Latent Factors of Visual Popularity Prediction

Spencer Cappallo[†] Thomas Mensink[†] Cees G. M. Snoek[†‡]
[†]University of Amsterdam [‡]Qualcomm Research Netherlands
{cappallo, tmensink, cgmsnoek}@uva.nl

ABSTRACT

Predicting the popularity of an image on social networks based solely on its visual content is a difficult problem. One image may become widely distributed and repeatedly shared, while another similar image may be totally overlooked. We aim to gain insight into how visual content affects image popularity. We propose a latent ranking approach that takes into account not only the distinctive visual cues in popular images, but also those in unpopular images. This method is evaluated on two existing datasets collected from photo-sharing websites, as well as a new proposed dataset of images from the microblogging website Twitter. Our experiments investigate factors of the ranking model, the level of user engagement in scoring popularity, and whether the discovered senses are meaningful. The proposed approach yields state of the art results, and allows for insight into the semantics of image popularity on social networks.

Categories and Subject Descriptors

I.4.9 [**Image Processing and Computer Vision**]: Applications

Keywords

latent model; visual popularity; twitter

1. INTRODUCTION

Popularity is a difficult quantity to measure, predict, or even define. It is the result of many sundry factors and lies mercy to the whim of the zeitgeist. Still, we observe that humans exhibit some capability to predict what others will enjoy. For example, one expects that a photo of a cute kitten will yield a more positive response than a blurry photo of the ground. This fact suggests that there is an objective commonality to the appeal of certain images. The topic of this paper is the prediction of the popularity of images on social networks, such as photo-sharing websites and microblogging services, with the aim to gain insight into what qualities make a particular image popular.

Popularity prediction has been explored in the textual domain, *e.g.* [19, 18], but limited work has been done on the challenging problem of predicting popularity of images based on image content. Prior work has investigated the relative effectiveness of various features, both social and visual, for predicting the popularity of images on social networks [9, 3, 15]. These works demonstrate that visual features can be used to predict popularity of images. We build upon this foundation to explore the visual cues that determine popularity, identifying latently both the visual themes associated with popularity as well as those bound to unpopular images.

Previous work on image popularity has viewed the problem through the lenses of several different paradigms. Both [3] and [9] address popularity prediction as a regression problem, using support vector regression and random forests respectively. Meanwhile, McParlane et al.[15] cast the problem as one of binary classification, and utilize an SVM with a nonlinear kernel to make predictions. The method proposed in this paper instead views popularity prediction as

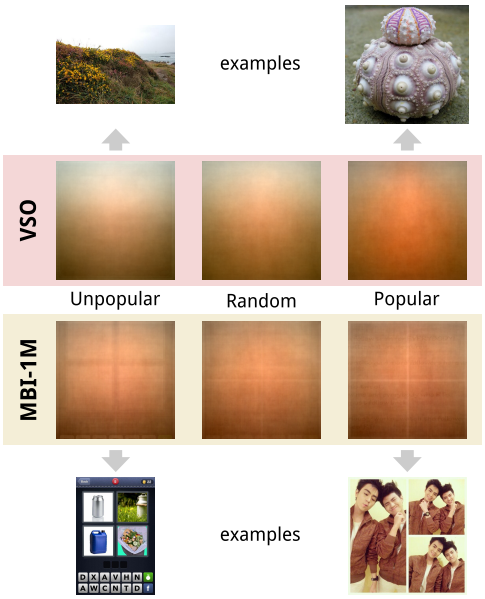

Figure 1: The average of 10,000 popular, unpopular, and random images for two datasets. Note the suggestion of a horizon line in the unpopular images of the Visual Sentiment Ontology dataset [2], while the popular images appear to have a more central composition. Meanwhile, our proposed dataset of 1M Micro-Blog Images from Twitter appear less organic, with highly defined structures indicative of many near-identical images. There is also horizontal striation suggestive of text. We report our visual popularity prediction experiments in the context of both photo sharing websites and microblogging.

ICMR '15, June 23 - 26, 2015, Shanghai, China

http://dx.doi.org/10.1145/2671188.2749405.

