Automatic Image Annotation using Weakly Labelled Web Data

Pravin Kakar, Xiangyu Wang and Alex Yong-Sang Chia

Social Media and Internet Vision Analytics Lab,
Institute for Infocomm Research,
#21-01, 1 Fusionopolis Way,
Singapore 138632.
{kakarpv, wangx, yschia}@i2r.a-star.edu.sg

Abstract. In this work, we propose and describe a method for localizing and annotating objects in images for the Scalable Concept Image Annotation challenge at ImageCLEF 2015. The unique feature of our proposed method is in its almost exclusive reliance on a single modality – visual data – for annotating images. Additionally, we do not utilize any of the provided training data, but instead create our own similarly-sized training set. By exploiting the latest research in deep learning and computer vision, we are able to test the applicability of these techniques to a problem of extremely noisy learning. We are able to obtain state-of-the-art results on an inherently multi-modal problem thereby demonstrating that computer vision can also be a primary classification modality instead of relying primarily on text to determine context prior to image annotation.

Keywords: visual recognition, scalable annotation, learning from noisy data

1 Introduction

The Scalable Concept Image Annotation challenge (SCIA) at ImageCLEF 2015 [15] is designed to evaluate methods to automatically annotate, localize and/or describe concepts and objects in images. In contrast to previous years, there have been several notable changes to the challenge. Some of them are highlighted below:

- Localization of objects within images has been introduced. As a results, the focus on more "object"-like concepts has increased this year.
- Use of hand-labelled data has been allowed. Although this is done to technically allow the use of deep learning models trained on the

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [12], it opens up the possibility of potential clean-up of the training data. Note that we have not done this in this work, but it appears to be legal within the regulatory framework of the challenge.
- The training and test sets are identical. Therefore, a method that is able to exploit the noisy training data (e.g. via data cleaning) could, in theory, benefit from potentially overfitting the training data.

From a computer vision perspective, SCIA is more challenging than the current benchmark challenge [12] in at least two senses - 1) the training data provided is fairly noisy, which makes learning a difficult problem and 2) the test set is 5× the size of [12]. While this does not indicate a clear increase in level of difficulty (for example, [12] has 4× the number of concepts of SCIA), certain aspects are definitely more demanding.

In the rest of these notes, we discuss our proposed method, including data collection, classifier training and post-processing tweaks. We also discuss the challenges posed due to the fact that test data is annotated via crowd-sourcing which adds another source of label noise to "ground-truth" data. Finally, we present our results on SCIA along with proposals for future research to improve the automatic annotation capabilities of techniques in this field.

2 Algorithm Design

As mentioned earlier, our algorithm is designed to mostly rely on visual data. We do not employ extensive ontologies to augment training data, nor do we use them during the training process, thus helping understand the importance of having a strong visual recognition pipeline. The various stages of our annotation pipeline are discussed below.

2.1 Data Collection

We do not use the provided training set for two main reasons: 1) as the training and test sets are identical, there is no penalty for overfitting on the training data, which could provide an artificial boost

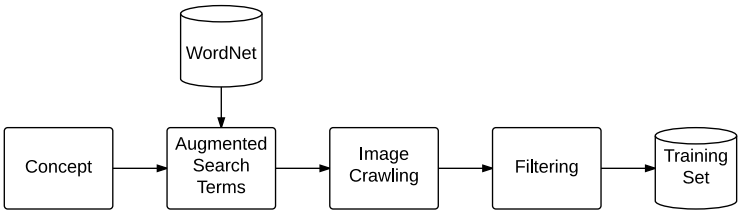

Fig. 1. Data collection pipeline

to performance results, and 2) there is little direct relationship between the target concepts and the image keywords in the training data, making it difficult to decouple the significance of a good ontology from that of a good visual learning mechanism. Therefore, we create our own training data of approximately the same size as the SCIA dataset.

The data collection pipeline is shown in Figure 1. We first consider the target concept names as keywords for which appropriate images need to be found. There is an issue of non-specificity of some of the concept names. For example, the concept "dish" can refer to both the vessel as well as the food content, although only the former is the target concept. Additionally, it is difficult to achieve both specificity and diversity using a single keyword when doing a web search for images. As an example, searching for "candy" yields generic images of candy, which while containing diverse instances of candy do not closely match single, specific instances of candy.

Both the above issues are conventionally dealt with by using ontologies to determine the coverage for each concept. We do not build our own challenge-specific ontology here, but instead simply rely on WordNet [10] to augment the individual keywords. In particular, this is done by also considering hyponyms (sub-categories) and lemmas (similar words) of the target concept. The hyponyms help target specific instances of the target concept, while the lemmas help increase the coverage of the target concept.

