_{spatial} + \lambda_2 L^2_{spatial} \tag{9}$$

Here λ_is are indicators which are 1 if the affine produced by either of the subnetworks is outside the boundaries of the image, and $L^i_{spatial}$ is the spatial loss incurred by the ith subnetwork.

3.4.3 Multiscale Pyramid Spatial Transformers. For the multiscale pyramid networks, since there are 3 spatial transformer networks per subnetwork, the spatial transformer loss is calculated for each of these networks, and the total loss for networks **ViralNet-M** and **ViralNet-MC** is given by

$$L_M = (\prod_{i=1}^{3}(1-\lambda_{1,i})) \cdot (\prod_{i=1}^{3}(1-\lambda_{2,i}))L_{rank} + \sum_{i=1}^{3}\lambda_{1,i} \cdot L^{1,i}_{spatial} + \sum_{i=1}^{3}\lambda_{2,i}L^{2,i}_{spatial} \tag{10}$$

Here, $\lambda_{i,j}$ denotes if the affine produced by the jth spatial network in subnetwork i is out of image boundaries, and $L^{i,j}_{spatial}$ is the spatial loss produced by the jth spatial network in subnetwork i. During training time, each input image is cropped with 90% area and different crops are fed to different scales, to prevent overfitting and to map to different transformations.

We train these networks using backpropagation on the relative attribute training data, and the category networks are trained beforehand either from publicly available models or using the data described in the earlier section. The category models are not fine-tuned during the attribute training process, their weights are kept fixed. Additionally, the spatial loss $L_{spatial}$ applies only to the spatial transformer networks and not the ranker networks. All initialization schemes are following [48].

4 RESULTS

Our experiments were carried out using a combined architecture of TensorFlow [1] written in Python and Caffe [30] with the Python framework. All experiments were run on NVIDIA Tesla k40 machines. All mentioned model weights were obtained from the Caffe Model-Zoo [15]. We follow the learning rate, initialization and implementation details are identical to [48] for the single-scale networks.

Algorithm	Percentage Accuracy		
	Viral-Complete	Viral-Pairs	Viral-RandomPairs
Chance	50		
SVM + Image Features [19]	53.40	61.60	58.49
RankSVM + 2x2 Dense HOG Features	52.75	58.81	56.92
RankSVM + AlexNet fc7 Features [35]	54.41	61.93	58.58
RankSVM + VGGNet-16 fc7 Features [47]	55.18	63.01	59.15
Human [19]	-	71.76	60.12
Popularity API [32]	-	-	51.12
Human Annotated Atts. -5 [19]	-	-	65.18
SVM + Deep Attributes-5 [19]	-	-	68.10
Xiao and Lee [52]	63.25	75.23	73.22
ViralNet [48]	65.87	76.20	74.38
ViralNet-M	66.15	77.01	75.52
ViralNet-C	**67.88**	**77.60**	**76.78**
ViralNet-MC	**68.09**	**78.38**	**76.95**

Table 1: Table summarizing our empirical results on the Viral Images dataset. The results on different baselines have been replicated from the referred papers in each entry. We observe that on this dataset we comfortably outperform the existing state-of-the-art by a relative improvement of 12% on the toughest split, and relative improvements of 24% and 23.3% on the easier splits.

For the multi-scale pyramids, the scale values were initialized to 1, 0.5 and 0.25 times the initialization for single-scale network scale values respectively.

Baselines: Our baseline technique **ViralNet** when applied to the relative attribute datasets, comprises the exact experiments reported by [48]. Additionally, we compare with the work of Xiao and Lee [52] in discovering spatial extents for attributes, the baselines reported by Deza and Parikh [19] and features extracted from popular image classification architectures such as AlexNet [35] and VGGNet [47] as well as shallow image features plugged into a RankSVM [31].

4.1 Viral Images Dataset

Our experiments begin with the Viral Images Dataset [19], which is a comprehensive inference dataset for virality prediction. We find that even the baseline network [48] outperforms the existing state-of-the-art on this dataset [19] comfortably, and our final model **ViralNet-MC** gives a significant improvement over the baseline. Specifically, we observe that the addition of category features from the network provide positive signal to the base networks. We visualize the region networks for our multiscale and baseline networks in more details in the next section. The results on this dataset are described in more detail in Table 1.

4.2 Popular Images Dataset

The next experiments we perform are on the Popular Images Dataset [32]. On this dataset as well, we outperform the existing state-of-the-art comfortably on all splits. However, we find that the category subnetwork, which is trained on the 'Viral Categories' dataset, provides less supervision compared to the previous dataset. Our complete results are reported in Table 2.