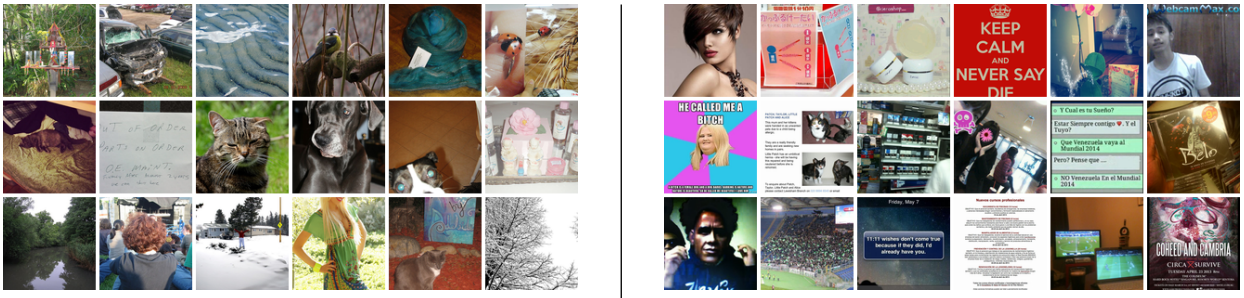

Photo-sharing **Micro-blogging**

Figure 2: Thumbnails of a random selection of photos from two datasets. Note the larger emphasis on natural images in the Flickr-based dataset on the left, versus the large number of graphics in the proposed Micro-Blog Images dataset on the right.

essentially a retrieval problem, and introduces a novel application of latent weights to improve performance and gain insight on visual popularity. Furthermore, previous papers have focused on a comparison and combination of visual and other features, while we seek to understand more about the nature of visual popularity itself.

There is a fundamental question as to whether popularity can be measured objectively. Following previously published work [18, 3, 9] we assume that recorded metrics like view count, comment count, and "likes" are indicative of overall popularity. However, there is no unified evaluation metric in the literature, which is in part due to the difference between viewing the problem of popularity prediction as either binary classification or regression. McParlane et al. [15] report the mean class accuracy, choosing a binarization of the annotations where the 20% most popular images belong to the positive class, and the remaining 80% belong to the negative class. Alternatively, Can et al.[3] report the root-mean-square error in the log domain, as they have approached the popularity prediction as a regression problem. Finally, Khosla et al.[9] address the problem as a ranking problem and report results in Spearman's correlation coefficient, a measure of the similarity between two rankings. Since our methods also consider popularity prediction as a ranking problem, we follow [9] and use Spearman's correlation coefficient for evaluation.

Both [9, 15] use large-scale datasets from the photo-sharing website Flickr. Khosla et al.[9] use the Visual Sentiment Ontology dataset [2], while McParlane et al.[15] use the MIR-1M dataset [13]. Despite using a large dataset, [15] experiments only on a small subset of the MIR-1M dataset of 1000 images. Flickr is a photo-sharing website, and thus the nature of its images may differ greatly from the nature of visual popularity on a microblogging website, like Twitter or Weibo. While Can et al.[3] investigate popularity prediction on the microblogging service Twitter, and present results on the improvement of adding visual features to social features, none of these previous works investigate how visual popularity changes between different types of social media-sharing networks.

In this paper we investigate what factors contribute to visual popularity prediction. We formalize the prediction as a ranking problem with a latent-SVM objective for both positive and negative senses of popularity. In an effort to generalize our conclusions beyond popularity on photo sharing websites, we propose a new image dataset which we harvest from the microblogging website Twitter (Figures 1 and 2). Our experiments on three datasets investigate factors of the ranking model, the level of user engagement in scoring popularity, and whether the discovered senses are meaningful. Before detailing our experiments, we first introduce our ranking model.

2. RANKING (UN)POPULAR IMAGES

2.1 Popular latent senses

We view image popularity prediction as a retrieval problem: Given a large set of images, we wish to rank the images in order of likely popularity. For this reason, we follow a ranking SVM [8] approach. We assume a set of n training examples $\{x_i, y_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^{1\times d}$ denotes the d-dimensional visual features of image i, which is based on deep convolutional networks [10] in our case, and where y_i denotes the popularity of an image, which is defined as $y_i = log(c_i + 1)$ where c_i is a measure of popularity, such as view count or number of comments.[1] The ranking SVM framework then minimizes the following objective function:

$$L_{\text{rnk}} = \sum_i \sum_j |1 - f(x_i) + f(x_j)|_+ \qquad (1)$$
$$\forall i, j \text{ s.t. } y_i > y_j$$

where $f(x) = w^\top x$ is the prediction function, and $w \in \mathbb{R}^{1\times d}$ is the weight vector to learn. Due to the size of our training sets, we train all our models using stochastic gradient descent [1] with mini-batches of around 100 images where all viable pairs are incorporated into the computed gradient.

