Will it play in Peoria? Predicting Image Popularity with Convolutional Neural Networks

Nicholas Dufour
Stanford University
450 Serra Mall, Stanford, CA
ndufour@stanford.edu

Abstract

We present the results of several experiments using convolutional neural networks (CNNs) to predict image popularity as measured by the number of views, 'favorites' and comments generated by the image. Images were derived from the Flickr 100M *dataset. The CNN was obtained from a previous study examining the classification of* Flickr *images by their 'style.' Several attempts to predict popularity are presented, with mixed success. An analysis of why this approach has failed thus far is put forth along with an outline of ongoing and future projects.*

1. Introduction

Automated assessment of image popularity is an attractive prospect for obvious reasons, both for the potential insights it would shed on human image perception as well as providing the ability to automatically filter uninteresting or otherwise unappealing content. Little is known about what makes an image popular; in part because it is clear that the predictors of popularity are not entirely encoded in the features of the image—or video, song, tweet, or whatever cultural item we consider. This is compounded by the *phenomenon of virality*, which is in part characterized by a brief but incredible increase in popularity of a cultural item. The phenomenon is extraordinarily complex; and has been attributed to mechanisms as diverse as spontaneous symmetry breaking to the existence of information that exploits fundamental properties of the brain. The superpopularity of these images is very unlikely to be due to the image alone—while some images may be, by some metric, tens or hundreds of thousands times more popular, this is not because the image is tens or hundreds of thousands of times 'better.'

Nonetheless it's clear, at least trivially, that it cannot be due entirely to exogenous features (i.e., features that are not contingent on the image itself). As proof, we offer the fact that the majority of imagespace is images of noise, yet very few images of noise are widely popular. Furthermore, images that are objectively aesthetic or attractive—either due to composition or content—are plainly more likely to be popular.

Thus far predicting popularity has largely defied formalization and automation (*see* Related Work). However the recent development of sophisticated machine vision algorithms which do not rely on hand-curated features—namely CNNs—offer an opportunity for unprecedented predictive power by more closely replicating the mechanism underlying human visual cognition while also not relying on human guesswork in the form of feature curation. Here, we leverage the power of CNNs to determine the degree of popularity of variance that can be accounted for by image content alone—if any—and explore what the potential predictors are.

2. Related Work

Predicting popularity has remained difficult despite advances in compute vision, potentially due to its percieved intractability. Numerous research programs have skirted its borders, but none (to the author's awareness) have done so as directly as the work presented in this paper. The bulk of such work has focused on images *aesthetics* or *quality*, that is, the percieved beauty of an image apart from its context. Among these, most have used used human-devised features [3, 5] and many focus on one specific aspect of image quality, such as camera distortion [11].

The field has not remained completely immune to the deep learning revolution. Recent work has used CNNs to assess image aesthetics, although still rely on human devised features [10]. Others have applied CNNs directly to the image themselves, but restricted themselves to curated images sets where images are annotated in terms of low-level features of the image that are believed to related to overall image quality [4].

Some have succeeded in predicting abstract features of an image (image 'style') over a largely unrestricted image domain by using a deepnet [8]. However, this work differs in two important ways: it relies on *post hoc* human annotation and treats the problem as one of classification rather than explicit prediction of a continuous value.

An large body of work has been driven by the availability of the Aesthetic Visual Analysis database [12], which has been used to automate prediction of image aesthetics [1] and memorability [6] and even the beauty of paintings [9], as well as the interestingness [7] and aesthetics [2, 13] of video data, most of which rely on human-crafted features and 'classical' learning algorithms. Further, none of these directly address popularity and all attempt to predict aggregated ratings of individual humans.

3. Data

Our independent variables consisted of images drawn from an online source, while our dependent measures of popularity were drawn from social media quantities.

3.1. Images

Images were obtained from Flickr, a widely used social photo-sharing website. Flickr made a large amount of data recently available (the 'Flickr 100M' dataset), which consists of metadata for some 90 million images and 10 million videos, such as title, tags (user generated content-related keywords) and photo acquisition data. To constrain the space of images somewhat, images were only drawn from those that included at least one of the top 100 tags, ordered by their frequency of use.

