Instagram Popularity Prediction
via
Neural Networks and Regression Analysis

Crystal J. Qian
cqian@princeton.edu

Jonathan D. Tang
jontang@princeton.edu

Matthew A. Penza
mpenza@cs.princeton.edu

Christopher M. Ferri
cferri@princeton.edu

Abstract

With over 700 million active users sharing content on Instagram, predicting the popularity of posts has attractive applications in social analysis and engineering. Previous research has determined that image content and social metadata have dominant predictive power; but to what extent does image composition (aesthetic value) impact popularity? We introduce a novel dataset of $3,411$ posts labelled as scenery for content-neutral social media analysis. To evaluate the predictive power of image composition on Instagram posts, we compare the popularity predictions of a neural network trained on aesthetic value to the predictions of regression models using social metadata.

1 Introduction

As social networks and content-sharing continue to grow rapidly in size and volume, predicting the popularity of social media content has become an important social problem with many potential applications. Successful popularity analysis can directly affect fields such as targeted advertisement, political strategy, and social engineering.

What factors determine an image's popularity, and how influential is each factor, specifically in the context of social networks? We divide a posted image's information into two parts; the

1. **Visual data**: the pixel composition of an image.
2. **Social data**: metadata image tags, such as the location, comment count, number of hashtags, and number of other media posts by the image's poster.

By using convolutional neural networks (CNNs) to predict popularity based on visual data and regression methods to predict popularity based on social data separately, we can quantify the influence of aesthetics on popularity. If a popular Instagram user posts with image metadata tailored specifically to her follower base, does the content of the post matter? If a well composed image (that is, an image whose composition style tends to result in higher popularity) does not take advantage of social engineering techniques, will its popularity be stunted?

Existing research on image popularity,[**?**] including studies on Instagram's topical interests,[**?**][**?**] suggest that the subject or image tags of an image contribute strongly to the popularity of an image. For example, images tagged with terms such as ”sexy legs”, ”women”, etc. receive significantly more attention than images containing ”sad clowns”.[**?**][**?**] To remove this topical bias and focus solely on the effect of aesthetic composition, we curated a set of $3,411$ images of user-identified landscapes from Instagram. Landscape images tend to be evaluated neutrally and consistently based on pure aesthetic beauty, as opposed to images of people, animals, and text, which may be evaluated subjectively and inconsistently based on personal recognition, humor, cuteness, etc.

With over 700 million monthly active users on mobile and web platforms,[?] Instagram is a leading photo-sharing social networkm. the majority of previous research in popularity prediction used images from the social network Flickr, but its monthly usage of only 17.5 million unique users[?] makes it a less robust sampling of posts. Additionally, while existing Flickr datasets[?] are large in size, the density of tagged/labelled images is relatively sparse. Instagram's user hashtags and content volume allowed us quickly to create a tagged image set of substantial size.

For this paper, we curated a dataset of $3,411$ landscape-labelled images and social metadata scraped and filtered from Instagram. The principle dataset used in popularity prediction, MIRFLICKR, contains 1 million Flickr images, but the number of labelled photos remains small (only 385 images labelled "landscape", 331 as "trees", 199 with "lake", etc.).[?] To our knowledge, our #scenery_lovers dataset is the largest social image database with a shared label. We also quantified the predictive power of aesthetic value and image composition on the popularity of landscape-labelled images.

2 Background

Online content popularity prediction has been the subject of substantial research over the last several years, mainly using social analysis rather than image composition analysis.

McParlane et al. published prediction methods tailed towards social image posts on Flickr with "no, or limited textual/interaction data", focusing specifically on image context, visual appearance, and user context.[?] Features analyzed included season of posting, device, use of flash, number of faces, number of contacts on Flickr, etc.

Our analysis falls more in line with that of Khosla et al., which expanded on McParlane et al.'s work by dividing analysis into image content and social context, similarly to our current approach.[?] Khosla et al.'s variant of improving popularity prediction focused on identifying the relative popularities of portions of images to determine which regions of given image helped boost popularity, whereas we evaluate popularity broadly across a wide set of images.

