Intrinsic Image Popularity Assessment

Keyan Ding
City University of Hong Kong
Hong Kong
keyanding2-c@my.cityu.edu.hk

Kede Ma
City University of Hong Kong
Hong Kong
kede.ma@cityu.edu.hk

Shiqi Wang*
City University of Hong Kong
Hong Kong
shiqwang@cityu.edu.hk

ABSTRACT

The goal of research in automatic image popularity assessment (IPA) is to develop computational models that can accurately predict the potential of a social image to go viral on the Internet. Here, we aim to single out the contribution of visual content to image popularity, *i.e.*, intrinsic image popularity. Specifically, we first describe a probabilistic method to generate massive popularity-discriminable image pairs, based on which the first large-scale image database for intrinsic IPA (I^2PA) is established. We then develop computational models for I^2PA based on deep neural networks, optimizing for ranking consistency with millions of popularity-discriminable image pairs. Experiments on Instagram and other social platforms demonstrate that the optimized model performs favorably against existing methods, exhibits reasonable generalizability on different databases, and even surpasses human-level performance on Instagram. In addition, we conduct a psychophysical experiment to analyze various aspects of human behavior in I^2PA.

CCS CONCEPTS

• **Information systems** → **Content analysis and feature selection**; • **Human-centered computing** → **Social media**.

KEYWORDS

Intrinsic image popularity, learning-to-rank, deep neural networks, human behavior analysis.

ACM Reference Format:
Keyan Ding, Kede Ma, and Shiqi Wang. 2019. Intrinsic Image Popularity Assessment. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351007

1 INTRODUCTION

Recent years have witnessed an accelerated proliferation of images and videos being uploaded to various social platforms such as Instagram[1], Flickr[2], and Reddit[3]. Some photos turn to be extremely popular, which gain millions of likes and comments, while some

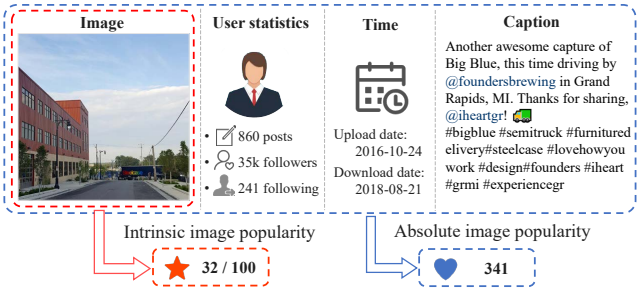


Figure 1: Absolute IPA versus I^2PA. Absolute IPA makes use of the image and all its relevant social and textual information to predict the number of received likes, while I^2PA relies solely on the image itself, exploiting the visual content for popularity prediction. 341 is the number of likes received by the image, and 32 is the predicted intrinsic popularity score by our model (with a re-scaled maximum score of 100).

are completely ignored. Even for images uploaded by the same user at the same time, their popularity may be substantially different. This interesting phenomenon motivates us to ask what the secret of image popularity is. It is generally believed that image popularity is determined by a combination of factors, including the visual content, the user statistics, the upload time, and the caption [14, 20, 30, 36].

Computational models for *absolute* image popularity assessment (IPA) attempt to predict the number of received likes/comments of an image by combining all visual and non-visual factors [20, 30]. Here we aim to single out the contribution of visual content to image popularity, namely *intrinsic* image popularity, and develop computational models for intrinsic IPA (I^2PA) for several reasons. First, by focusing on the visual content, I^2PA is a cleaner and easier-to-interpret problem than absolute IPA (see Fig. 1). Second, computational models for I^2PA guide the identification of potentially popular images with no social and textual contexts, and hold much promise in optimizing social image management and recommendation systems in the long run. For example, more computation and storage resources may be allocated to images with high intrinsic popularity. Third, from the users' perspective, I^2PA model predictions are ideal indicators of which images in their personal albums are worth uploading to gain great attention, when they just join the social network and have no social interactions. Moreover, the users may gain inspiration regarding how to filter and prioritize photos assisted by the model instead of their own biased opinions. Last, analyzing how image attributes such as image quality, aesthetics, contexts, and semantics contribute to intrinsic image popularity is by itself an interesting problem for human and computer vision study (see Sections 4.2 and 4.3).

*Corresponding author.
[1]https://www.instagram.com
[2]https://www.flickr.com
[3]https://www.reddit.com

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351007

In this paper, we conduct a systematic study of I^{2}PA based on Instagram, a leading photo-sharing social network with over one billion monthly active users on the web and mobile clients [9]. We first develop a probabilistic method to construct a new form of data - popularity-discriminable image pairs (PDIPs), which contain rich information about intrinsic image popularity by reducing the influence of non-visual factors. We show that such PDIPs can be generated at very low cost and high accuracy, leading to the first large-scale image database for I^{2}PA. We then train deep neural networks (DNNs) to predict intrinsic image popularity by learning-to-rank [5, 8, 19] millions of PDIPs in the proposed database. Experimental results on Instagram and several other social platforms show that our model predicts intrinsic image popularity accurately, outperforms state-of-the-art methods (*e.g.*, the commercialized products Virality Detection [34] and LikelyAI [24]), and generalizes reasonably. Moreover, we conduct a psychophysical experiment to collect human opinions on intrinsic image popularity, and find that our method slightly surpasses the human-level performance on Instagram.

Our contributions are four-fold. First, we revisit the concept of intrinsic image popularity with a new problem formulation. Second, we construct the first large-scale image database for I^{2}PA, consisting of more than two million PDIPs with reliable annotations. Third, we develop a computational model for I^{2}PA based on a DNN, which delivers human-level performance. Fourth, we conduct a psychophysical experiment to analyze various aspects of human behavior in I^{2}PA.