Visual popularity is a challenging problem, and the visual clues which determine popularity may cover a wide range of image types. As an example, consider photos of celebrities and photos of beautiful landscapes. Both types of photo are more likely to be popular on social media than photos of spatulas, but they have two distinct sets of visual cues associated with them. The problem becomes needlessly difficult if approached with a model that is not sufficiently complex to exploit the very different properties of landscapes and celebrity portraits.

Therefore we turn to the idea of learning latent popularity senses, where each image is ranked according to the highest scoring latent sense, leading to a latent-SVM objective.

[1]Note that the log transform does not influence the standard ranking SVM objective, but later on we redefine the margins based on the values of y_i.

The latent-SVM objective is probably best known for object detection, where it is used in the seminal Deformable Parts Model [6] to find the location of the part filters. In our paper, we do not aim to find the best location of a part, but instead to find a visual popularity sense to which the image belongs. This is conceptually similar to latently learning multiple models for image classification [16] and for query-based image retrieval [12]. For the latter, a latent ranking algorithm is introduced for binary classification which uses latent weights to capture the distinct visual senses of a label. For example, 'jaguar' may correctly refer to either the animal or the automobile, which have very different visual appearances which would be challenging to fully capture in a single representation without latent senses. We draw inspiration from these approaches and adapt them to the problem of popularity prediction, with the intent to both improve prediction performance and gain additional insight into the visual themes of popular images.

Following this rationale, we alter the ranking SVM objective in the following manner:

$$L_{\text{lat}} = \sum_i \sum_j |1 - f_s(x_i) + f_s(x_j)|_+ \quad (2)$$
$$\forall i, j \text{ s.t. } y_i > y_j$$

where $f_s(x) = \max_{s \in S} w_s^\top x$, corresponding to selecting the score of latent sense with the maximum response to x, and S is the set of all latent weights.

The method selects pairs of images with different popularity annotations and identifies the latent weights which respond most strongly to each image. It then updates the weights accordingly, by punishing the latent sense that responds strongly to the less popular image, while encouraging the latent sense that responds most strongly to the more popular image.

Due to the fickle nature of popularity, it is unclear whether it is wise to learn as large a separation between two images with similar popularity value as between two images with extremely different popularity scores. For this reason, we replace the fixed margin with a soft-margin that is dependent on the difference in popularity:

$$L_{\text{sft}} = \sum_i \sum_j |\Delta(y_i, y_j) - f_s(x_i) + f_s(x_j)|_+ \quad (3)$$
$$\forall i, j \text{ s.t. } y_i > y_j$$

where $\Delta(a, b)$ defines the margin between between a and b, which is similar to the margin rescaling formulation used in structural SVMs [20], and $f_s(x) = \max_{s \in S} w_s^\top x$, as before.

Several approaches for selecting Δ are possible, including:

$$\Delta(i, j) = y_i - y_j$$

where in our case y corresponds to the log of the popularity annotation. However, especially in the log domain, the separation between y_i and y_j can be very small. To encourage greater separation between examples, we also propose the following alternative:

$$\Delta(i, j) = \max(\alpha, y_i - y_j)$$

where α is a predefined constant minimum margin between examples. In preliminary experiments, we found that maintaining a constant minimum margin yielded the best results, and we use $\alpha = 1.0$ in all our experiments.

	MBI-1M
Images	1,007,197
Average Favorites	0.866
Average Re-tweets	1.133
Average Re-tweets + Favorites	1.998

Table 1: Statistics of our proposed Micro-Blogging Images dataset

	VSO	**MIR-1M**
Images	891,297	865,833
Average Views	422.2	904.2
Average Comments	4.6	12.5
Average Views + Comments	426.8	912.5

Table 2: Statistics of the photo-sharing datasets

2.2 Unpopular latent senses

The proposed latent model discovers categories of images which are informative for ranking visual popularity. It focuses on amplifying those visual cues which correspond positively to popularity. However, it is reasonable to assume that there are also visual cues which suggest an image is unpopular. To draw from the earlier example, photographs of spatulas might be plentiful, yet consistently unpopular on social media. If the model focuses solely on optimizing for popular cues, it may underutilize this informative data. Furthermore, on a qualitative side, identifying common visual aspects to what makes an image *unpopular* is equally as interesting as identifying what makes an image *popular*.