Gelli et al. showed that social features have more predictive power than visual content, but also that visual content does have limited predictive power when no user metadata is provided.[?]

We contribute the following variants/improvements to this research space:

- We use a larger and more active social network, and present findings more relevant to the current social media scene.
- By curating a substantial dataset of images with the same aesthetic theme, i.e. landscapes, we can analyze the success of image composition with minimal content bias.
- We make additional improvements/changes to the means of classification, updating visual analysis with the use of more complex CNN classifiers and applying various regression models for social analysis.

3 Data

Our dataset consists of $3,411$ images of landscapes downloaded from Instagram via a customized web scraper that we developed for this project, based on some tools from a pre-existing Python-based Instagram downloader; these existing tools were especially helpful because of limited available documentation.[?] Our scraper issues requests directly to Instagram's web server using GraphQL and processes the JSON responses, without the need of an Instagram API key, saving the image metadata in a csv file. Our web scraper was limited by the Instagram server to a few hundred requests every 5-10 minutes.

We cleaned up our dataset by filtering through the following specifications:

1. **Hashtag filtering**: We scraped images with the public Instagram hashtag #scenery_lovers, which contains around $42,000$ images as of May 2017. Because we wanted to eliminate

social bias by choosing solely pictures of landscapes, we exploited this hashtag to obtain a large set of user-labelled data. Furthermore, the #scenery_lovers niche of Instagram is frequented by users that tend to take aesthetic and photographic composition seriously, so the dataset is fairly homogenous in terms of aesthetic quality.

2. **Popularity sampling**: We inspected the top $11,492$ most popular images with this hashtag based on Instagram's image ranking algorithm, which sorts images by their level of engagement in terms of likes and user comments.
3. **Size filtering**: Of these, we chose the $3,783$ images that were square (which revealed the insight that the majority of #scenery_lovers images on Instagram are non-square). The images in our dataset needed to have the same aspect ratio for consistency and ease of configuration as we use the image matrices in our CNN. We observed that a significant portion of these square photos actually had user-added borders with white padding, which allow landscape-oriented photos to be presented in the shape of a square. Because adding this border is a user's conscious decision, we can use this sample to determine (from solely visual analysis) if orientation affects image popularity.
4. **Manual noise filtering**: We manually removed another 370 "noisy" images from our dataset: images that prominently feature animals or people, images that include prominent text or in-image captions, collages of multiple pictures, and other non-landscape images.

Figure 1: Images posted to Instagram with the #scenery_lovers hashtag. *Left*: image not filtered due to non-square constraints, because the user intentionally added a white border to post the landscape-oriented photo. *Middle*: "noisy" image filtered due to animal bias. *Right*: "noisy" image filtered due to text bias.

The $3,411$ images are then split randomly, with 70 percent used as a training set and 30 percent used as a test set; this split is consistently used in both the visual and social analyses.

4 Approach

We define the *popularity* of an image to be the likefollower ratio LTF, i.e. the number of likes on an image divided by the number of followers of the image's poster. LTF values can vary quite widely in this dataset, with the top three largest LTFs being 7.733, 6.333, and 4.777, corresponding to photos that received significantly more likes on the author's image than the number of people following the author, while the three lowest LTFs are 0.00128, 0.00577, and 0.00611, corresponding to photos that were particularly unpopular and posted by authors with larger numbers of followers.

Because the linear-fit correlation between the number of likes on an image and the number of followers on its author is strong ($R = .78$, $R^2 = .61$), we want to normalize popularity so that it is not biased towards users of Instagram with high follower counts. If we were to use a metric like the view count or the amount of interaction (likes and comments), users with few followers would be put at a disadvantage, since their photos would be displayed to fewer people.[**?**][**?**] We are interested in calculating *relative* popularity, i.e. the success of an image in comparison to the size of its social network, which reduces biases from the image poster's popularity.