2 RELATED WORK

Popularity assessment of social media content (*e.g.*, texts, photos, and videos) has been an active research field in the past decade. Traditional computer vision and natural language processing methods focused on handcrafting image and text features [10, 26], which requires extensive human expertise. Recently, there has been a roaring wave of developing DNNs that emphasize automatic hierarchical feature learning for IPA [18, 24, 34, 40].

Most studies investigate image popularity based on the professional photo-sharing site Flickr. McParlane *et al.* [30] predicted image popularity (*i.e.*, the number of comments and views) using image and user contexts. Khosla *et al.* [20] predicted the normalized view counts, and analyzed the impact of low-level features (color patch variance, gradient, and texture), middle-level features (GIST [33]) and high-level features (semantics) on the prediction accuracy. Gelli *et al.* [14] conducted a qualitative analysis regarding which visual factors may influence image popularity. Wu *et al.* [41–43] incorporated multiple time-scale dynamics in predicting image popularity. Zhang *et al.* [45] proposed a user-guided hierarchical attention network for multi-modal content popularity prediction.

There are also several studies of IPA using other social networks. Mazloom *et al.* [29] predicted the popularity of brand-related posts on Instagram, and later extended their method to account for category specific posts [28]. Zhang *et al.* [46] addressed user-specific popularity prediction on Instagram with a dual-attention mechanism. Deza and Parikh [12] cast IPA as a classification problem based on images collected from Reddit, a popular website composed of many interest-centric sub-communities. Hessel *et al.* [17] compared multi-modal content with social contexts in predicting relative popularity on Reddit. They found that visual and textual features tend to outperform user statistics.

Due to its commercial value, many companies have developed computational models for IPA. LikelyAI [24] is such a product that assesses the popularity of Instagram posts. Trained on millions of images, LikelyAI is claimed to recognize popular patterns on Instagram. Virality Detection [34] is a similar tool to score images based on their potentials to become popular on social media. Virality Detection is trained on a massive corpus of web images, and achieves high accuracy on the AVA dataset [32], suggesting close agreement between image aesthetics and image popularity.

Most of the above-mentioned methods try to predict absolute image popularity by combining various visual and non-visual features. In contrast, I^{2}PA receives little attention despite broad practical applications. The closest studies to ours are due to Cappallo *et al.* [7], and Dubey and Agarwal [13]. Cappallo *et al.* learned a visual popularity predictor from both popular and unpopular images on Flickr using RankSVM [19]. However, their training pair generation process does not exclude the impact of non-visual factors such as user statistics and textual information. Moreover, their model is not end-to-end optimized, and may result in suboptimal performance. Dubey and Agarwal [13] modeled image popularity with pairwise spatial transformer networks, whose training pairs suffer from similar problems [7]. In addition, the image content based on Reddit [12] is not diverse enough, and may hinder the generalizability of the learned networks.

3 METHOD

In this section, we first describe the probabilistic method for PDIP generation and ways to reduce the impact of non-visual factors. Based on the method, we build the first large-scale image database for I^{2}PA. Last, we describe the specification and learning of our DNN-based computational model for I^{2}PA.

3.1 PDIP generation and database construction

At the beginning, we crawl more than one million active users on Instagram using the snowball sampling [4] through the chain of followers. To decorrelate the sampled users, we randomly remove 80% of them, and collect the information of each post of the remaining users, including the download time, the post URL, the user ID, the content type, the upload time, the caption (including emojis, hashtags, and @ signs), the number of likes, and the number of comments. As a result, we obtain over 200 million distinctive posts as the candidates to build our database. Note that all data are collected via HTTP requests for research purpose only.

We present in detail the probabilistic method for PDIP generation. In agreement with [1, 28, 29], the log-scaled number of likes S is considered as the ground truth for absolute image popularity, based on which we make two mild assumptions.

- S obeys a normal distribution (assuming the Thurstone's model [39])

$$p(S|\mu) \propto \exp\left(-\frac{(S-\mu)^2}{2\sigma^2}\right), \tag{1}$$

Table 1: Statistics of the proposed large-scale image database for I^{2}PA

Attribute	Value
Number of PDIPs	2.5×10^6
Number of users involved	1.1×10^5
Average likes per image	5.3×10^3
Average $P(Q_A \geq Q_B \mid S_A, S_B)$	0.978
Average upload time interval	4.8 days
Proportion of no hashtag	45.6%
Proportion of no @ sign	47.9%
Proportion of no caption	11.1%
Average length of descriptive text	2.1 words

with mean μ and standard deviation (std) σ. Here μ is a random variable, which can be viewed as the average number of likes received by an image in the log scale, if the image were uploaded and rated multiple times. Without any prior knowledge, we assume $p(\mu)$ is flat with a finite positive support. To simplify the derivation, we treat σ as a positive constant to be determined.

- The intrinsic image popularity Q is a monotonically increasing function of μ.

Using the Bayes' theorem, we have

$$p(\mu|S) \propto p(S|\mu)p(\mu) \propto p(S|\mu)\,, \tag{2}$$

where the second proportion follows from the assumption that $p(\mu)$ is flat. That is, conditioning on S, μ is Gaussian with mean S and std σ. To ensure that Image A is intrinsically more popular than Image B in a PDIP, we compute the probability

$$P(Q_A \geq Q_B|S_A, S_B) = P(\mu_A \geq \mu_B|S_A, S_B) \tag{3}$$

$$= P(\mu_A - \mu_B \geq 0|S_A, S_B)\,, \tag{4}$$

where Eq. (3) follows from the assumption that Q is a monotonically increasing function of μ. Assuming the variability of intrinsic popularity across images is uncorrelated, and conditioning on S_A and S_B, the difference $\mu_{AB} = \mu_A - \mu_B$ is also Gaussian

$$p(\mu_{AB}|S_A, S_B) \propto \exp\left(-\frac{(\mu_{AB} - (S_A - S_B))^2}{4\sigma^2}\right). \tag{5}$$

Combining Eq. (3) with Eq. (5), we have

$$P(Q_A \geq Q_B|S_A, S_B) = \Phi\left(\frac{S_A - S_B}{\sqrt{2}\sigma}\right), \tag{6}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. $P(Q_A \geq Q_B|S_A, S_B)$ indicates the probability that Image A is intrinsically more popular than Image B. In practice, we choose a large threshold $P(Q_A \geq Q_B|S_A, S_B) \geq T$ to ensure the popularity discriminability of PDIPs.