Images were down or upsampled such that the smallest dimension (horizontal or vertical) was 256 pixels, and then center-cropped to 256-by-256 pixels. Before being transformed into input for the deepnets, they underwent random cropping to 227-by-227 pixels and randomly flipped horizontally. In total, over 400,000 images were collected.

3.2. Popularity Metrics

Image popularity was assessed by three primary measures: the by-image number of views, the number of 'favorites,' and the number of comments. Views refer simply to the number of times an image was accessed, at any time and by anyone, since it was first uploaded. 'Favorites' quantify the number of times a Flickr user elected to favorite an image, thereby adding it to a collection of favorite images that the user curates. Flickr users may engage in discussion over an image, the amount of which is captured by the 'comments' quantity.

It is unclear which popularity metric best relates the popularity of an image. Views is the rawest measure; but is not restricted to Flickr users, and hence is the most subject to potentially-confounding viral phenomena through hotlinking and external sharing, if virality is viewed as a combined function of intrinsic popularity and the luck-of-the-draw, so to speak. On average, images in our dataset were viewed 146 times, while the most popular image was viewed more than 320,000 times (*see* Figure 1).

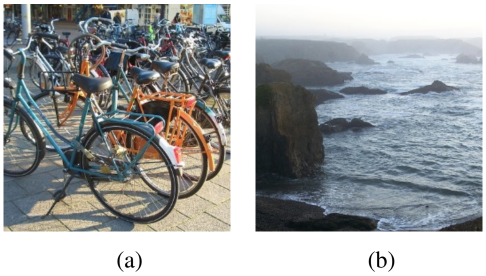
(a) (b)

Figure 1: Comparison of an average-popularity image with 200 views (a) and a super-popular image with over 320,000 views (b)

The other two measures come with orthogonal difficulties. Favorites and comments are far sparser than views. While the majority of images have some number of views, only one in ten images has a favorite or a comment, which makes them potentially more difficult targets for automated learning due to their relative sparsity. In terms of robustness to virality, comments appear to be—at least initially—the most attractive: they are not open to non-users of Flickr, and require the most work to create. However they are actually *more* frequent than favorites, which simply requires the push of a button. The three measures of popularity are summarized in Table 1. The covariance structure of the metrics, while not directly related, is of sufficient interest to bear mention. The number of favorites was predictive of the number of comments and the number of views, but views and comments were not highly predictive of each other. Data are summarized in Table 2. This finding was unexpected, as all measures were expected to vary in correlation with each other—as we consider them all measures of popularity. It is possible the capture both popularity and the degree to which the image polarizes opinion, to varying

	Mean	STD	Max
Views	146.37	323.58	321,528
Favorites	0.32	2.21	1,533
Comments	0.50	5.51	1,175

Table 1: Summary of popularity metrics

	Views	Favorites	Comments
Views	—	0.44	0.12
Favorites	—	—	0.44

Table 2: Covariance structure among popularity metrics

degrees, with popularity generating views and ‘polarizability’ driving comments, with favorites capturing aspects of both. Although this explanation is purely speculative and not entirely satisfying in the intuitive sense.

4. Methods

A number of experiments were carried out to determine if deepnets could learn the features of images that predict their popularity and, as a function of this, determine if this information is encoded in images at all.

4.1. Predicting Popularity as Regression

The most straightforward way of accomplishing the popularity prediction task is as a problem of regression. The goal of this is to predict, given an image, the value of some measure of popularity. The number of views was selected, as this metric was able to distinguish the largest number of images (recall that the majority of images had zero comments or favorites). As discussed, views was subject to ‘viral phenomena’ and accordingly exhibited a distribution that was strongly governed by a power law (*see* Figure 2).