4.3 Relative Attribute Datasets

We also evaluate our algorithm on relative attribute datasets to observe the performance benefits of category and multiscale interventions in end-to-end spatial transformer networks. As expected, we observe that the category supervision provides positive signal to the attribute prediction. Additionally, the multiscale intervention also provides additional signal in datasets which have multiple scales of localization, such as **OSR** and **LFW-10**. We report the mean prediction accuracy over all additional relative attribute datasets in Table 3. Note here that the architecture for **ViralNet** is identical to the architecture introduced by [48].

4.4 Qualitative Experiments

To further understand the qualitative nature of our results and interpret the decision surfaces produced by our models, we perform several visualization experiments. Singh et al [48] provide a qualitative localization experiment by visualizing the region of interest fed to the ranker networks from the spatial transformer networks, for both the single-scale and multi-scale pyramidal networks. We observe that our scale initialization is critical for the networks to feed different regions of interests at various scales in these networks. The visualizations in the virality and popularity datasets provide significant insights into the regions the convolutional networks are focusing on during prediction. We find that the networks latch on to areas of interest which can be visualized as important to virality - the presence of text, bright colors and humans in images promote virality, and the localization areas returned by these networks agree with these propositions. Figure 4 provides sample visualizations.

To understand in detail the contribution of the category features in attribute prediction, we also perform another qualitative visualization experiment. We visualize the nearest neighbors in the penultimate layers of both the **ViralNet** and **AlexNet** models for different values of virality strengths. In images classified by either networks, we examine the nearest neighbors in the training set for each, and observe that the improvement is achieved because

Algorithm	Percentage Accuracy		
	One-per-user	User-mix	User-specific
GIST [32]	50.32	56.61	57.74
Color Histogram [32]	56.13	57.58	59.07
Texture [32]	59.91	61.01	63.02
Color Patches [32]	60.22	61.09	64.19
Gradient [32]	62.12	63.04	64.47
DeCAF [20] Features [32]	63.89	65.05	64.89
Objects [32]	62.01	65.17	65.32
Combined [32]	64.02	65.98	66.18
RankSVM + AlexNet fc6 Features [35]	66.01	67.75	67.02
RankSVM + VGGNet-16 fc6 Features [47]	67.15	68.97	67.28
Xiao and Lee [52]	70.02	71.32	72.86
ViralNet [48]	71.89	74.08	75.33
ViralNet-M	72.57	75.59	76.99
ViralNet-C	72.84	75.52	76.97
ViralNet-MC	**73.01**	**76.07**	**77.78**

Table 2: Table summarizing our empirical results on the Popular Images dataset. The results on different baselines have been replicated from the referred papers in each entry. We observe that on this dataset we comfortably outperform the existing state-of-the-art by a relative improvement of 8%.

Algorithm	Mean Percentage Accuracy		
	LFW	Zappos	OSR
Parikh and Grauman [41] + CNN [48]	74.61	94.69	94.98
Xiao and Lee [52]	84.66	95.47	92.16
End-to-End Localization [48]	86.91	95.78	97.02
ViralNet-M	86.94	**96.01**	96.68
ViralNet-C	87.15	-	97.02
ViralNet-MC	**87.67**	-	**97.55**

Table 3: Table summarizing our empirical results on the several relative attribute learning datasets. The results on different baselines have been replicated from the referred papers in each entry. We observe that the category supplement networks provide positive signal in prediction, increasing accuracy across datasets.

of localization in the space of abstract attribute information. See Figure 5 for sample visualizations.

4.5 Ablation Studies

Our proposed model takes focuses on the usage of spatial transformer networks to localize regions important for the prediction of virality and popularity as attributes. Additionally, our formulation of tackling the prediction task as a relative attribute learning task focuses on the relative nature of attribute values as a stronger signal for modeling virality. In order to justify our two interventions in this problem, we conduct ablation studies for both interventions, first by removing the spatial transformers completely and using just a Siamese convolutional network (similar to [17]) with the RankNet loss [13], initalized with the AlexNet [35] and VGGNet-16 [47] model weights. Secondly, we also compare the straightforward formulation of training a regressor on existing architectures AlexNet and VGGNet-16 on the virality values. We summarize the results of this ablation experiment in Table 4, and our results justify both the choices as we see that the ablation models perform significantly poorly compared to the **ViralNet** architectures.

Algorithm	Percentage Accuracy		
	Complete	Viral-Pairs	Random-Pairs
AlexNet [35] Regression	53.01	57.72	54.19
VGGNet-16 [47] Regression	56.12	60.11	58.23
AlexNet [35] Siamese	58.17	61.11	59.87
VGGNet-16 [47] Siamese	60.23	66.15	63.15
ViralNet	**65.87**	**76.20**	**74.38**

Table 4: Table summarizing the results of our ablation studies on the Viral Images dataset. We find that both our interventions to the existing models provide significant improvements in virality prediction.