Enumerating all possible pairs that only satisfy the probability constraint is not enough, as image popularity may be affected by other non-visual factors. Therefore, it is desirable to further constrain the two images in a pair to have similar textual and social contexts. According to the mechanism of Instagram, we consider three major non-visual factors.

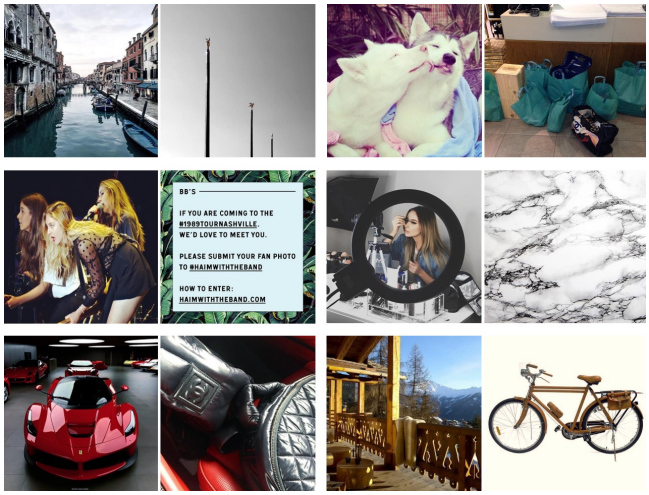


Figure 2: Representative PDIPs from the proposed database. The left image in each pair is expected to be intrinsically more popular than the right one, which has been confirmed by our psychophysical experiment. More than 80% of the subjects think the left images would receive more likes on social media than the right ones.

- **User statistics.** Several studies have showed that the popularity of an image is highly correlated with the user who uploads it [6, 14, 44]. The most obvious reason is that different users have different numbers of followers. Images posted by the users with more followers have higher chances of receiving more likes. The number of active followers and their preferences make the relationship more complicated. Considering the above issues, we restrict images from the same user for PDIP generation.
- **Upload time.** A user often has a different number of followers at different times. To reduce this effect, the post time difference of the two images in a PDIP is set to a maximum of ten days. In addition, it is helpful to exclude images just uploaded to the social network as the number of likes has not reached a stable state. According to the analysis in [1], the number of likes for most images stops to increase after four weeks. As such, we exclude images posted within one month. In addition, the upload time in a day may also affect image popularity. Failing to model this issue may result in minor label noise. However, as will be clear in Section 4.2, our learning process is quite robust to label noise in PDIPs.
- **Caption.** Image captions have a noticeable impact on image popularity, especially those containing hashtags and @ signs. A hot hashtag contributes significantly to image popularity because of the extensive exposure to viewers beyond followers. Generally the more hashtags of a post, the greater chances of receiving more likes. @ signs may also affect image popularity. For example, images @ a celebrity would probably receive more likes than those @ an ordinary user or without the @ sign. To remove the textual bias, we restrict the hashtag and @ sign of the images in a PDIP to be the

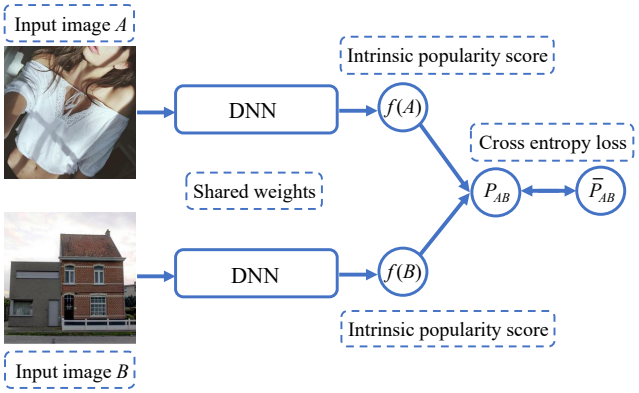


Figure 3: Computational models for I^{2}PA based on DNNs. The two streams have the same network architecture, whose model parameters are shared and optimized by minimizing the binary cross entropy loss. Either stream can be used to predict intrinsic image popularity.

same (in both content and number). Moreover, the length of the caption (excluding the hashtag and @ sign) is restricted to a maximum of six words.

We summarize the four constraints of images for PDIP generation as follows:

- $P(Q_A \geq Q_B | S_A, S_B) \geq T$;
- from the same user;
- posted more than one month and within ten days;
- caption with a maximum of six words and the same hashtag/@ sign.

An Instagram post may contain multiple images, videos, or a mixture of them. Here we only consider single image posts because it is difficult to allocate the number of likes across multiple images in a post. In addition, we exclude images with less than 50 likes to reduce the boundary effect. From more than 200 million candidate images on Instagram, we construct the first large-scale database for I^{2}PA, which contains approximately 2.5 million of PDIPs, satisfying all of the above constraints. To ensure the content diversity, one image only participates in one PDIP. Table 1 summarizes the statistics of the proposed database, and Fig. 2 shows six sample PDIPs.