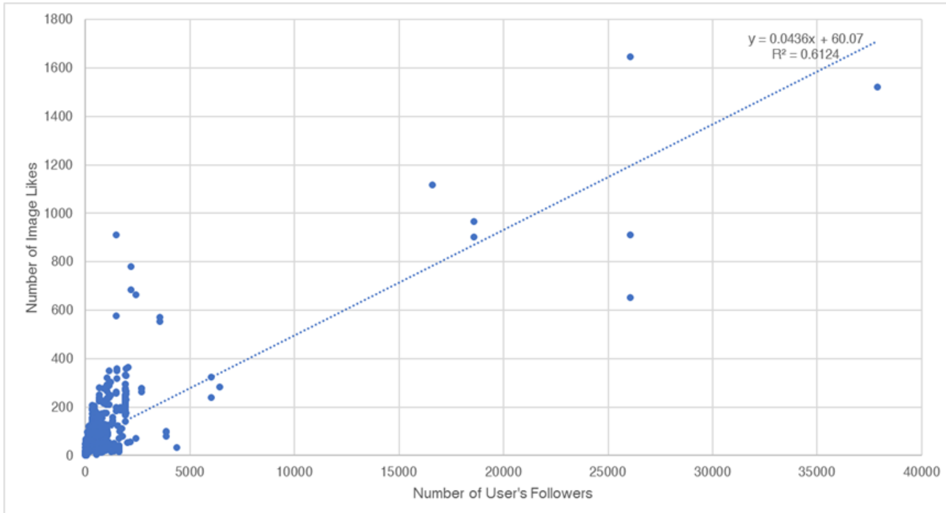

Figure 2: Number of image poster's followers v. number of image likes.

4.1 Social Analysis: Metadata

We predict the LTF of an image given a prior of 8 metadata features:

1. The **side length** (i.e. height and width) of the square image. Images with greater side length could correspond to prettier, higher quality, and hence more popular images.
2. A **location** ID that maps Instagram's geotag data to integers. A frequently used location ID could indicate more frequented locations, which could indicate greater popularity.
3. The **comment count** on the image; more comments indicates more user interaction with the image, which could correspond to more likes and therefore greater popularity.
4. The **time elapsed** from when the image was posted to the time our web scraper downloaded it. It could be the case that shorter time elapsed means less popularity, since people would have less time to interact with and like that image. It could also conversely be the case that popularity begins to wane after a certain length of time.
5. The **hashtag count** in the image caption. More hashtags means that the image is exposed, publicly, to more people outside of the poster's follower base, increasing popularity. Alternatively, too many hashtags could indicate a spam post, which could signal lower popularity.
6. The **caption length** of the image, i.e. the number of characters in the caption, a feature which overlaps to a small degree with the number of hashtags. Effective image descriptions tend to bias a viewer into liking an image.
7. The **author following**, or number of people followed by the image's poster. A user who follows a large number of accounts is more likely to be a spammer, which would typically indicate less popular images.
8. The **media count**, i.e. the total number of images and videos posted by the user. A high media count could indicate a prolific Instagram user, which might indicate lower popularity, as the user's followers would tend to have their attention distributed over many different photos, rather than concentrated on one photo.

The prior used to determine LTF does not use the image's composition or aesthetics; if we can predict an image's popularity reasonably well given only this data, it could indicate that image aesthetics or composition have little impact on image popularity. We perform linear and ridge regression on a training set of these priors and corresponding LTF as a continuous class, using resulting classifiers to predict popularity on the test data.

4.2 Visual Analysis: Image Features

We again predict the LTF of an image, this time given the pixel composition of the image itself. Predictions with error rates indicate correlation of image popularity and aesthetic quality. Previous studies show that the image composition itself is not as strong as a predictor for image popularity as social features.[**?**][**?**] However, these studies use old neural networks, such as Alexnet, which consists of only 8 layers with low-level image features like color, GIST, and content features.