For these reasons, we introduce latent senses for identifying unpopular images which are learned in parallel to those optimizing for popularity. To do this, the latent senses learned by the model are split into two halves. The weights of the first half are learned in the manner described in the previous section, while the weights of the second half are learned in a contrary manner to encourage the discovery of visual cues that correspond inversely to popularity. These two sets of weights are modified accordingly:

$$L_{\text{p\&n}} = \sum_i \sum_j \Big[|\Delta(y_i, y_j) - f_{s_+}(x_i) + f_{s_+}(x_j)|_+ + |\Delta(y_i, y_j) - f_{s_-}(x_j) + f_{s_-}(x_i)|_+ \Big] \quad (4)$$
$$\forall i, j \text{ s.t. } y_i > y_j$$

At test time, the popular and unpopular senses are combined to form a single score:

$$f_{\text{p\&n}}(x) = \max_{p \in S_+} w_p^\top x - \max_{n \in S_-} w_n^\top x \quad (5)$$

where S_+ and S_- are the sets of senses focusing on popular and unpopular images, respectively.

3. EXPERIMENTAL SETUP

3.1 Datasets

We explore two domains for popularity prediction in our experiments, micro-blogging and photo-sharing, for which we make use of three large scale image datasets. The datasets

are described below. See Tables 1 and 2 for an overview of basic statistics.

MBI-1M We introduce a new and challenging dataset for the task of popularity prediction, the 1M Micro-Blog Images (MBI-1M). This new dataset consists of over 1M images taken from Twitter. The images selected are taken from a subset of the 240 million tweets collected for the TREC 2013 microblog track [11] and these datasets can thus be used in conjunction. Users on Twitter are able to share images through Twitter's image hosting service, and 1 million tweets from February and March 2013 which contained such images were selected. The images were collected through the Twitter API, and up-to-date retweet and favorite counts (our measures of popularity on Twitter) were collected for the 1 million selected Tweets. MBI-1M has been released.[2]

VSO The Visual Sentiment Ontology (VSO) dataset [2] consists of 930k images from Flickr, which were collected based on a search through the Flickr API for 3,244 adjective-noun pairs. Two examples of these adjective noun pairs are "beautiful girl" and "little house". This dataset was used in [9], and thus serves a fair comparison point to establish state of the art. It further serves as a testbed for making comparisons regarding the nature of visual popularity on different social networks. We choose the total number of views and total number of comments as our measures of popularity on Flickr. However, since these have not been made available for this dataset, we collected our own data through the Flickr API, which we have released.[2] Due to deleted photos and other changes, the meta data was available for 891,297 of the images. This subset is used in our experiments.

MIR-1M We also provide some results on the MIR-1M [13] dataset, a dataset of 1M images from Flickr with a creative commons license. The images are selected based on their Flickr interestingness scores. The MIR-1M dataset was introduced for benchmarking image retrieval, however McParlane et al. [15] has made available popularity data for 865,833 images. For this dataset we use the total number of views as our measure of popularity.

Visual features For all experiments, we use features from a ConvNet structured after AlexNet [10] and trained to identify 15k ImageNet classes [4]. The features used are 4096-dimensional features from the last fully connected layer of the network, and have proven to work well as general and state-of-the-art features for many computer vision tasks [7, 14], including view count prediction [9].

Data Splits For all three datasets, we randomly allocate the data into training, validation, and test sets, consisting of 70%, 10%, and 20% of the images respectively.

The nature of the images on Flickr and Twitter are very different. As can be seen in Figure 2, Flickr has a greater focus on photography, while Twitter contains a large number of graphics and images of text. This is further illustrated in Figure 1, where the mean of popular, random, and unpopular images is displayed. For the Flickr-based VSO dataset, gradients suggestive of landscape photos appear unpopular, while the most popular images appear to have some central subject like a face. Meanwhile, a lot of structure is evident in both the most and least popular Twitter images, suggesting a repetition of certain types of graphics.