3.2 DNN-based computational models for I^{2}PA

In this subsection, we describe a DNN-based computational model for I^{2}PA by learning-to-rank millions of PDIPs. As a machine learning technique, learning-to-rank was extensively studied in the context of information retrieval [25], and later found its way to computer vision [8], image processing [27], and natural language processing [23]. Pairwise learning-to-rank approaches assume that the relative order between two instances is known (or can be inferred), and aim to minimize the average number of incorrectly ordered pairs. The PDIPs in our database fit the pairwise learning-to-rank scheme naturally, and we use them to drive the learning of a Siamese architecture for I^{2}PA (see Fig. 3).

The input A of a PDIP to the first stream is an RGB image, and the output is the predicted intrinsic popularity score $Q_A = f(A)$. Similarly, the second stream inputs the other image B and predicts $Q_B = f(B)$. The network architectures of the two streams are the same, whose weights are shared during training and testing. We compute the predicted score difference $O_{AB} = f(A) - f(B)$, and convert it to a probability using a logistic function

$$P_{AB} = \frac{\exp(O_{AB})}{1 + \exp(O_{AB})}. \tag{7}$$

Denote the ground-truth binary label of a PDIP as $\bar{P}_{AB}$, where $\bar{P}_{AB} = 1$ indicates Image A is intrinsically more popular than Image B and otherwise $\bar{P}_{AB} = 0$. We adopt the binary cross entropy as the loss function

$$\ell = -\bar{P}_{AB} \log P_{AB} - (1 - \bar{P}_{AB}) \log(1 - P_{AB}) \tag{8}$$

$$= -\bar{P}_{AB} O_{AB} + \ln(1 + \exp(O_{AB})). \tag{9}$$

After training, the optimal predictor $f^\star$ (either stream in the Siamese architecture) is learned. Given a test image X, we perform a standard forward pass to obtain the predicted intrinsic popularity score

$$Q_X = f^\star(X). \tag{10}$$

4 EXPERIMENTS

In this section, we first describe the implementation details, including the default DNN architecture and the training procedure. We then quantitatively compare our model with the state-of-the-art. We also conduct a qualitative analysis of our model, and have a number of interesting observations. Last, we perform a psychophysical experiment to analyze human behavior in this task.

4.1 Implementation details

We adopt ResNet-50 [16] as our default DNN architecture, and replace the last layer with a fully connected layer of one output, representing the predicted intrinsic popularity score. The initial weights are inherited from models pre-trained for object recognition on ImageNet [11], expect for the last layer that is initialized by the method of He *et al.* [15]. The two parameters T and σ that govern the reliability of PDIP generation are set to 0.95 and 0.3, respectively. During both training and testing, the short side of the input image is rescaled to 256, from which a $224 \times 224 \times 3$ sub-image is randomly cropped. The training is carried out by optimizing the cross entropy function using Adam [21] with an ℓ_2 penalty multiplier of 10^{-4} and a batch size of 64. The learning rates for the pre-trained DNN layers and the last layer are set to 10^{-5} and 10^{-4}, respectively. After each epoch, we decay the learning rates linearly by a factor of 0.95. Training takes approximately one day on an Intel E5-2699 2.2GHz CPU and an NVIDIA Tesla V100 GPU. Our model takes 450 ms and 20 ms to process an image of size $224 \times 224 \times 3$ on CPU and GPU, respectively. To facilitate research in I^{2}PA, we make the PyTorch implementation of our model and the large-scale image database publicly available at https://github.com/dingkeyan93/intrinsic-image-popularity.

4.2 Quantitative evaluation

Main results on Instagram. We adopt pairwise accuracy as the quantitative measure, which is defined as the percentage of correctly ranked pairs. From the 2.5 million of PDIPs in the proposed database,

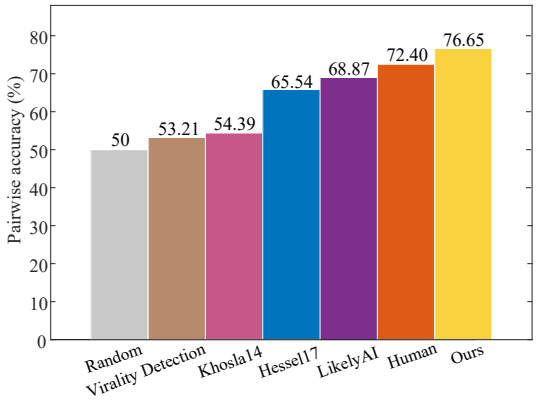


Figure 4: Pairwise accuracy on the test set consisting of 50,000 PDIPs. Note that the human-level performance is measured on 1,000 PDIPs randomly sampled from the test set due to the high cost of the psychophysical experiment.

we randomly choose 50,000 pairs for validation, 50,000 pairs for testing, and leave the rest for training. The weights that achieve the highest pairwise accuracy on the validation set are used for testing.

As a relatively new problem, it is difficult to find computational models specifically for I^{2}PA in the literature. We try our best to compare our model with four most relevant and state-of-the-art methods, whose implementations are publicly available for testing only. These are Khosla14 [20], Hessel17 [17], Virality Detection [34], and LikelyAI [24]. Khosla14 makes one of the first attempts to predict absolute image popularity. It also provides an API to assess intrinsic image popularity. Hessel17 is a multi-modal content popularity predictor based on Reddit. Six category-specific models are trained, and the one (for the *pics* category) that achieves the highest pairwise accuracy on our test set is used for comparison. Virality Detection and LikelyAI are two commercialized products, aiming to predict image popularity in a variety of practical scenarios (with or without social and textual contexts).