This augmented set of keywords per concept is then passed into an image search engine. We use Bing Image Search [7] in this pipeline.

Note that we search for the hyponyms and lemmas of the target concept by appending the target concept, in order to ensure that the correct sense of images is being search for. For example, searching for "truffle" rather than "truffle candy" results in a very different set of images that include fungi, which fall outside the scope of the target concept.

We gather up to 4000 images per target concept from our crawling engine. These images are passed through a filtering step where images that are corrupted, too small or almost completely uniform in appearance are discarded. The remaining images then form our training dataset - an automatically created, noisily labelled dataset.

2.2 Feature Extraction

For the images collected by the above process, we extract features that will be useful for image classification. We choose to use the features from the winner of the latest ILSVRC classification challenge - GoogLeNet [13], a deep learning model trained on the ILSVRC dataset and consisting of a highly-interconnected network-in-network architecture. Nevertheless, their model size is small enough to fit within our available computing resources of a single GeForce GTX 480 GPU with 1.5 GB of memory.

For each training image, we scale it down to 256×256 pixels and use the center crop of 224×224 pixels. The intuition behind this is that as the images are retrieved using specific keywords, it is likely that the object of interest is the focus of the image and should be dominant. This also reduces the computational complexity of the feature extraction process considerably. We extract features from the pooled 5B layer of the GoogLeNet model (see [13] for details), yielding a 1024-dimensional vector per training image. Each feature vector is then normalized to unit length.

We then train linear SVM classifiers [3] in a one-versus-all fashion. This is not strictly correct, as some concepts will almost certainly overlap (e.g. "face" will contain "eye", "nose", "mouth", etc.). However, making such an independence assumption greatly simplifies the learning process. Moreover, it allows us to avoid using a relationship ontology to determine the appopriate weight of every concept for each classifier. This is also in line with the goal of the challenge to

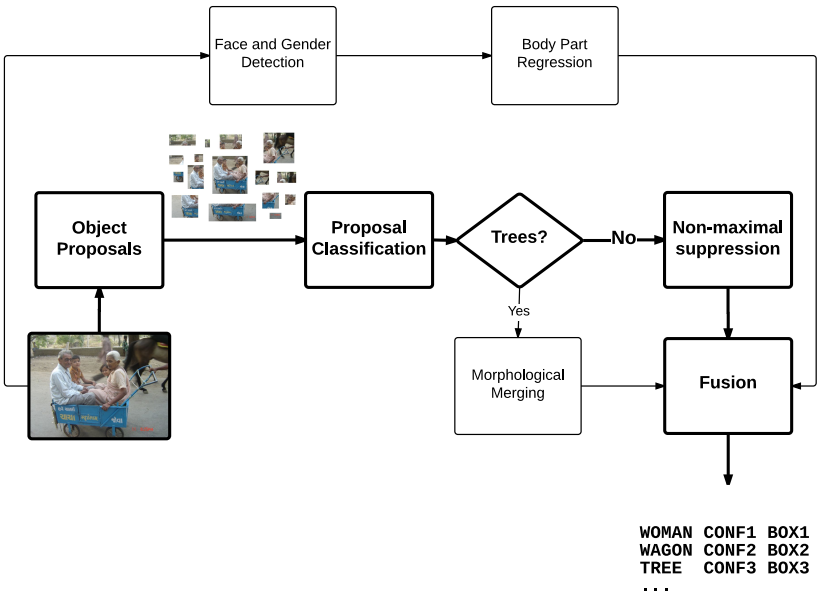

Fig. 2. Image annotation pipeline

design a scalable system, as the addition of a new target concept does not necessitate a recomputation of the weights against every existing concept. In order to manage space constraints, we uniformly sample the negative training samples for each concept, only selecting 60,000 of them.

Thus, we train a single 1024-dimensional linear classifier per target concept to use for annotating test images.

2.3 Annotation Pipeline

Figure 2 shows our processing pipeline for a single test image.

We first create object proposals using the technique of [14] that uses selective-search to find likely "object" regions. Experimentally, it is observed that many of these proposals are near-duplicates. In order to increase diversity, we limit the returned proposals to those that overlap others by at most 90%. This is found to be an acceptable value for controlling the tradeoff between increasing diversity,

and losing significant objects. We restrict the number of proposals returned to the first 150, in decreasing order of size.