4.3 Spotlight Method: Neural Networks

We use convolutional neural networks (CNNs) to make the predictions based solely on image composition. CNNs are one of the most powerful tools currently in use in the field of computer vision, and have revolutionized the field of image recognition. State-of-the-art neural networks perform even better than humans on classification in the ImageNet competition.[**?**]

A neural network consists of multiple neurons stacked in layers. In a simple feed-forward network, each neuron receives the output of every neuron in the previous layer. The neuron contains a weight variable w_i for each input to the current neuron. It takes this weight value and multiplies it with the respective input, and sums over all of the multiplied values. Finally, a bias term b is added and the final value is put into an activation function f. The most common activation function, called ReLu, is simply $f(x) = \max(0, x)$. Putting this all together, a general neuron comes in the form of Equation 1.

$$y = f\left(\sum_{i=0}^{n} w_i x_i + b\right) \tag{1}$$

From this equation, we are able to reduce an entire layer of a network into a matrix W, the bias into a vector $\vec{b}$, and the input into a vector $\vec{x}$. This means that one layer of a neural network can be expressed as Equation 2.

$$\vec{y} = f\left(\mathbf{W}\vec{x} + \vec{b}\right) \tag{2}$$

To train a feed-forward network, we must learn the values of the the W matrices and $\vec{b}$ vectors. To do this, we use gradient descent which optimizes the weights of the latent variables. We calculate the gradient at the output layer, and from there, we use an algorithm called backpropagation to find the gradients for each individual layer. The algorithm relies on the derivative chain rule to calculate the gradient of a specific layer based on the gradient of the layer after it. This can be seen in Equation 3.

$$\delta_j^{l-1} = \left(\sum_{i=1}^{n_l} \delta_i^l W_{ij}^l\right) f'\left(\vec{u}^l\right) \tag{3a}$$

$$\Delta \mathbf{W}_{ij}^l = \eta \delta_j^l x_i^{l-1} \tag{3b}$$

$$\Delta b_i^l = \eta \delta_j^l \tag{3c}$$

From this equation, we can start at the last layer L and slowly update the layers by moving backwards through the feed-forward network. Once again, this equation is usually represented in matrix-vector form, allowing backpropagation to be represented by a large number of matrix multiplications. This is also the reason why neural networks are known as one of the most computationally intensive machine learning algorithms: millions upon millions of large matrix multiplications are required during the training phase. Consequently, graphics processing units (GPUs) are regularly used for hardware acceleration in this process.

In an effort to improve time performance and accuracy, CNNs were first proposed by Yann LeCunn with the creation of the LeNet neural network.[**?**] Convolution is a complex mathematical process that allows us to use kernels, which are learned in the training process. The benefit of this approach is that networks significantly reduce the number of connections per neuron.

For our project, we will be using a specific neural network model called Inception-v3.[**?**] Developed by Google, this is a state-of-the-art model that consists of over 30 layers and multiple paths through the network, allowing it to achieve high accuracy on classification on the Imagenet dataset. However, even on a state-of-the-art, multi-GPU server, this model takes a week to train, so we use the transfer learning technique to only retrain the output layer.

5 Results

5.1 Social Analysis

Figure 3 below enumerates the weights of the features used in our regression analysis. A positive weight for a continuous feature correlates to an increase in predicted popularity, Weights of higher absolute value have greater predictive power.