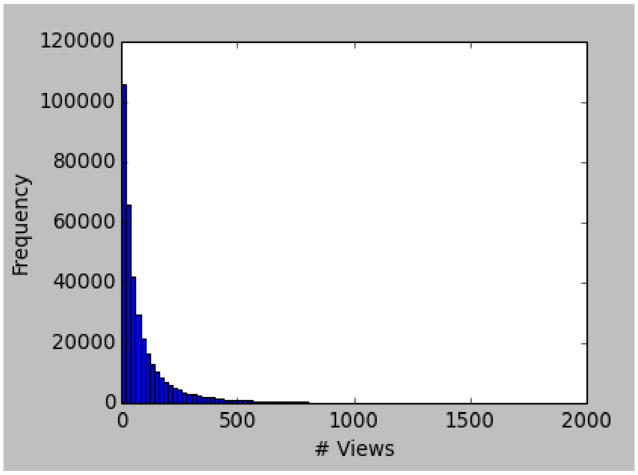

Figure 2: Histogram of image views depicting a clear Power Law relationship (images with more than 2,000 views are not shown)

This is to be expected, as the distribution of popularity of any type of item or concept almost always follow a power law distribution. Having a small number of training items that have such enormous values for the quantity being predicted makes training extremely difficult; they act like land mines in the training process, causing the loss to temporarily explode and derailing gradient descent. To avoid this, any image whose popularity metrics—views, favorites or comments—was more than six standard deviations above the mean was removed, resulting in some 500 images being excluded from analysis.

Regression was performed by a pretrained deepnet designed for image ‘style’ classification (also using Flickr images) [8], consisting of five convolutional layers followed by three fully connected layers with 0.5 dropout. The network is fully described by:

1. CONV 11x11, 4 stride - ReLU - Pool 3x3, 2 stride - Norm
2. CONV 5x5, 2 pad - ReLU - Pool 3x3, 2 stride - Norm
3. CONV 3x3 - ReLU
4. CONV 3x3 - ReLU
5. CONV 3x3, 2 stride, ReLU, Pool 3x3, 2 stride
6. FC - ReLU
7. FC - ReLU
8. FC
9. Euclidean Loss Layer

The net was run for 20,000 iterations, a little over 4 times over the dataset, with a base learning rate of 1×10^{-8}. The final fully connected layer had a learning rate multiplier of 20, all other layers had a multiplier of 1 (since it had been pretrained).

4.2. Predicting Popularity as Classification

The purely regression-based approach was potentially problematic for several reasons: it is heavily subject to ‘viral phenomena,’ it has a power-law distribution, and there exists a distinction between the popularity due to the intrinsic properties of an image (i.e., the relative aesthetics) versus extrinsic properties (i.e., whether or not the image was taken by a famous individual, or shared at the right time, etc). Thus, predicting views is not only difficult but may only measure a proxy of popularity, at least in terms of the kind of features we’re trying to learn.

To address these issues, a second approach was taken in which the image favorites were used. Favorites—as mentioned—is restricted to users of Flickr, and requires more ’work’ so to speak, which may reduce the system noise. Images were considered of class 0 if they were never favorited, and of class 1 if they had been favorited at least once before. Favorites were substantially more selective than views, of the 393,691 images analyzed, only 43,071 had been favorited at least once. The same neural net from 4.1 was used, although a softmax loss layer was used in lieu of a euclidean loss layer.

4.3. Predicting Popularity by Hybridizing Regression and Classification

The last experiment conducted attempted to fuse regression and classification. To eliminate the ‘power-law problem,’ data were divided into 10 bins where each bin contained an equal number of images, where the 0^{th} bin contained images with the lowest number of views while the 9^{th} bin contained images with the highest number of views.

Purely classifying images as belonging to one of these bins is of course a perfectly valid task, although it discards critical information. For instance, misclassifying a 2 as a 3 is far less grievous than misclassifying a 1 as an 8, though a softmax layer will treat both mistakes with equal seriousness. Thus, while all images were labeled, the euclidean loss layer was introduced once again, to differentially penalize ‘misclassification.’