[2] Released datasets can be found at:
`http://staff.fnwi.uva.nl/s.h.cappallo/data.html`

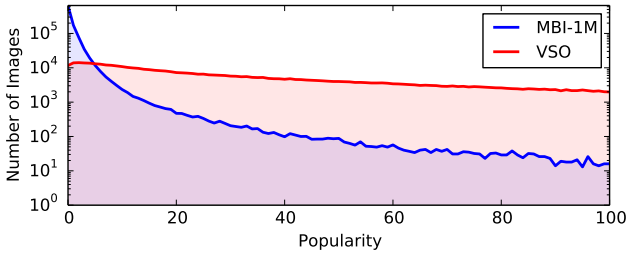

Figure 3: Distribution of view count for VSO dataset and the sum of retweets and favourites for the proposed MBI-1M dataset, plotted with a logarithmic y-axis. The Twitter-based MBI-1M dataset is dominated by images with low or zero popularity counts, while Flickr-based VSO exhibits a broader distribution.

3.2 Evaluation criteria

As stated previously, we see popularity prediction primarily as a ranking problem. Therefore the results are evaluated with Spearman's rank correlation coefficient, ρ. Spearman's correlation coefficient is a measure of the dependence between two rankings [17]. We use it to correlate the predicted ranking $\hat{R}$ with the ground-truth ranking R, as follows:

$$\rho = 1 - \frac{6 \sum_i^n (\hat{r}_i - r_i)^2}{n(n^2 - 1)} \tag{6}$$

where n is the number of datapoints, $\hat{r} \in \hat{R}$, $r \in R$, and the sum term measures the square difference between the predicted rank and ground-truth rank for all examples. The value ρ therefore measures how closely the predicted rank aligns with the ground truth ranking. In the case of a perfect, monotonic relationship, $\rho = 1.0$, and in the case of a random relationship it receives a value of $\rho = 0$.

3.3 Experiments

We evaluate the effect of incorporating latent senses with and without a fixed margin term, including latent senses focusing on unpopular images. We investigate the difference in predictive ability of Flickr views and comments, as well as Twitter retweets and favorites. Finally, we also explore the semantics of the latently learned senses. We structure our experiments by four questions.

Do latent senses improve popularity prediction? This experiment explores whether adding latent senses to the proposed model improves performance. For these experiments, the model attempts to enforce a fixed margin of 1.0 between example pairs during training. We test this approach with no latent senses, and with a varied number of latent senses.

The nature of popularity is such that, as seen in Figure 3, there is a very long tail. The difference in popularity between the most popular and least popular images can be very great, while the difference in popularity among the bulk of the images is small. As we seek to identify and understand which images become very popular, it is perhaps desirable to exploit this variability. For this reason, we also investigate the effect of enforcing a larger separation between very popular and unpopular images, while reducing the penalty for examples with very similar popularity scores. We do this by allowing for a dynamic margin based on the delta in

	MBI-1M		VSO		MIR-1M
Latent Senses	**Fixed Margin**	**Soft Margin**	**Fixed Margin**	**Soft Margin**	**Soft Margin**
None	0.161	0.160	0.330	0.331	0.330
5	0.166	**0.165**	0.333	0.334	0.346
10	0.165	0.165	0.335	**0.337**	0.346
20	**0.167**	0.165	**0.336**	0.335	**0.346**

Table 3: The effect of adding latent senses to the ranking SVM with both a fixed margin of 1.0 and a soft margin. All values reported as Spearman's ρ. Note a unilateral improvement with adding latent senses, along with diminishing returns with increased number of senses. Due to similarity between VSO and MIR-1M datasets, only soft margin experiments are reported for the latter.

popularity values.

Does engagement level affect predictability? Popularity is a multi-faceted and highly abstract concept; it is impossible to quantify directly. Instead, highly correlated values like view count or re-tweets are employed as proxy measures. No single metric among these has an absolute correspondence to popularity.

The common thread between these disparate popularity measurements is that they all represent some countable interaction between a broadcaster and recipients. We put forth the notion that all these measures lay along a spectrum of engagement, defined by the level of interaction required by the recipients. The popularity measures discussed can be lumped into two broad categories along this spectrum: "High Engagement" measures, where recipients have a high degree of interaction with or investment in the image that is broadcast, and "Low Engagement" measures, where recipients have a small or passive level of interaction or investment. For the purposes of the datasets and popularity measurements discussed within this paper, the act of commenting on or re-broadcasting an image is seen as high engagement, while merely viewing or "liking" an image are labelled low engagement.