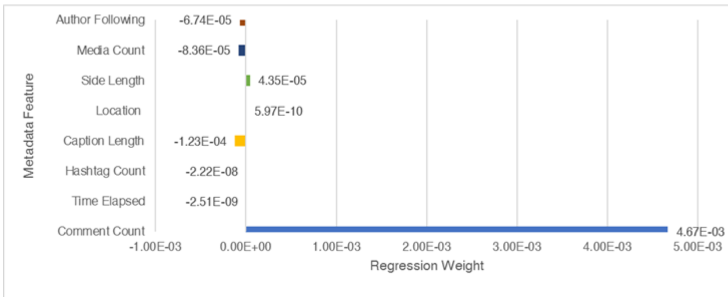

Figure 3: Regression Weights of Metadata Features

The linear and ridge regression methods converged to the same coefficient weights. We calculate the mean-squared error (MSE), the mean of the squared values of the errors (with the ideal MSE being 0.0), which gives us a fairly objective evaluation metric for continuous-valued outcomes. We achieved an **MSE** of 0.0596 (normalized to 0.0573) and an **RMSE** (root mean squared error) of 0.2442 (normalized to 0.0913), indicating that social metadata alone can predict popularity accuracy extremely well. For example, an MSE of 0.0596 means that, for a user with 500 followers, our predictions would tend to be within $0.0596 \cdot 500 = 29.8$ likes of the true value.

5.2 Visual Analysis

We trained the Inception-v3 final layer in approximately 10 minutes on a total of 16 buckets of LTF value ranges. However, the model exhibited extreme overfitting, as shown by training loss decrease and validation loss increase, and was only able to achieve an accuracy of 11 percent at best.

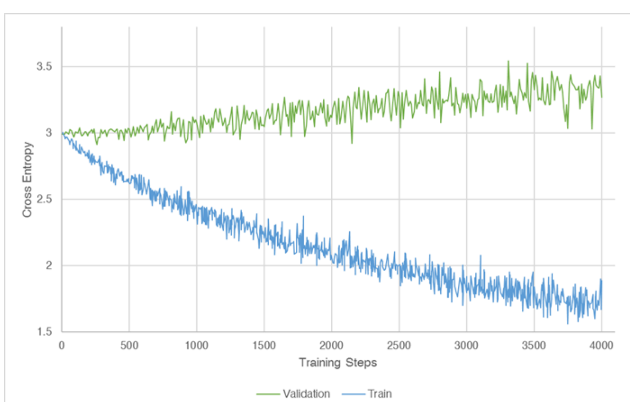

Figure 4: Cross Entropy Loss of Inception-V3 Model

6 Conclusions and Future Work

6.1 Social Analysis

Comment count has the strongest predictive power, corresponding positively to popularity. This result makes sense; since more comments indicate grater user interaction, and users cannot downvote or ”dislike” posts, more interaction tends to indicate more positive feedback. A greater side length also corresponds positively to popularity, likely because larger images tend to be of higher quality.

As expected, a large caption length, media count, and author following indicate less popularity, as this behavior is typical of spammers, as well as new users who are often ”desperate” for likes. As an example, figure 5.2 is from an author with 3895 followers, but only yielded 112 likes! The caption with length 775 is as follows:

“Early morning drama for you ... I’m addicted to my early morning power runs now even though you would just laugh your head off at me as I can just barely run down the road and then I have to just stop!!!”

The time elapsed seems to have little impact on the popularity prediction, which matches research indicating that after a rapidly-growing initial seed period, social media posts tend to ”settle down” and stop accumulating likes.[**?**]

Figure 5: *Left*: 5.1, the image with the highest LTF in the dataset, 7.7333, accumulating 116 by an author with 15 followers. *Middle left*: 5.2, an image with a low $LTF = 0.0288$ score. *Middle right*: 5.3, a image with low $LTF = 0.2036$ accurately predicted with .0001 error, calculated as the absolute value of the subtraction of the real LTF from the predicted LTF. *Right*: 5.4, the image that was least accurately predicted, with high $LTF = 4.0$ and error 3.7063.

The photo referenced in figure 5.3 was accurately predicted because it was moderately large in the following features: comments count, caption length, width; and quite large in media count and number of users followed by the author. Overall, this photo had more attributes associated with less-popular images and indeed ended up being less popular, so our prediction was accurate.