5. Results

5.1. Experiment 1: Regression

The deepnet was run on the data for 6,000 iterations, covering the entire dataset roughly one and a half times. Loss was very high, with a mean training error of 49,992 for the final 1,000 iterations. Although, this is substantially better than the expected error of approximately 87,000. The best validation accuracy achieved was 50,090.

5.2. Experiment 2: Classification

The deepnet which classified the images as either having been favorited at least once or never favorited ran for 86,000 iterations, although the primary gains in training accuracy were achieved in the first third of training (*see* Figure 3).

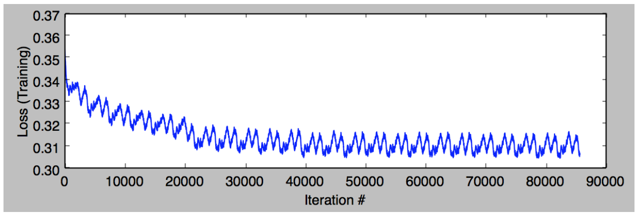

Figure 3: Training loss from experiment 2. The values are a moving average with a window of 400 iterations.

The origin of the oscillatory behavior is unclear, although the oscillation period is approximately equivalent to traversing the entire dataset.

With 10.9% of the dataset having the label 1 and the rest being labeled 0, a neural net that simply guesses a label of zero would achieve an accuracy of 89.1%, and that is what was seen when the deepnet was tested against the validation set, with a maximum accuracy of only 89.6% and a mean accuracy during the last 5 testing episodes of 89.1%.

5.3. Experiment 3: Hybrid Approach

The deepnet for the final experiment was trained for 97,000 iterations (*see* Figure 4), with primary reductions in training loss happening in the first third of the training process once again.

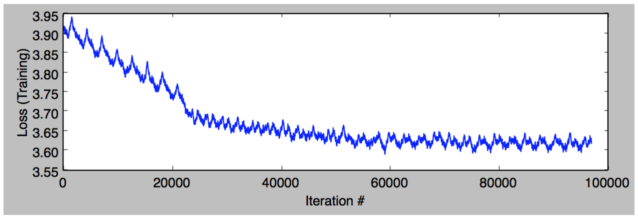

Figure 4: Training loss from experiment 3. The values are a moving average with a window of 600 iterations.

Oscillatory behavior was once again observed, albeit with less regularity. Caffe calculates Euclidean Loss according to:

$$\frac{1}{2n}\sum_{i=1}^{n}(x_i - \bar{x_i})^2$$

Given that there are 10 ‘labels’ which evenly partition the data into equal-sized groups, the expected Euclidean loss from guessing the mean ‘label’ is exactly 4.125. The deepnet did significantly outperform chance, with a mean validation-set accuracy of 4.01 and a minimum of 3.94.

6. Conclusions and Future Directions

We would be hard-pressed to claim that the deepnets were successful in our endeavor to predict the popularity of images. Here, we discuss several reasons for the failure of the deepnets roughly arranged in increasing order of the difficulty in overcoming them.

6.1. The deeepnets need more time or data

Training the deepenets longer would have necessarily improved the training loss, however, in all three experiments the validation accuracy was stable by the time training was halted, suggesting that the decrease in training loss was due to overfitting rather than true learning.

Of course, more data is always better, although with 400,000 images our dataset is roughly twice the size of the AVA dataset [12], which has been used successfully for research programs that are similar to this one (*see* Related Work), albeit with less abstract prediction goals. It is possible that our training data was simply too unconstrained, and that the task could be accomplished by focusing on a more specific domain of images (for instance, images of the beach or the ocean).

6.2. The deepnet's design was not appropriate

It remains possible that the deepnet employed (*see* Section 4.1) was not appropriate to this task, and a different architecture would have greatly improved the accuracy. This is of course always a possibility—one that any group working with deepnets can never be truly sure they have avoided. Nonetheless, we remain confident that the deepnet was appropriate.