5 CONCLUSION

In this paper, we presented a novel interpretation of the problem of predicting virality as a pairwise attribute learning task, and presented a novel multiscale pyramidal deep architecture to predict virality with a significant improvement (19.1% average relative increment) over the existing state-of-the-art. Additionally, our network is able to localize several different areas of interest for each image which provides significant insights into the qualitative understanding of virality, which are vital in the construction and easy caching of viral content on the Internet.

The augmentation with category information indicates the value of supervision via well-defined labels on the performance of an amorphous recognition task, and is suggestive of similar improvements on other abstract learning tasks such as memorability prediction and urban perception. It has long been understood that the number of shares and connectivity of a content outlet has a significant impact on a shared image's popularity, and our work provides evidence that there is significant contribution of the content as well in its popularity.

In conclusion, we provide a qualitative and quantitative improvement over the existing work on image virality prediction and abstract attribute learning, that can be generalizable to any abstract classification task. We believe this technique of visualizing network regions will be useful in understanding relative attributes and

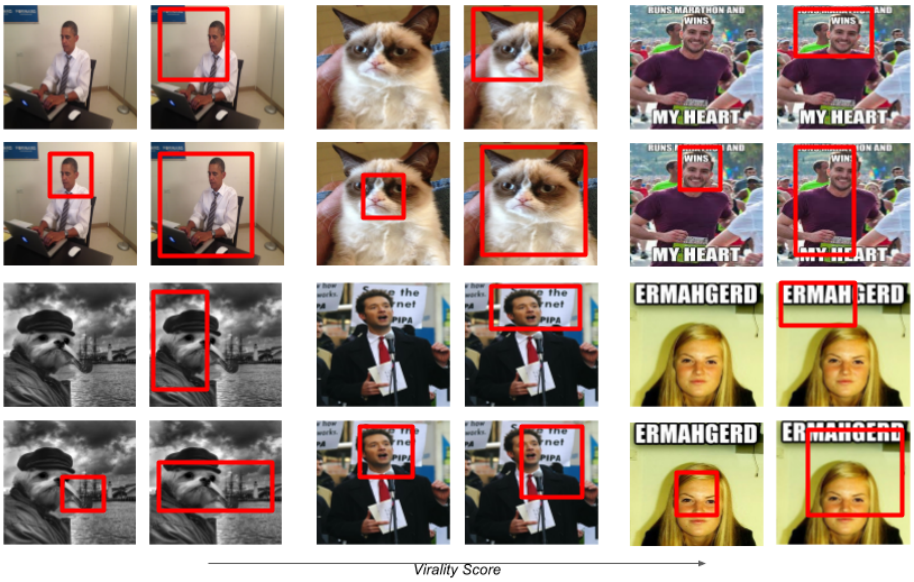

Figure 4: We examine the qualitative effectiveness of the spatial transformer networks in this experiment. For our ViralNet-MC model, we feed images of varying degrees of virality metric and observe the regions fed into the ranking networks at different scales. We find that at different scales, these regions focus on different aspects of input images. For each image group, the top-left is the original image, bottom-left is the localization at 0.25 scale, top-right is the localization at 0.5 scale and bottom-right is the localization at original scale. We see that the model latches on to faces at smaller scales, and text if present.

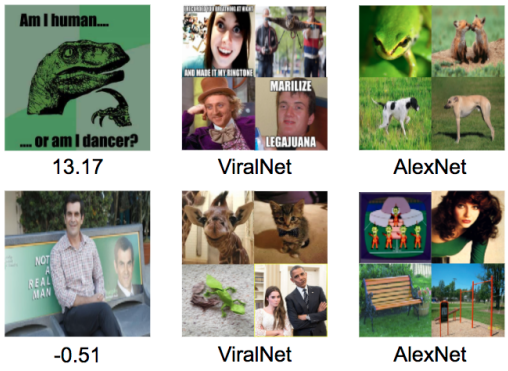

Figure 5: Nearest Neighbours for two sample inputs to both images in subspace of pre-final activations with their representative virality scores. We see that our architecture ViralNet localizes more closely to the viral categories, whereas AlexNet localizes more closely to visual representations and objects.

classification, as well as further our understanding what promotes virality on the Internet.

REFERENCES

[1] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and others. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. *Software available from tensorflow. org* (????).

[2] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *OpenFace: A general-purpose face recognition library with mobile applications.* Technical Report. CMU-CS-16-118, CMU School of Computer Science.

[3] Roy M Anderson, Robert M May, and B Anderson. 1992. *Infectious diseases of humans: dynamics and control.* Vol. 28. Oxford university press Oxford.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[5] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 519–528.

[6] Frank M Bass. 1969. A simultaneous equation regression study of advertising and sales of cigarettes. *Journal of Marketing Research* (1969), 291–300.

[7] Jonah Berger. 2011. Arousal increases social transmission of information. *Psychological science* 22, 7 (2011), 891–893.