Each object is then passed through the same feature extraction pipeline as in Section 2.2, and the classifiers trained therein are run to yield the most likely concepts per region. In general, non-maximal suppression is done per concept across all regions to limit the effect of overlapping proposals reporting the same concept for the same object.

There are two branches from this primary pipeline that we employ based on our observations on the SCIA development set. Firstly, we observe that many object proposals are labeled as "tree" if they contain heavy foliage. While not incorrect for the individual region, it may be incorrect for an overall image, where it is often difficult to localize a single tree. In order to mitigate this effect, we perform morphological merging for all tree boxes, taking the convex hull for each merged region as the bounding box of the tree and assigning it the highest confidence of all the merged boxes. We observe this to help improve the localization performance for the "tree" concept on the development set. We also believe that this idea can be extended to other non-uniformly shaped, difficult-to-localize concepts such as "rock", "leaf", "brick", etc. but we do not have sufficient annotations in the development data to verify the same.

Secondly, we observe that for any generic image dataset, humans are an important object. This is true for SCIA as well as for [12,8,2]. Note that this is in contrast to domain-specific datasets such as [9,11]. To this end, we use face and gender detection from [1] to detect persons with frontal faces in images. We supplement this with a simple regression to an upper-body annotation using the data from [4]. Finally, we use the information from [6] to determine the locations of various other person attributes.

A fusion step merges the results from the primary and two secondary pipelines. Specifically, person results from multiple pipelines are suppressed or modified based on overlaps between returned localizations for the same concepts. Additionally, localizations that have too high or low aspect ratios are suppressed, along with localizations that fall below a preset score threshold. Finally, if all localizations have been suppressed, then we report a single localization comprising of the entire image, corresponding to the global scene classification.

This is based on the premise that all the development set images contain at least one concept, and we extend that assumption to all the test images.

Optionally, the fusion section can also contain multiple textual refinement steps. One option is to search URL filenames for concept names, and if found, assign them to the entire image. Another approach uses correlation between concepts from an ontology. This is done to test the impact of simple context addition to the annotation pipeline. Details of this latter approach are provided in the following subsection.

2.4 Ontology and correlation

With the feature extraction and annotation pipeline, a set of bounding boxes $\{B_i\}$ are obtained for each test image. We denote the prediction scores for the target concepts in B_i as $S_i = [s_{i1}, ..., s_{im}]$, where m is the total number of target concepts. By combining the prediction scores for all the bounding boxes $\{B_i\}$, the prediction score for the image is calculated as $S = [s_1, ..., s_m]$ where $s_i = \max_j s_{ji}$.

Due to the fact that concepts do not occur in isolation (e.g. bathroom and bathtub, mountain and cliff), semantic context can be used to improve annotation accuracy. Following a way similar to [5], we adopt semantic diffusion to refine the concept annotation score. We denote $C = \{c_1, ..., c_m\}$ as the set of target concepts. Let W be the concept affinity matrix where W_{ij} indicates the affinity between concepts c_i and c_j, and D denote the diagonal node degree matrix where $D_{ii} = d_i = \sum_j W_{ij}$. Then the graph Laplacian is $\Delta = D - W$ and the normalized graph Laplacian is $L = I - D^{-1/2} W D^{1/2}$. In this problem, we measure the concept affinity based on Wikipedia dataset. Let M denote the total number of pages in Wikipedia. For concept c_i, we denote $y_{ik} = 1$ if concept keyword c_i appears in page k, and $y_{ik} = 0$ otherwise. The affinity W_{ij} between concept c_i and c_j can then be computed using Pearson product moment correlation as:

$$W_{ij} = \frac{\sum_{k=1}^{M}(y_{ik} - \mu_i)(y_{jk} - \mu_j)}{(M-1)\sigma_i \sigma_j} \tag{1}$$

where μ_i and σ_i are the sample mean and standard deviation for c_i, respectively. Based on our study, the original prediction should

be quite precise. We employ only positive correlation to boost the concepts to improve the recall.

Let $g \in \mathcal{R}^{m \times 1}$ denote the refined score vector, the values g_i and g_j should be consistent with W_{ij} (the affinity between concepts c_i and c_j). Motivated by the semantic consistency, we formulate the score refinement problem by minimizing a loss function

$$\varepsilon = \frac{1}{2} \sum_{i,j=1}^{m} W_{ij} || \frac{g_i}{d_i} - \frac{g_j}{d_j} ||^2 \tag{2}$$

The loss function can be rewritten as

$$\varepsilon = \frac{1}{2} tr(g^T L g) \tag{3}$$

The loss function can be optimized using gradient descent algorithm as

$$g = g - \alpha \nabla_g \varepsilon \tag{4}$$

where $\nabla_g \varepsilon = Lg$, and α is the learning rate.