Conversely, the photo referenced in figure 5.4 had a low comment score and media count, and moderately large caption length, width, and number of users followed by the author. These are qualities that tend to be associated with less-popular images, but the photo actually ended up being extremely popular, with a very high LTF of 4.0. Because of this major discrepancy in attributes, the prediction error for this photo is very high. The photo’s LTF value seemed to be somewhat skewed by the small number of people following the author, which was 9, so that a relatively small number of likes, 36, ended up having a fairly large impact.

Other social analysis on image popularity evaluate captions and hashtags in greater detail through lexicographic analysis, perhaps by interpreting tags as a bag-of-words.[**?**] We only added numeric metadata about captions and hashtags since our metadata prediction was creately mainly to compare against image-based prediction, but additional training based on caption/hashtag content could increase our prediction accuracy.

Because comment counts could also be an evaluative metric for popularity, running a regression without this feature in the prior would yield results that are more robust to other interpretations of popularity (not just the LTF). We could even redo these experiments with popularity defined as the sum of the like count and comment count, divided by the number of the poster’s followers.

For future work, it would be interesting to extract additional Instagram metadata to use as regression features; for example, Instagram internally records the image filter, as well as color and contrast adjustment settings, of every photo post, but this metadata seems to be unreachable from the web endpoint that we currently use in our scraper. Getting our hands on this data would make our analysis more robust and also produce some interesting results, allowing us, for example, to figure out which Instagram filters and color settings tend to be most successful. In addition, extracting features from the Instagram photos using an external image recognition library, such as the color distribution and number of edges detected, could likewise improve our regression analysis.

6.2 Visual Analysis

The CNN was generally unable to act as a reliable predictor of image popularity. We suspect that due to the uniform aesthetic subculture of the #scenery_lovers Instagram community and careful filtering done when curating our dataset, we collected a fairly uniform set of aesthetically pleasing images, all of similar content. The consequence of this homogeneity is twofold: first, training a neural network to recognize subtle aesthetic differences becomes more difficult, and second, often significant differences in metadata between photos would provide more differentiating information than the visual data. Unlike other applications of the Inception-v3 model, we are trying to teach the model not to classify discrete objects, but the aesthetic of the image. For example, Inception-v3 might perform well when classifying the concrete categories of cars versus pedestrians, but seems to perform badly for the general task of aesthetic photographic evaluation. It is not immediately obvious, even to a trained human observer, whether a particular photo in our dataset should be more popular than another.

Retraining all the network's layers as opposed to applying transfer learning would likely improve our results. However, to do so would be prohibitively costly, as it would take multiple days, if not weeks to train the full model. If we were to train the network through all layers, we could build image kernels that are much better suited to the task of judging images aesthetically, rather than identifying objects.

6.3 Overall Thoughts

We chose landscape-themed images for our dataset to eliminate bias due to image content and train our classifier on the aesthetic quality of the image alone. This backfired, in a way; the visual homogeneity of the content resulted in low predictive capabilities for the neural network classifier. Retraining the classifier could improve this result, but this could also suggest that aesthetic quality is too subjective to correlate with popularity. It could also verify existing research that popularity is largely dependent on the image content.[**?**][**?**]

We could combine our neural network and regression methods into a single ensemble method. Using both approaches simultaneously would allow the classifier to utilize both the image content and metadata of a post. If the resulting classifier performed better than our metadata-based regression approach, it would indicate that our quantification of aesthetic value and image composition does indeed correspond to improved popularity predictions, thereby expanding upon Gelli et al.'s research, which showed that visual content has nonzero predictive powers *only* when no user metadata is present.[**?**]

In summary, we scraped and curated a dataset of labelled images with social metadata for popularity prediction. Additionally, we determined that the aesthetic value of images of the same label (in this case, landscape photos) has minimal predictive power. If one were to post a landscape-themed image on Instagram with the #scenery_lovers hashtag, the resulting popularity depends largely on image metadata (number of hashtags, resolution of the image, etc.). While this result needs to be verified in greater detail by improving our classification approaches, we have contributed data showing that aesthetic value cannot reliably predict popularity; that is, it is highly subjective and non-uniform among similarly labelled landscape images.