To explore the effect of these alternate measures on the predictability of visual popularity, we test our model on low and high engagement metrics, as well as the sum of both types of metric. Results are presented for both the Flickr-based VSO dataset and the proposed MBI-1M dataset.

How do unpopular latent senses alter prediction? Beyond the existence of popular imagery, it is likely that there also exists consistently unpopular imagery. For this reason, a modification to our approach is tested wherein the existing latent senses are split into two distinct sets: senses attempting to maximize visual popularity, and senses attempting to maximize consistently unpopular imagery.

Can semantics be extracted from latent senses? In the process of training, the proposed model attempts to learn latent visual senses which are informative for predicting popularity. Such senses can be viewed as categories describing the common visual themes of popularity. The question arises, then, as to whether anything meaningful can be said about or extracted from the photos contained within these latent categories.

The visual cohesiveness of such latent categories is difficult to evaluate empirically. To that end, we explore whether any semantics can be attached to these categories. The VSO dataset contains adjective-noun pair ground truth annotation for its images. We evaluate whether there are any dominant adjectives or nouns within a visual category, which would be suggestive of cohesive visual semantics.

For MBI-1M, there is no ground truth annotation to evaluate latent category semantics, but each image has accompanying text from the original tweet. To test whether we can pull any descriptive semantics from this text, we perform *tf-idf* on the nouns of the top most highly ranked tweets per category. The combined text from these tweets in each latent category is treated as a document for the purposes of *tf-idf*, and the words with the highest *tf-idf* values per sense are viewed as the most descriptive.

Finally, the proposed method is also compared against previously published results.

4. RESULTS

4.1 Latent senses

As seen in Table 3, the inclusion of latent senses within the ranking SVM model improves performance, but there are limited or diminishing returns after only a few senses have been added. It is important to note that this is merely a reflection of the overall ranking performance of the model, and does not necessarily say anything about whether additional senses may improve the semantic interpretability of popularity. Whether these latent senses have any semantic cohesiveness will be investigated in a later section.

Also visible in Table 3, introducing a more dynamic margin has a small effect on the performance of the model. For the MBI-1M, we see broadly similar, though slightly smaller ρ values. Only 35% of the tweets in the dataset have received any re-tweets or favourites whatsoever, and of those, over half have received only one re-tweet or favourite. This means that only 18% of all examples have a popularity value other than 0 or 1. It is posited, therefore, that the benefits of a dynamic margin are largely irrelevant when the variability of popularity values is so limited. The results on the VSO dataset lend weight to this argument, as a small improvement over the fixed margin is observed. The VSO dataset has a distribution with a much wider spread across popularity scores, and therefore the incorporation of this information within the model is expected to be more useful than in the case of MBI-1M.

During learning, a gradient is only calculated for those latent senses which exhibit the maximum response to the example images. This raises a concern that a random initialization of the weights could result in overly dominant senses being subject to the majority of the learning. Latent senses with less fortuitous initializations might then languish, being rarely or never selected during training. To help ensure that all latent senses have a chance to learn, we use the cen-

Figure 4: An example of three latent senses learned on the VSO dataset, showing the test images each sense predicts as most popular. Note that, while the model learns latent visual categories, there is no guarantee of semantic similarity. In particular, the top row exhibits a visual cohesion but a semantic diversity.

troids of a k-means clustering on the top 1000 most popular images to initialize our latent weights. We observe an improvement from $\rho = 0.162$ to $\rho = 0.165$ on MBI-1M when latent weights are initialized with k-means rather than being randomly initialized.

Adding latent senses improves popularity prediction. However, there is little difference in the overall predictive capability as the total number of latent senses moves beyond only a handful. The use of a soft margin has a only small impact on popularity prediction, and whether that impact is positive or not appears to depend on the properties of the distribution of the annotations.

4.2 Engagement level

Interestingly, as seen in Table 4, low engagement popularity metrics are more predictable across both datasets than high engagement popularity. Especially marked is the large difference between view counts and comment counts for predictability on the VSO dataset. It appears that the factors which determine total comment count are further removed from the visual features of the image than those of view count. This is perhaps indicative of the influence of an image's larger context on a user's decision to comment on a photo. For example, an embarrassing photo of a political leader may engender lengthy discussion in its comments, but be visually indistinct from any number of other portraits. In contrast, a photo of a beautiful sunset may be more universally appealing in content, but lacks the grander context that would drive users to comment.