Intially, $g = S$. By iteratively optimizing the loss function, we obtain the refined smooth score vector g for the image. A threshold τ is chosen, so that we consider concept c_i appears if $g_i > \tau$, otherwise we think the concept does not appear in the image (consequently not in any of the bounding boxes B_i). That is, for each bounding box in an image, we report the concept with the maximum confidence given the concept appears in the image.

3 Dataset Limitations

We tune the various parameters of our algorithm by validating its performance on the development set. Unfortunately, the development set (and by extension, the SCIA test set) has multiple problems that make it very difficult to correctly gauge the effect of tuning. Most of these problems arise from the limitations of crowd-sourcing ground-truth annotation, and need to be addressed to make SCIA a more consistent evaluation. We summarize the major issues involved below with the 4 I's.

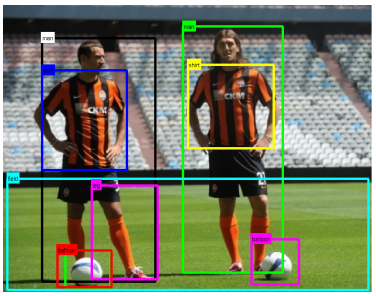

Fig. 3. Image demonstrating inconsistency, incompleteness and incorrectness of annotations. Annotations are from the SCIA development set.

Inconsistency Many images in the development set exhibit inconsistent annotations. An example of this is shown in Figure 3. Despite there clearly being 4 legs in the picture, only 1 is annotated. This is inconsistent as none of the unannotated instances are any less of "leg"s than the one annotated. Other examples of inconsistencies seen in the development set include unclear demarcations between when multiple instances of the same concept are to be grouped into a single instance or vice versa and annotating partially-visible instances in some images while not annotating completely visible instances in other images.

Incompleteness Several images in the development set are incompletely annotated. This is most prevalent in the case of humans where various body parts are skipped altogether in the annotations. Apart from this, there appears to be a certain level of arbitrariness in choosing which concepts to annotate. For instance in Figure 3, "shirt" is annotated, but "short pants" is not when clearly both have about the same level of "interestingness". Additionally, concepts like "chair", "sock", "shoe", "stadium", etc. which are also present in the image are not annotated. This makes it extremely difficult to judge the performance of a proposed technique on the development set. Moreover, it seems to run counter to the challenge assertion that the proportion of missing or incomplete annotations is insignificant.

Incorrectness Althought not as prevalent as the previous two problems, many annotations are incorrect. In Figure 3, the two balls are labelled as balloons, which is clearly wrong. There are other cases of

wrong annotations in items of clothing (shirt/jacket/suit) as well as gender and age of persons ("man" labelled as "male child", "woman" labelled as "man", etc.).

Fig. 4. Image demonstrating impossibility of annotations.

Impossibility This issue is the least prevalent of the four discussed in this section. The image shown in Figure 4 was flagged by our image annotation pipeline as having a very large number of object instances. It can be seen that the image contains more than 200 faces. This implies that there are more than 100 instances of at least one of "man" or "woman"[1]. Within the rules of the challenge, each concept is limited to 100 instaces, making it impossible to annotate all instances correctly. Grouping multiple instances into a single instance, if one were inclined to do so, is not straightforward as there is no clear group of men and women as in some other images.

4 Evaluation

In this section, we evaluate the performance of different settings of the algorithm on the development and test datasets. It is to be noted that there appear to be significant differences in the quality of the annotations between the two sets, so results on one are not indicative of results on the other. Moreover, as there were no particulars

[1] A quick inspection of the image shows no children, eliminating the possibility of instances of "male child" and "female child"

provided about the measures being used for evaluation beyond "performance" on both concepts and images, we used the F-score as a measure on the development set, which formed the basis of 2 out of 3 measures in the previous iteration of the challenge. As it turns out, the evaluation measure used on the test set was the mAP and so, results between the development and test sets are again not directly comparable. These statistics are shown in Table 1.

Table 1. Performance statistics for various runs.

