Exploiting Visual-based Intent Classification for Diverse Social Image Retrieval

Bo Wang[1], Martha Larson[1,2]
[1]Delft University of Technology, Netherlands
[2]Radboud University, Netherlands
b.wang-6@student.tudelft.nl,m.a.larson@tudelft.nl

ABSTRACT

In the 2017 MediaEval Retrieving Diverse Social Images task, we (TUD-MMC team) propose a novel method, namely an *intent-based* approach, for social image search result diversification. The underlying assumption is that the visual appearance of social images is impacted by the underlying photographic act, i.e., why the images were taken. Better understanding the rationale behind the photographic act could potentially benefit social image search result diversification. To investigate this idea, we employ a manual content analysis approach to create a taxonomy of intent classes. Our experiments show that a CNN-based neural network classifier is able to capture the visual difference between the classes in the intent taxonomy. We cluster images of the Flickr baseline based on predicted intent class and generate a re-ranked list by alternating images from different clusters. Our results reveal that, compared to conventional diversification strategies, intent-based search result diversification is able to bring a considerable improvement in terms of cluster recall with several extra benefits.

1 INTRODUCTION

The recent advances in deep learning, especially convolutional neural networks, have been successfully applied in various computer vision and multimedia tasks such as object recognition and scene labeling [4]. However, recognition of the literally depicted content of multimedia documents (i.e., what is visible in the image) has absorbed most of the research attention. In contrast, less research has focused on social, affective and subjective properties of data, for example, why the image was taken.

In this paper, we focus on user intent, i.e., the goals that users are pursuing when they take photos. We assume that intent has visual reflexes that can be captured by automatic visual classifiers. Intent classes can be further applied to search result diversification. The goals of the photographer provide a simple, easily understandable explanation for the differences observed between photos [7]. However, given the lack of intent taxonomies (definitions of intent classes) and data sets annotated with intent labels, we will start with creating a taxonomy of intent classes, which we turn to next.

2 INTENT DISCOVERY

2.1 Data Set Generation

The intent taxonomy was created using a manual content analysis [5] approach on the basis of YFCC100M [10], the largest social image collection that has ever been released. Since we are interested in building a taxonomy of intent classes with higher abstraction level that goes beyond concept detection, we choose to use NUS-WIDE concepts (81 concepts) [2]. We use these concepts as queries to retrieve images from the YFCC100M data set (using a tag-based retrieval system). For each query, we collect the top-200 relevant images. We use the entire results list if less than 200 images are found. After querying for all NUS-WIDE concepts, we arrive at a data set containing 15618 images.

MediaEval'17, 13-15 September 2017, Dublin, Ireland

2.2 Intent Labeling

The intent taxonomy and labeled data set were produced by an expert annotator, who examined each image in turn. The manual content analysis approach used by the annotator consists of several steps. For each image, the annotator first assigns a preliminary intent label. Each new image is then judged as either belonging to an existing intent class, or requiring the creation of a new intent class. Before introducing a new class, the annotator returns to the previous annotated images to ensure that it is not possible to accommodate the new image by updating the description of an existing class. If no existing class can be extended to incorporate the new image, a new intent class is introduced. The final 14 classes intent taxonomy are described in [11].

3 INTENT CLASSIFICATION

We adopt a conventional transfer learning scheme to predict the intent class of an image. Transfer learning trains models on one task, and leverages them for a different, but related task [6]. In our case, we used VGGNet [9] to extract visual content features from our images (originally trained on ImageNet [3]). The last fully connected layer (between 2048 neurons and 1000 class scores) was removed and the rest of the network serves as a feature extractor. We retrained a Softmax classifier using a cross-entropy Softmax loss on our image data set annotated with 14 intent classes. We used 70% of the data for training and held 25% of the data out for validation purposes. (The remaining 5% is not used here.) Before we trained, we re-sized all images to 224x224 pixels, and applied data augmentation (random horizontal flipping, chopping and rescaling). Our model achieved 71% accuracy on the validation set, suggesting that intent classes are visually stable enough to allow a classifier to generalize over them.

4 DIVERSIFICATION

The intent-based search result diversification works as follows: The first step is to create a refined initial ranked list by re-ranking the Flickr baseline using textual features (vector space model with tf-idf weights) with the aim of increasing precision. After that,

Table 1: Results in terms of Precision, Cluster Recall and F1 score with respect to 4 different runs on Dev and Test set.

Data Set	Evaluation	**visual (run1)**	**text-rerank+text (run2)**	**text-rerank+visual (run3)**	**text-rerank+intent (run4)**
Dev Set	P@20	61.52%	67.72%	67.72%	67.69%
	CR@20	49.29%	52.36%	53.61%	55.61%
	F1@20	54.73%	59.05%	59.83%	61.07%
Test Set	P@20	66.01%	70.36%	70.71%	72.62%
	CR@20	56.98%	61.42%	58.09%	61.25%
	F1@20	58.30%	63.43%	61.21%	64.62%

the top N images in the re-ranked list are classified by our intent classifier. In our case, N is 50. To generate the final results list, we apply a round-robin approach. We consider each intent class to be a cluster of images, and pick the top-ranked photo from each intent cluster (without replacement) in turn. This approach applies the assumption that new clusters reflect diversity as captured by photographer's intent.

In addition to the intent-based approach, we also submitted three runs: *visual* (run1), *text-rerank+text* (run2) and *text-rerank+visual* (run3) for search result diversification. The intent based approach is designated *text-rerank+intent* (run4).