Fig. 4 shows the results, where we see that our model achieves the best performance with a pairwise accuracy of 76.65%. Khosla14 and Virality Detection marginally outperform the random guess baseline, which may be due to the distribution mismatch between training (Flickr) and testing (Instagram) images. Specifically, Instagram is a community conductive to self-expression, while Flickr focuses more on photographs of high visual quality and aesthetics. Hessel17 suffers from the similar issue, whose training data are crawled from Reddit rather than Instagram. Trained on Instagram images, LikelyAI performs slightly better than Hessel17, but is inferior to our model by a large margin. We believe this performance improvement arises because the PDIPs used for training contain reliable information about intrinsic image popularity, and our end-to-end optimized model is able to capture the features and attributes of images that are highly relevant to intrinsic image popularity. Our results also suggest that a fine-grained treatment of I^{2}PA based on different social platforms may be needed to combat data distribution mismatch.

Table 2: Pairwise accuracy as a function of simulated label noise level

Noise	0%	10%	20%	30%	40%
Accuracy (%)	76.65	75.61	74.35	73.10	68.55

Table 3: Pairwise accuracy as a function of network architecture

DNN	AlexNet	VGGNet	ResNet-50	ResNet-101
Accuracy (%)	73.22	76.15	76.65	76.87

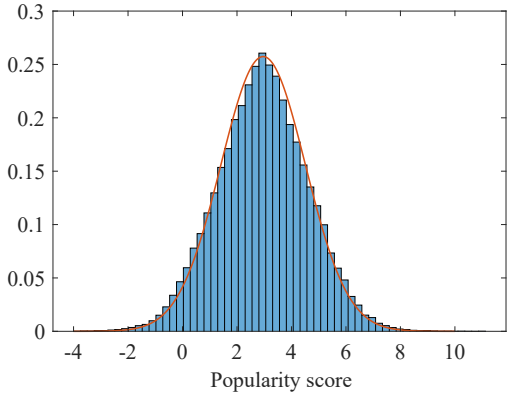
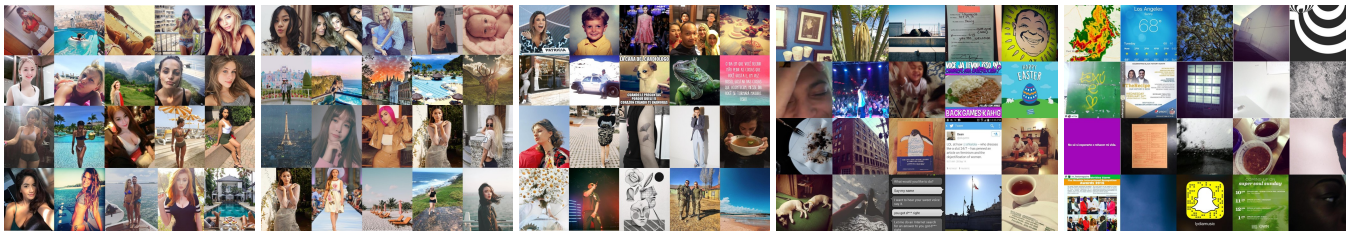


Figure 5: The normalized histogram of the predicted intrinsic popularity scores by our model on the test set with 50,000 PDIPs. Red line: fitted Gaussian curve.

Generalizability on Reddit and Flickr. To probe the generalizability of our model trained on Instagram, we test it on two other social platforms - Reddit and Flickr. The Reddit database [17] contains over 100,000 pairs of popular and unpopular images categorized by six sub-datasets. We choose the largest sub-dataset *pics* due to its diverse content variations. Our model achieves a pairwise accuracy of 58.9% (on 44,343 pairs), and is slightly worse than Hessel17 [17] (60.0%), which is trained on the same Reddit database.

Next, we test our model on images from Flickr. Due to the lack of intrinsic image popularity databases on Flickr, we decide to build a small one for testing. Specifically, we choose the social media headline prediction challenge database [42] as the starting point. The database contains over 340,000 posts from over 80,000 users. For simplicity, we select the most popular 50,000 images and the most unpopular 50,000 images according to the normalized number of views, and pair them randomly. When considering the visual content only, our model achieves a pairwise accuracy of 63.3%, and is slightly better than 62.4% of Khosla14 [20], which is trained on the same Flickr database.

As previously discussed, the performance drop of our Instagram-based model on Reddit and Flickr may be because the dataset distributions are different. In addition, without reducing the effect of non-visual factors, the test image pairs are much noisier.



(a) (b) (c) (d) (e)

Figure 6: Images with different predicted intrinsic popularity levels. Some images are resized without keeping aspect ratios for neat display. (a) Excellent. (b) Good. (c) Fair. (d) Bad. (e) Poor.

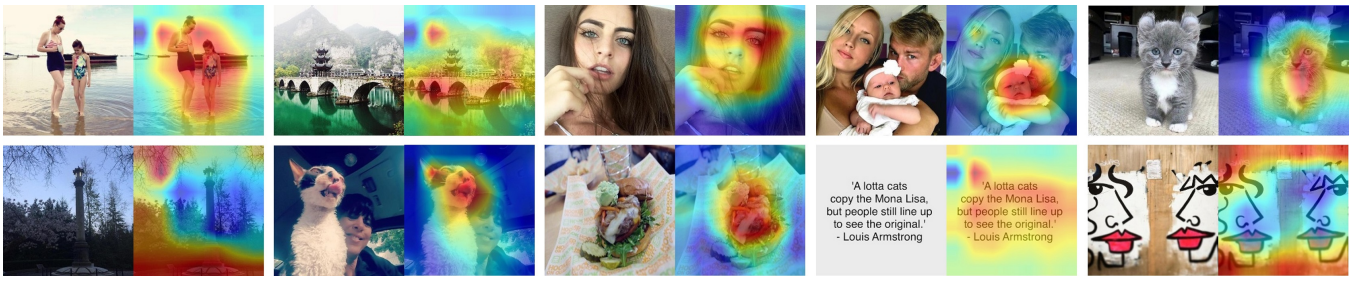


Figure 7: Heatmaps of sample images generated by Grad-CAM [37]. A warmer region contributes more to I^{2}PA. First row: images of high predicted popularity. Second row: images of low predicted popularity.