As mentioned, the deepnet was adapted from a work which attempted, and succeeded, at classifying images based on their 'style' [8]. While classifying image style is different, perhaps even orthogonal, to predicting image popularity, the tasks are similar in that they rely on detecting and learning abstract or very high-level features of an image, or the confluence of low-level features on a global, image-wide scale.

For this reason we do not believe our choice of deepnet was inappropriate—and indeed we remain confident that it was the *most* apt given the time constraints required we use a pretrained model. In the future, when more time is available, we hope to architect a custom network and train it from scratch.

6.3. Images do not encode predictors of popularity

The most concerning possibility is that images do not encode the features that are predictive of their popularity—i.e., it is not possible, even in principle, to predict how popular and image is without the context of the image. This seems unlikely and is intuitively unsatisfying.

In the introduction, we made the argument that clearly at least *some* of the popularity of an image is encoded in the image itself, and offered as proof the fact that there are no popular images that are of white noise. While this is true, it is a distinct possibility that the intrinsic image features predict popularity up to a certain point. For instance, they might be able to inform one of an images *potential to become popular*, but whether or not they actually become popular is contingent on extrinsic features.

6.4. Future Directions

Despite the predictive failure of the deepnets in these experiments, much remains that can be done to potentially learn more. For instance, are other abstract features of an image predictive of popularity? I.e., general aesthetics, or emotional content, etc. Examining this covariance structure would allow us to focus on more tractable features of an image (i.e., aesthetics, which has been predicted with moderate success) as intermediaries, since they are known to predict popularity. Furthermore, this would permit use of existing datasets (like AVA).

As mentioned previously, designing custom deepnets (with increased depth to afford greater capacity for abstraction) could enable the detection of more subtle features relevant to popularity.

Finally, identifying more stable measures of popularity—perhaps on sites where content is generally shared among a known group (for instance, social contacts on a website like Instagram) would alleviate the difficulty presented by 'viral phenomena' (at least, partly).

References

[1] T. Aydin, A. Smolic, and M. Gross. Automated aesthetic analysis of photographic images. 2013.

[2] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah. Towards a comprehensive computational model foraesthetic assessment of videos. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 361–364, New York, NY, USA, 2013. ACM.

[3] R. Datta and J. Z. Wang. Acquine: Aesthetic quality inference engine - real-time automatic rating of photo aesthetics. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR '10, pages 421–424, New York, NY, USA, 2010. ACM.

[4] Z. Dong, X. Shen, H. Li, and X. Tian. Photo quality assessment with dcnn that understands image well. In X. He, S. Luo, D. Tao, C. Xu, J. Yang, and M. Hasan, editors, *MultiMedia Modeling*, volume 8936 of *Lecture Notes in Computer Science*, pages 524–535. Springer International Publishing, 2015.

[5] Z. Dong and X. Tian. Effective and efficient photo quality assessment. In *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, pages 2859–2864, Oct 2014.

[6] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1469–1482, 2014.

[7] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang. Understanding and predicting interestingness of videos. In *AAAI*, 2013.

[8] S. Karayev, A. Hertzmann, H. Winnemoeller, A. Agarwala, and T. Darrell. Recognizing image style. *CoRR*, abs/1311.3715, 2013.

[9] C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *Selected Topics in Signal Processing, IEEE Journal of*, 3(2):236–252, 2009.

[10] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*, MM '14, pages 457–466, New York, NY, USA, 2014. ACM.

[11] A. Mittal, R. Soundararajan, and A. Bovik. Making a 'completely blind' image quality analyzer. *Signal Processing Letters, IEEE*, 20(3):209–212, March 2013.

[12] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2408–2415, June 2012.

[13] Y. Wang, Q. Dai, R. Feng, and Y.-G. Jiang. Beauty is here: Evaluating aesthetics in videos using multimodal features and free training data. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 369–372. ACM, 2013.

7. Appendix - Code

Limited code was written for this project; Caffe was used, but run from the command line, and using a pretrained network. All scripting done was non-standalone scripts written for convenience purposes and hence are omitted here.