Method	Development set		Test set	
	F (0.5)	F (0)	mAP (0.5)	mAP (0)
Better precision (BP)	**0.1463**	0.2921	0.3039	0.4571
Better recall (BR)	0.1462	0.2934	0.2949	0.4466
BP + $\geq$ 1 pred (BP1)	0.1459	0.2940	0.2949	0.4466
BR + $\geq$ 1 pred (BR1)	0.1459	0.2927	0.3039	0.4569
BR1 + URL search	0.1450	**0.2948**	0.2948	0.4465
BR1 + Agg. NMS	0.1380	0.2894	0.3536	0.5045
BR1 + hair + mouth + URL search	0.1442	0.2923	0.5707	0.7603
BR1 + hair + mouth	0.1450	0.2899	0.5786	0.7685
BR1 + body parts + URL search	0.1279	0.2615	0.6024	0.7918
BR1 + face parts	0.1407	0.2818	**0.6595**	**0.7954**
Runner-up	NA	NA	0.5100	0.6426

We run two base versions of our pipeline, one aimed at garnering better precision (BP) and one aimed at getting better recall (BR). These are shown in the first two rows of the table. Following this, we notice that in some images, no concepts were predicted as their confidences scores fell below their respective thresholds. In these cases, we forced at least 1 prediction to be made ($\geq$ 1 pred.) giving rise to two more variants, BP1 and BR1.

URL-search corresponds to the URL filename-concept name match discussed earlier. Agg. NMS refers to aggressive non-maximal suppression that employs a NMS threshold of 0, resulting in all overlapping bounding boxes for the same concept to be reduced to a single one. From human attributes, we either report hair + mouth, which showed no deleterious effects on the development set in the face of incomplete annotations, or face parts which also adds eyes and noses, or body parts which further adds in faces, heads, necks and arms. In the case of ontologies, while we obtain slightly better

results on the development set, output errors in the submission cause the performance to be quite low, which is an outlier.

From the results, it can be seen that all human attributes significantly help boost performance. Moreover, URL-search causes a drop in performance, while aggressive NMS again boosts performance. Hence, a possible solution that yields even better performance could be BR1 + Agg. NMS + body parts.

It is also to be noted that the runner-up in the challenge attains a performance about 15% lower than ours. As the details of their technique are not available, it is difficult to pinpoint the cause of the large difference, but we believe that the use of an external training set, combined with human part extraction played an important role.

5 Conclusions and Future Work

In this work, we have presented our proposed method for the ImageCLEF Scalable Concept Image Annotation challenge. Our method places heavy emphasis on the visual aspect of annotating images and demonstrates the performance that can be achieved by building an appropriate pipeline of state-of-the-art visual recognition techniques. The interconnections between the techniques are modified and enhanced to improve overall annotation performance by branching off secondary recognition pipelines for certain highly common concepts.

We also highlight the limitations of the current challenge dataset with respect to the ground-truth annotations, categorizing the major shortcomings. Despite these and our technique's general lack of reliance on textual data, we are able to outperform competing methods by a margin of at least 15%. In the future, we plan to refine our annotation pipeline based on the analysis of the results. As most of the target concepts in this iteration of the challenge were localizable in a well-defined manner, it will be interesting to examine localization for other, more abstract concepts. We also hope to combine advances in natural language processing and semantic ontologies to appropriately weigh training instances in learning classifiers as well as look at the problem from a multi-modal point of view.

References

1. Open biometrics, http://openbiometrics.org/
2. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html
3. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. The Journal of Machine Learning Research 9, 1871–1874 (2008)
4. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8. IEEE (2008)
5. Jiang, Y.G., Wang, J., Chang, S.F., Ngo, C.W.: Domain adaptive semantic diffusion for large scale context-based video annotation. In: Computer Vision, 2009 IEEE 12th International Conference on. pp. 1420–1427. IEEE (2009)
6. Jusko, D.A.: Human figure drawing proportions, http://www.realcolorwheel.com/human.htm
7. Microsoft: Bing image search (2015), http://www.bing.com/images
8. Opelt, A., Pinz, A., Fussenegger, M., Auer, P.: Generic object recognition with boosting. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(3), 416–431 (2006)
9. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3498–3505 (2012)
10. Princeton University: About wordnet (2010), https://wordnet.princeton.edu/wordnet/
11. Quattoni, A., Torralba, A.: Recognizing indoor scenes. Computer Vision and Pattern Recognition (2009)
12. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) (2015)
13. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. CoRR abs/1409.4842 (2014), http://arxiv.org/abs/1409.4842
14. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision 104(2), 154–171 (2013)
15. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., de Herrera, A.G.S., Bromuri, S., Amin, M.A., Mohammed, M.K., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., del Mar Roldán García, M.: General Overview of ImageCLEF at the CLEF 2015 Labs. Lecture Notes in Computer Science, Springer International Publishing (2015)