Ablation study. We provide a baseline for our model - a DNN trained for absolute IPA and used to assess intrinsic image popularity. Specifically, we first summarize the visual content by a scalar using the same ResNet-50 [16], and combine the image-based score with six non-visual features: the number of followers, followings, posts, hashtags, @ signs, and the length of caption. The seven-dimensional feature vector is fed to a fully connected neural network with a 7-256-128-64-1 structure to predict the number of likes. We train the entire model end-to-end on 4.8 million images used to generate training PDIPs. We adopt the mean squared error (MSE) as the loss function. The training procedure is the same as in Section 4.1. The optimized model predicts absolute image popularity reasonably, as evidenced by a Pearson correlation of 0.83 on the training set and 0.80 on the test set, respectively. However, when tested on the 50,000 PDIPs, this baseline model only achieves a pairwise accuracy of 66.5%. The performance drop may be because strong features (*e.g.*, the number of followers) tend to dominate the learning, leaving less room for exploiting the visual content.

We analyze the impact of T and σ in PDIP generation on the final performance. However, due to the limited computation and storage resources, we perform a similar but simulated label noise experiment. In particular, we select a percentage $q \in \{10\%, 20\%, 30\%, 40\%\}$ to randomly flip the binary labels of PDIPs in the training set and retrain the models, whose test results are listed in Table 2. It is clear that the training is robust to label noise. When the label noise level is 30%, our model is still competitive with humans. Although it seems impossible to eliminate the impact of non-visual factors in our current PDIP generation process, our experiment suggests that the induced label noise does not seem to hinder the learning of a robust model for I^{2}PA. In addition, we may relax the constraints of PDIP generation (*i.e.*, decrease T or increase σ) to obtain pairs with more diverse content.

We also investigate the impact of different DNN architectures on pairwise accuracy, including AlexNet [22], VGGNet [38], ResNet-50 [16] (default), and ResNet-101 [16]. From the results in Table 3, we find that there is still room for improvement on top of ResNet-50 if deeper and more advanced networks are adopted.

4.3 Qualitative analysis

We provide a qualitative analysis of I^{2}PA from three perspectives: global image content (Fig. 6), local image content (Fig. 7), and comparison with human data (Fig. 9). It should be noted that these results would become less obvious if we predict absolute image popularity and do not single out the contribution of image content.

We first exam the histogram of predicted intrinsic popularity scores by our model on the test set, which provides a good coverage of representative Instagram images. The histogram can be well fitted by a Gaussian distribution with mean 2.96 and std 1.55. A higher value indicates better intrinsic popularity (see Fig. 5).

To better analyze how global image content affects I^{2}PA, we define five popularity levels - "excellent", "good", "fair", "bad", and "poor" that evenly cover the predicted score range. Fig. 6 shows representative images of each level. We find that images in the excellent level are often beautiful and attractive people, which is in close agreement with Park and Lee's conclusion [35]. Images in the good level tend to be brilliant selfies and spectacular sceneries. The high-score selfies are often accompanied by beautiful faces, which is consistent with the result that photos with faces are 38% more

Table 4: Pairwise accuracy of different strategies and factors. Majority: performance obtained by majority vote. Single: performance obtained by averaging individual subjects. G_I: subjects with Instagram experience. G_{II}: subjects with little Instagram experience. G_A: subjects spending more than three hours on social media per day. G_B: subjects spending less than three hours on social media per day

Accuracy (%)	Majority	Single	Female	Male	G_I	G_{II}	G_A	G_B
Mean	72.4	66.6	67.0	66.8	68.5	65.1	67.1	66.4
Std	—	4.4	4.2	4.5	3.7	4.5	3.8	4.1

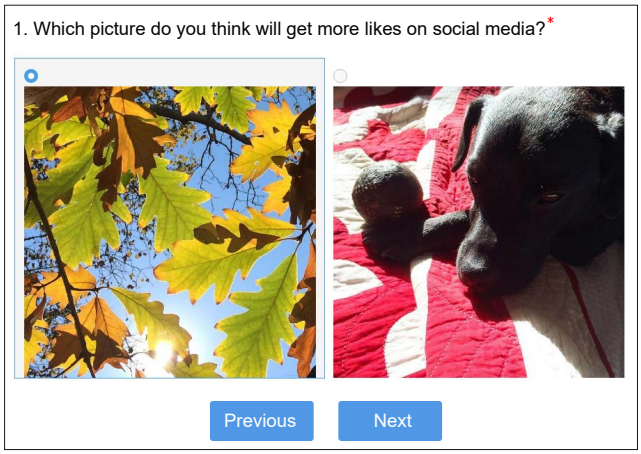


Figure 8: Web-based user interface for the psychophysical experiment.

likely to receive likes [2]. Images in the fair level look ordinary and forgettable, whose common characteristics are difficult to summarize because of the content diversity. Images in the bad level are less prominent, and may lack interesting and distinguishable features. Images in the poor level mainly consist of empty backgrounds with few salient objects.

We also investigate how local image content contributes to I^{2}PA. Specifically, we generate the heatmaps of sample images by Grad-CAM [3, 37], a visual explanation for deep networks via gradient-based localization. A warmer region in the heatmap indicates that it plays a more important role in I^{2}PA. From the first row of Fig. 7, we find that several elements such as fine architectures (the second image), pretty faces (the third image), lovely kids (the fourth image), and cute animals (the fifth image) are often activated, leading to high popularity predictions. By contrast, for images in the second row, poor quality regions (the first image), unsightly expressions (the second image), textual descriptions (the fourth image), and empty backgrounds (the fifth image) tend to dominate, leading to low popularity predictions.