It also seems likely that a similar effect may occur in the microblog popularity data. For example, though our model can identify the general visual properties of images of text, it does not incorporate the content of that text, which may play a large role in whether or not a user will re-broadcast a particular image. Meanwhile, a photograph of a broadcaster's family member may warrant a *favorite* from a recipient, but is unlikely to be re-broadcasted by that recipient unless it also holds special relevance for them. Notably, the combination of high and low engagement measures outperforms either alone on MBI-1M, which suggests that *retweets* and *favorites* may be more closely linked than *views* and *comments*.

The level of engagement of a popularity metric has an effect on its popularity. Low engagement measures are easier to predict than high engagement ones, possibly due to the greater context of an image playing a larger role in recipient participation for high engagement metrics.

Engagement	MBI-1M (ρ)	VSO (ρ)
Low	0.151	**0.337**
High	0.130	0.110
Combined	**0.165**	0.328

Table 4: The relative predictive capability of high and low engagement measures of popularity. Low engagement measures outperform high engagement measures on both datasets, perhaps indicative of the wider context of an image playing a larger role in high engagement measures. There is a larger discrepancy for Flickr-based dataset VSO, compared with Twitter-based MIB-1M, indicative of the differences between social networks and their popularity measures.

Top words: pancakes, smoothie, pancake, bread, pork

Figure 5: Example of a latently learned negative sense, along with most prevalent words in accompanying tweets.

4.3 Unpopular latent senses

As shown in Table 5, there is an improvement when including latent senses identifying unpopular imagery over focusing solely on popular imagery. When learning only popular latent senses, the focus is on those visual cues which signify a popular image. Visual cues which are uniformly indicative of unpopular images are learned only to the ex-

Latent Senses	MBI-1M ρ	VSO ρ	MIR-1M ρ
Popular	0.165	0.337	0.346
Popular and Unpopular	**0.172**	**0.345**	**0.365**

Table 5: The effect on popularity prediction of adding latent senses focusing on unpopular images. Values reported here for models with 10 latent senses, either entirely popular or split into 5 popular and 5 unpopular. Learning to identify unpopular images in parallel to popular ones yields improvement across all three datasets.

tent that no popular senses should respond strongly to them; once the responses for those images are pushed down sufficiently far, learning halts. This has perhaps little effect on the top ranked results, but has an effect on the overall ranking. Overlooking these visual cues that suggest an image is unpopular means that the tail end of the ranking is not optimized.

Incorporating latent senses to detect the visual cues of unpopular imagery is important for the overall performance of a ranking approach. The learned senses are informative for identifying those images that are categorically unpopular. Furthermore, as seen in Figure 5, these unpopular senses can be interesting in their own right, telling us something about the types of images which are commonly posted and routinely ignored.

4.4 Semantics from latent senses?

In Figure 6a, the top predicted test images for two latent senses are displayed alongside the adjectives and nouns that most frequently appear within the top 500 predicted images of the senses. Note that while both latent categories appear to have cohesive visual styles, the most common words are only descriptive for one of them. This suggests that the adjectives and nouns within the VSO annotations happen to capture the visual semantics of one of the categories, but not the other.

On the other hand, in Figure 6b are the top ranked test images of latently learned categories from MBI-1M alongside the most common nouns from the accompanying tweets of the top 10,000 predicted test images. As the words are not restricted to the small set of VSO annotations, we observe that the top words for a category of text and graphics is representative of its content.

The model proposed within this paper seeks to group visually similar images together which aid in predicting popularity. For this reason, it differs from a straightforward clustering approach which seeks merely to group visually similar images. This means that the latent categories are likely to have broader content variety than you might see with clustering, especially with a low number of latent senses. In the earlier section, it was shown that increasing the total number of latent senses learned did not have a significant impact on overall performance. However, it is important to note that such a change does have an impact on the visual diversity of each latently learned category.

The categories which are latently learned appear to have largely cohesive visual semantics. Through utilizing accompanying text, we are able to automatically extract some descriptive text for these semantics. The ability to assign

Nouns: dog, animals, cat, face, eyes, puppy, animal
Adjectives: wild, happy, funny, little, cute, great, dirty

Nouns: girls, smile, eyes, face, lady, hair, dress
Adjectives: young, sexy, little, beautiful, hot, happy

(a) *VSO*

Top words: dress, shoes, fashion, wear, bacon, corn

Top words: followers, conversation, text, quote, stupid

(b) *MBI-1M*

Figure 6: Top-ranked images for two latent senses of the VSO dataset (a) and two from the MBI-1M dataset (b), along with the most common words from the ground truth annotation and text of the tweets, respectively. Note that the VSO ground truth does not have sufficient semantic overlap with the bottom category, but performs well for the top category. In contrast, because the text of tweets is less semantically restricted, the top words of both categories for MBI-1M relate to their visual themes.