For *visual* (run 1), we directly apply k-means clustering to the CNN-based descriptors provided by the task organizers [12]. We employed a heuristic approach to initialize k. Specifically, we treat k as a variable and initialize $k \in (1, n]$ and apply k-means clustering for n times. For each k, we evaluate clustering performance with silhouette analysis [8] and select the best k with respect to the achieved silhouette score.

Our *text-rerank+visual* (run3) adopts the same general strategy as the visual-based approach. The difference is that instead of directly apply k-means clustering, we first re-rank the Flickr baseline with tf-idf weights and then cluster.

For our *text-rerank+text*(run2) approach, again, we first re-rank the Flickr baseline with tf-idf weights. Since in this case, we are not allowed to use visual descriptors, the most critical issue is to learn a good representation for each "short document" consisting of title, description and tags. To achieve this, we adopted the idea of *weighted word embedding aggregation* proposed by Cedric et al. [1]. More concretely, for each term associated with an image, we use its 50-dimensional word embedding vector. (Word embedding vectors were supplied by the organizers.) Each image is thus represented as a set of vectors. For an image with m terms, we have set of m 50-dimensional vectors. To model an image, we take the coordinate-wise maximum and minimum of the set of m vectors. We concatenate the two resulting vectors (min and max) to arrive at a 100-dimensional vector, which is our final text-based image representation. For each query, we have a set of 300 image vectors, to which we apply k-means clustering with silhouette analysis.

5 RESULTS AND ANALYSIS

Table 1 reports the results in terms of the official MediaEval 2017 evaluation metrics P@20, CR@20 and F1@20. In general, higher precision is usually associated with relatively higher cluster recall and F1 scores because non-relevant images have no associated diversity cluster label. This phenomenon can be clearly observed comparing *visual* and *text-rerank+visual*. What is surprising is that the text-based image representation achieves a better clustering result on the test set compared with the visual CNN representation. The text-based approach *text-rerank+text* and our intent-based strategy *text-rerank+intent* perform comparably on the test set. The intent-based approach appears to give a boost to relevance as measured by P@20 and F@20.

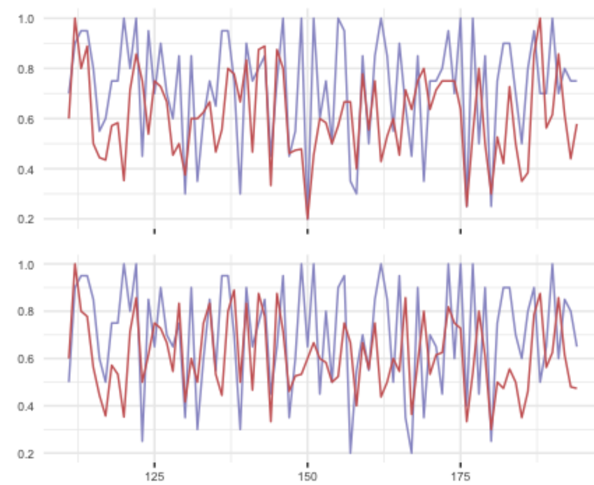

Figure 1: Comparison between *text-rerank+intent* (run4) (above) and *text-rerank+text run* (run2) (below) over all query id (x-axis), purple: P@20, red: CR@20.

Figure 1 shows that both metrics fluctuate widely with respect to different queries. We measured the Pearson coefficient between P@20 and CR@20 for *text-rerank+intent* (run4) (0.41) and *text-rerank+text* (run2) (0.35), which reveals that the intent-based approach is more sensitive to initial ranking precision. The standard deviations are comparable: $\sigma = 0.17$ for *text-rerank+text* and $\sigma = 0.18$ for *text-rerank+intent*.

We point out three other aspects of the intent-based diversification approach that make it practically useful. First, intent-based diversification has the advantage of better understandability since the classification result is able to directly provide a user-interpretable indication of the reason behind the ranking. The retrieval system can provide the user with an explanation for its prioritization of search results. Second, once the model has been trained, we do not necessarily need to fine-tune the hyper parameters, i.e., the position to cut the dendrogram (for hierarchical clustering) or the initial k (for k-means clustering). Third, image labels are generated off-line at indexing time, and a clustering step at query time, which increases the system response time, is not necessary.

REFERENCES

[1] Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. 2016. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters* 80 (2016), 150–156.

[2] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009, Santorini Island, Greece, July 8-10, 2009*.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. 248–255.

[4] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, and Gang Wang. 2015. Recent Advances in Convolutional Neural Networks. *CoRR* abs/1512.07108 (2015).

[5] Kimberly A Neuendorf. 2016. *The content analysis guidebook*. Sage.

[6] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22, 10 (2010), 1345–1359.

[7] Michael Riegler, Martha Larson, Mathias Lux, and Christoph Kofler. 2014. How 'How' Reflects What's What: Content-based Exploitation of How Users Frame Social Images. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*. 397–406.

[8] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[9] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

[10] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: the new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[11] Bo Wang and Martha Larson. 2017. Beyond Concept Detection: The Potential of User Intent for Image Retrieval. In *Proceedings of the ACM MM'17 Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes MUSA'17*. to appear.

[12] Maia Zaharieva, Bogdan Ionescu, Alexandru-Lucian Gînsca, Rodrygo L.T. Santos, and Henning Müller. 2017. Retrieving Diverse Social Images at MediaEval 2017: Challenges, Dataset and Evaluation. In *Working Notes Proceedings of the MediaEval 2017 Workshop, Dublin, Ireland, September 13-15, 2017.*