4.4 Psychophysical experiment

To understand human behavior in I^{2}PA and to make the results more interpretable, we conduct a psychophysical experiment. Specifically, we randomly select 1,000 PDIPs from the test set, and invite 30 subjects to perform a two-alternative forced choice (2AFC) using a web-based platform (see Fig. 8). The subjects (14 females and 16 males) are all college students of age 20 to 30, among whom 40% have used Instagram before. Besides, 30% subjects spend more than three hours on various social platforms every day. At the beginning of the experiment, eight training pairs are displayed to help the subjects build the concept of intrinsic image popularity. After that, they are free to make decisions based on their own understanding of image popularity. To reduce the fatigue effect, the experiment is divided into four sessions, each of which is limited to a maximum of 30 minutes. The subjects are encouraged to participate in multiple sessions. In the end, the subjects are allowed to review and compare their choices with the ground truths.

Table 4 lists the subjective results, where we see that the majority vote strategy significantly outperforms individual subjects, reflecting the difficulty of this task for a single observer. We next analyze the influence of the gender, the Instagram experience, and the online social time per day on I^{2}PA. From table 4, we find that 1) both male and female subjects tend to perform at a similar level; 2) subjects with Instagram experience (denoted by G_I) perform statistically better (based on t-statistics [31]) than those with little knowledge about Instagram (denoted by G_{II}); 3) subjects who spend more than three hours on social media per day (denoted by G_A) perform statistically better than those with less online social experience (denoted by G_B).

We further compare the subjects' choices against our model predictions and the ground truths using four types of pairs, as shown in Fig. 9. Pairs (a) and (b) in Part I have clear popularity discriminability based on different visual appearances, leading to easy predictions for both subjects and our model. Most PDIPs in the proposed database belong to this category.

For the pair (c) in Part II, 67% of the subjects predict the food image to receive more likes than the house image. However, we find that the user has posted too many food images, of which the followers may get tired. The more likes received by the house image indicates that the followers are more interested in viewing images with novel content. This contextual interactions among posts complicate I^2 PA because humans and our model do not get access to such information. For the pair (d), many subjects pay too much attention to image aesthetics (*i.e.*, they think the left image is more beautiful), which often results in selection bias. On Instagram, brilliant selfies generally receive more likes, which has been successfully captured by our model.

The pairs in Part III are difficult for our model because extremely abstract attributes such as the peculiar gesture (the pair (e)) and the creative/funny content (the pair (f)) need to be parsed and transformed to the concept of popularity. By contrast, humans have a better understanding of these concepts, and are able to make consistent choices easily.

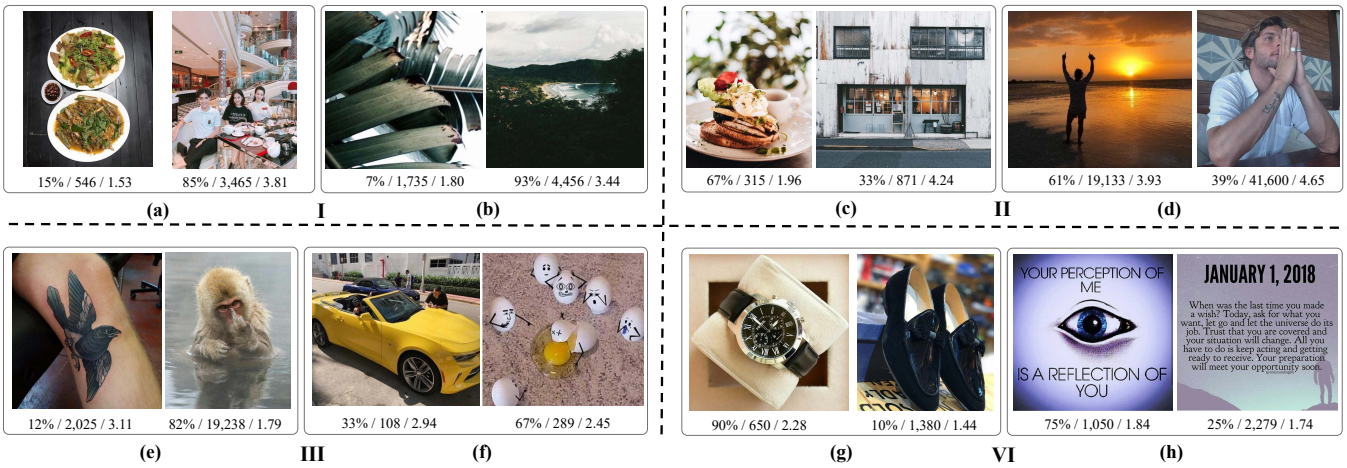


Figure 9: Eight representative PDIPs, within which the right image is intrinsically more popular than the left one. Below each image show the subject percentage in favor of the image, the number of received likes, and the popularity score predicted by our model, respectively.

For the pair (g) in Part VI, nearly all subjects prefer the elegant watch than the ordinary shoes, and our model agrees on this point. However, the shoes image receives more likes. We conjecture that the shoes may convey a special meaning (*e.g.*, as a memorable gift), of which the subjects in the psychophysical experiment are unaware. The number of likes may also be boosted by the Internet vendor for sales promotion. For the pair (h), the number of likes is mainly determined by the texts in the image, which are difficult for most subjects to comprehend due to the cultural differences. Our model also fails to understand the words, and tends to give text images low popularity scores.

5 CONCLUSION

We have conducted a systematic study of I^{2}PA. The principle behind I^{2}PA is to predict image popularity based on the image content only, and the concept of PDIP is introduced to reliably infer intrinsic popularity. The first large-scale image database for I^{2}PA is established, and a DNN-based computational model is further proposed, which achieves human-level performance. In addition, we have carried out a psychophysical experiment to understand how humans tend to behave in this task.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Zhuo (Jimmy) Wang for insightful discussions on the probabilistic formulation of I^{2}PA. This work is supported by Hong Kong RGC Early Career Scheme 9048122 (CityU 21211018) and City University of Hong Kong under Grant 7200539/CS, which are gratefully acknowledged.