Method	VSO (ρ)
Khosla et al. [9]	0.315
This paper	**0.345**

Table 6: Comparison of our proposed method to the regression model of Khosla et al.[9] using our features.

words to these latent categories allows insight into the visual themes that determine whether an image is popular or unpopular.

4.5 Comparison to others

First we compare our proposed latent ranking method to the approach of Khosla et al. [9] in Table 6 on the VSO dataset. Following [9], we use Decaf image features [5] with a linear support-vector regression model to obtain a ρ of 0.28. Our deep-learning ConvNet features, trained on 15K instead of 1K classes, using the same linear support-vector regression model obtains a larger ρ of 0.315. However, using our latent ranking method we obtain a ρ of 0.345, an improvement over the current state-of-the-art, which is not explainable merely through the use of improved features, but by using a more complex popularity prediction model.

We also compare the proposed latent ranking against the results published in McParlane et al.[15] on the MIR-1M dataset. In an attempt to make our comparison as fair as possible, we evaluate our trained ranking models on a binarized testset. McParlane et al. binarized their data in such a way that the top 20% most popular images were assigned to the positive, "popular" class, and the remainder were assigned to the "unpopular" class. They report 53% accuracy on a test set of 1,000 images comprised of 50 % popular and 50% unpopular images, and 59% accuracy using a combination of visual, social and textual features. To approximate their testing environment, we calculate accuracy on a subset of our test images, which is comprised of the 20% most popular images, and an equal number of random images from the remaining 80%, resulting in over 60K test images. To binarize the prediction of our ranking model, we use $l_i = [\![\hat{y}_i > \text{thr}]\!]$, where $\hat{y}_i$ is the predicted popularity, we find the optimal threshold value based on the performance on the training data. Our model performs with 64% accuracy, which compares favourably to the published results of McParlane et al.

5. CONCLUSIONS

We have explored various factors of image popularity on social media: popular and unpopular visual senses, low and high engagement popularity measurements, and different sources of social media. Our proposed ranking model outperforms the current state-of-the-art in predicting popularity, yet the community has a long way to go before image popularity prediction is fully understood. By making available the MBI-1M dataset and our full evaluation protocol, we hope to engage other researchers into the challenging quest of image popularity prediction as well.

Acknowledgments This research is supported by the STW STORY project and the Dutch national program COMMIT.

6. REFERENCES

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Good practice in large-scale learning for image classification. *IEEE Trans. PAMI*, 36(3):507–520, 2014.

[2] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *MM*, 2013.

[3] E. F. Can, H. Oktay, and R. Manmatha. Predicting retweet count using visual cues. In *CIKM*, 2013.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. Technical report, arXiv preprint 1310.1531, 2013.

[6] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. PAMI*, 32(9):1627–1645, 2010.

[7] D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors. *Proceedings of the European Conference on Computer Vision*. Springer, 2014.

[8] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.

[9] A. Khosla, A. Das Sarma, and R. Hamid. What makes an image popular? In *WWW*, 2014.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[11] J. Lin and M. Efron. Overview of the trec2013 microblog track. In *TREC*, 2013.

[12] A. Lucchi and J. Weston. Joint image and word sense discrimination for image retrieval. In *ECCV*, 2012.

[13] B. T. Mark J. Huiskes and M. S. Lew. New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In *ICMR*, 2010.

[14] A. Martinez, R. Basri, R. Vidal, and C. Fermuller, editors. *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 2014.

[15] P. J. McParlane, Y. Moshfeghi, and J. M. Jose. Nobody comes here anymore, it's too crowded; predicting image popularity on flickr. In *ICMR*, 2014.

[16] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Trans. PAMI*, 2013.

[17] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[18] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Comm. ACM*, 53(8):80–88, 2010.

[19] M. Tsagkias, W. Weerkamp, and M. De Rijke. Predicting the volume of comments on online news stories. In *CIKM*, 2009.

[20] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.