[8] Jonah Berger. 2013. *Contagious: Why things catch on.* Simon and Schuster.

[9] Jonah Berger and Chip Heath. 2007. Where consumers diverge from others: Identity signaling and product domains. *Journal of Consumer Research* 34, 2 (2007), 121–134.

[10] Jonah Berger and Katherine L Milkman. 2012. What makes online content viral? *Journal of marketing research* 49, 2 (2012), 192–205.

[11] Jonah Berger and Eric M Schwartz. 2011. What Drives Immediate and Ongoing Word of Mouth? *Journal of Marketing Research* 48, 5 (2011), 869–880.

[12] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 223–232.

[13] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*. ACM, 89–96.

[14] BuzzFeed. 2016. BuzzFeed. https://buzzfeed.com. (2016). [Online; accessed 10-Feb-2016].

[15] BVLC. 2016. Caffe Model-Zoo. https://github.com/bvlc/caffe/wiki/model-zoo. (2016). [Online; accessed 10-Feb-2016].

[16] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*. ACM, 129–136.

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

[18] James Coleman, Elihu Katz, and Herbert Menzel. 1957. The diffusion of an innovation among physicians. *Sociometry* 20, 4 (1957), 253–270.

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

[20] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2013. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531* (2013).

[21] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. 2016. Deep learning the city: Quantifying urban perception at a global scale. In *European Conference on Computer Vision*. Springer, 196–212.

[22] Gawker. 2016. Gawker. https://gawker.com. (2016). [Online; accessed 10-Feb-2016].

[23] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

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

[25] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Gool. 2013. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1633–1640.

[26] Lichan Hong, Gregorio Convertino, and Ed H Chi. 2011. Language Matters In Twitter: A Large Scale Study.

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

[28] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and others. 2015. Spatial transformer networks. In *Advances in Neural Information Processing Systems*. 2017–2025.

[29] Puneet Jain, Justin Manweiler, Arup Acharya, and Romit Roy Choudhury. 2014. Scalable Social Analytics for Live Viral Event Prediction.

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

[31] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*. 133–142.

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

[33] Aditya Khosla, Jianxiong Xiao, Phillip Isola, Antonio Torralba, and Aude Oliva. 2012. Image memorability and visual inception. In *SIGGRAPH Asia 2012 Technical Briefs*. ACM, 35.

[34] Aditya Khosla, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2012. Memorability of image regions. In *Advances in Neural Information Processing Systems*. 305–313.

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

[36] Himabindu Lakkaraju, Julian J McAuley, and Jure Leskovec. 2013. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. *ICWSM* 1, 2 (2013), 3.

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*. Springer, 740–755.

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

[39] Amandianeze O Nwana, Salman Avestimehr, and Tsuhan Chen. 2013. A latent social approach to youtube popularity prediction. In *Global Communications Conference (GLOBECOM), 2013 IEEE*. IEEE, 3138–3144.

[40] Funny or Die. 2016. FunnyORDie. https://funnyordie.com. (2016). [Online; accessed 10-Feb-2016].

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

[42] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. 2011. RT to Win! Predicting Message Propagation in Twitter.

[43] Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. 2013. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 365–374.

[44] Reddit. 2016. Reddit. https://reddit.com. (2016). [Online; accessed 10-Feb-2016].

[45] Ramachandruni N Sandeep, Yashaswi Verma, and CV Jawahar. 2014. Relative parts: Distinctive parts for learning relative attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3614–3621.

[46] David A Shamma, Jude Yew, Lyndon Kennedy, and Elizabeth F Churchill. 2011. Viral Actions: Predicting Video View Counts Using Synchronous Sharing Behaviors.. In *ICWSM*.

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

[48] Krishna Kumar Singh and Yong Jae Lee. 2016. End-to-End Localization and Ranking for Relative Attributes. In *European Conference on Computer Vision (ECCV), 2016*.

[49] Yaser Souri, Erfan Noury, and Ehsan Adeli-Mosabbeb. 2015. Deep Relative Attributes. *arXiv preprint arXiv:1512.04103* (2015).

[50] Naman Turakhia and Devi Parikh. 2013. Attribute dominance: What pops out?. In *Proceedings of the IEEE International Conference on Computer Vision*. 1225–1232.

[51] John Wihbey. 2014. The Challenges of Democratizing News and Information: Examining Data on Social Media, Viral Patterns and Digital Influence. *Viral Patterns and Digital Influence (June 9, 2014)* (2014).

[52] Fanyi Xiao and Yong Jae Lee. 2015. Discovering the Spatial Extent of Relative Attributes. In *Proceedings of the IEEE International Conference on Computer Vision*. 1458–1466.

[53] Aron Yu and Kristen Grauman. 2014. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 192–199.

[54] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.