REFERENCES

[1] Khaled Almgren, Jeongkyu Lee, and Minkyu Kim. 2016. Predicting the future popularity of images on social networks. In *Multidisciplinary International Social Networks Conference on Social Informatics, Data Science*. 1–6.
[2] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. 2014. Faces engage us: Photos with faces attract more likes and comments on Instagram. In *SIGCHI Conference on Human Factors in Computing Systems*. 965–974.
[3] Adam Bielski and Tomasz Trzcinski. 2018. Pay attention to virality: Understanding popularity of social media videos with the attention mechanism. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2335–2337.
[4] Patrick Biernacki and Dan Waldorf. 1981. Snowball sampling: Problems and techniques of chain referral sampling. *Sociological Methods & Research* 10, 2 (1981), 141–163.
[5] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *International Conference on Machine Learning*. 89–96.
[6] Ethem F Can, Hüseyin Oktay, and R Manmatha. 2013. Predicting retweet count using visual cues. In *ACM International Conference on Information & Knowledge Management*. 1481–1484.
[7] Spencer Cappallo, Thomas Mensink, and Cees GM Snoek. 2015. Latent factors of visual popularity prediction. In *ACM International Conference on Multimedia Retrieval*. 195–202.
[8] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 539–546.
[9] Josh Constine. 2018. Instagram hits 1 billion monthly users, up from 800M in September. https://techcrunch.com/2018/06/20/instagram-1-billion-users
[10] Marc Damashek. 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science* 267, 5199 (1995), 843–848.
[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.
[12] Arturo Deza and Devi Parikh. 2015. Understanding image virality. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1818–1826.
[13] Abhimanyu Dubey and Sumeet Agarwal. 2017. Modeling image virality with pairwise spatial transformer networks. In *ACM on Multimedia*. 663–671.
[14] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. 2015. Image popularity prediction in social media using sentiment and context features. In *ACM International Conference on Multimedia*. 907–910.
[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-Level performance on ImageNet classification. In *IEEE International Conference on Computer Vision*. 1026–1034.
[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[17] Jack Hessel, Lillian Lee, and David Mimno. 2017. Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. In *International Conference on World Wide Web*. 927–936.
[18] Chih-Chung Hsu, Ying-Chin Lee, Ping-En Lu, Shian-Shin Lu, Hsiao-Ting Lai, Chihg-Chu Huang, Chun Wang, Yang-Jiun Lin, and Weng-Tai Su. 2017. Social media prediction based on residual learning and random forest. In *ACM on Multimedia*. 1865–1870.
[19] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 133–142.

[20] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. What makes an image popular?. In *International Conference on World Wide Web*. 867–876.

[21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014).

[22] 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.

[23] Hang Li. 2011. Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies* 4, 1 (2011), 1–113.

[24] LikelyAI. 2017. Predict the popularity of your Instagram posts. https://www.likelyai.com

[25] Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval* 3, 3 (2009), 225–331.

[26] David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 2 (2004), 91–110.

[27] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. 2017. dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing* 26, 8 (2017), 3951–3964.

[28] Masoud Mazloom, Iliana Pappi, and Marcel Worring. 2018. Category specific post popularity prediction. In *International Conference on Multimedia Modeling*. 594–607.

[29] Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, and Willemijn Van Dolen. 2016. Multimodal popularity prediction of brand-related social media posts. In *ACM International Conference on Multimedia*. 197–201.

[30] Philip J McParlane, Yashar Moshfeghi, and Joemon M Jose. 2014. Nobody comes here anymore, it's too crowded; predicting image popularity on Flickr. In *ACM International Conference on Multimedia Retrieval*. 385–391.

[31] Douglas C Montgomery and George C Runger. 2013. *Applied Statistics and Probability for Engineers (6nd. ed.)*. Wiley, New York.

[32] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2408–2415.

[33] Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42, 3 (2001), 145–175.

[34] ParallelDots. 2018. Virality Detection. https://www.paralleldots.com/virality-detection

[35] Hyanghee Park and Joonhwan Lee. 2017. Do private and sexual pictures receive more likes on Instagram?. In *IEEE International Conference on Research and Innovation in Information Systems*. 1–6.

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

[37] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization.. In *International Conference on Computer Vision*. 618–626.

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

[39] Louis L. Thurstone. 1927. A law of comparative judgment. *Psychological Review* 34, 4 (1927), 273–286.

[40] Tomasz Trzciński and Przemysław Rokita. 2017. Predicting popularity of online videos using support vector regression. *IEEE Transactions on Multimedia* 19, 11 (2017), 2561–2570.

[41] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, and Tao Mei. 2016. Time matters: Multi-scale temporalization of social media popularity. In *ACM Multimedia Conference*. 1336–1344.

[42] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Huang Qiushi, Li Jintao, and Tao Mei. 2017. Sequential prediction of social media popularity with deep temporal context networks. In *International Joint Conference on Artificial Intelligence*.

[43] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition.. In *AAAI Conference on Artificial Intelligence*. 272–278.

[44] Kota Yamaguchi, Tamara L Berg, and Luis E Ortiz. 2014. Chic or social: Visual popularity analysis in online fashion networks. In *ACM International Conference on Multimedia*. 773–776.

[45] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. 2018. User-guided hierarchical attention network for multi-modal social image popularity prediction. In *International Conference on World Wide Web*. 1277–1286.

[46] Zhongping Zhang, Tianlang Chen, Zheng Zhou, Jiaxin Li, and Jiebo Luo. 2018. How to become Instagram famous: Post popularity prediction with dual-attention. In *IEEE International Conference on Big Data*. 2383–2392.