Automatic Image Annotation with Ensemble of Convolutional Neural Networks

Anastasia Timofeeva[1] [0000-0001-5813-582X], Oleksii Kudin[1] [0000-0002-5917-9127]

[1] Zaporizhzhya National University, Zaporizhzhya, Ukraine
nastyatima1@gmail.com, avk256@gmail.com

Abstract. This paper discusses the models and methods of machine learning that are employed to solve the problem of automatic image annotation. Today, the systems which have the ability to extract meaning from visual data are increasingly developed and used both in academia and industry. One of the practically important directions within the scope of this problems is the development of automatic systems for understanding of visual scenes. In this paper, we propose a brief survey of the state-of-the-art machine learning approaches and methods that have been suggested for automatic image annotation. We study the mathematical foundations of the overviewed methods and analyze their strengths and limitations. Further, we develop a proof-of-concept system for the image annotation using convolutional neural networks and construct a neural network ensemble using the snapshot approach. In the image processing stage, we resize image for computation acceleration and use image augmentation method. In addition, we outline a direction for further development of image annotating systems based on both theoretical and experimental models.

Keywords: Automatic Image Annotation, Convolutional Neural Networks, Ensemble Methods, Resizing

1 Introduction

Automatic image annotation (AIA) is the process in which a computer automatically specifies metadata to a digital image. Typically, in this process, metadata is assigned to an image in the form of titles or keywords. Automatic image annotation has a growing area of applications in computer vision, data meaning and web search.

Image processing differs from text data processing. Therefore, the development of effective methods for navigating and searching in large image databases is a complex problem.

The objective of the paper is to develop a concept of an automatic annotation system that includes transfer learning and snapshot ensemble as approach of combining neural networks.

The novelty of the article is due to adapting pre-trained neural net VGG-16 [1] for image annotation using transfer learning and using snapshot ensemble method to improve generalization ability of the system.

CEUR-WS.org/Vol-2403/paper3.pdf

2 Related Works

Text-based image annotation continues to be an important practical as well as fundamental problem in the computer vision and information retrieval. A number of approaches and surveys have been proposed in the past to address the annotation task.

According to the research conducted in [2] AIA methods were classified into five categories: generative model-based image annotation, nearest neighbor-based image annotation, discriminative model-based image annotation, tag completion-based image annotation, deep learning-based image annotation.

In order to improve the annotation performance of existing AIA approaches, a hybrid AIA approach based on visual attention mechanism (VAM) and the conditional random field (CRF) is developed in [3]. VAM is employed to determine salient regions of the image, and CRF is applied to optimize the initial label set. The experimental results confirm that the suggested hybrid AIA approach has the highest annotation performance.

The general approach uses SVM combined with KNN, as in works [3, 4]. The objective of SVM is to create a model that only targets the value of data instances with attributes in a test set.

Ameesh Makadia et al. in [5, 6] described Multiple Bernoulli Relevance Model, which based on the CRM (Continuous-Space Relevance Model). The key idea is to transform the image into a set of attribute vectors. Further, these vectors are employed simulation of possible words for annotation. Venkatesh N. et al. [7] is created the hybrid model SVM-DMBRM, which combines the method of reference vectors acting as a discriminative model and the method of discrete multiple Bernoulli matching of the generating model. In [8] authors used Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) methods, where images are segmented into superpixels, and visual features are extracted from each superpixel region. Boosted classifiers are then trained for each class, and the output of boosted classifiers are quantized as boosted visual words.

In works [9, 10, 11], convolutional neural networks (also the model AlexNet) are employed for image annotation. A very deep convolutional neural network extracts visual imaging features, ranging from very basic functions, such as edge detectors, and then gradually creating more complex features, such as detecting a shape.

Also Convolutional Neural Networks (CNNs) can be combined with Recurrent Neural Networks (RNNs). So, in [12, 13] the LSTM model (Long short-term memory) was used as RNN, which is responsible for creating an annotation to the image. This model provides accurate annotations that cover the scene of the image, as well as information about all objects and things in this image. Here the image is described not just by a set of keywords, but an associated sentence is generated.

Jiwei Hu et al. in [14] designed a new hierarchical model for image annotation, based on constructing a novel, hierarchical tree, which consists of exploring the relationships between the labels and the features dividing labels into several hierarchies for efficient and accurate labeling.

3 Segmented and Annotated IAPR-TC12 Dataset

The IAPR-TC12 collection, is an established image retrieval benchmark composed of about 20,000 images manually annotated with free-text descriptions with hierarchical organization in three languages; 96,234 regions compose the segmented collection, for which 256 labels have been used. The hierarchy was manually defined by the authors after carefully analyzing the images, the annotation vocabulary and the vocabulary of manual annotations. The vocabulary plays a key role in the annotation process because it must cover the most of concepts that we can find in the collection of images. At the same time, the vocabulary shouldn't be too large because AIA performance depends on the number of considered labels. The annotation vocabulary was organized mostly using *is-a* relations between labels. However, relations like *part-of* and *kind-of* were also included.

According to the suggested hierarchy, an object can be in one of six main branches: *'animal'*, *'landscape'*, *'man-made'*, *'human'*, *'food'*, or *'other'*. This is the high level of the hierarchy. The ten more common labels are 'sky-blue' (5,176), 'man', (3,634), 'group-persons' (3,548), 'ground' (3,284), 'cloud' (2,770), 'rock' (2,740), 'grass' (2,609), 'vegetation' (2,455), 'woman' (2,339), and 'trees' (2,291) [15].

4 System Concept

The automatic image annotation system is based on key ideas transfer learning and snapshot ensemble as approach. Pre-trained convolutional neural net VGG-16 is used for tagging raw image by six classes: *'animal'*, *'landscape'*, *'man-made'*, *'human'*, *'food'*, or *'other'*. The main purpose of this labels is to facilitate and improve the annotation process by reducing labels amount. Transfer learning is used for train VGG-16 on the IAPR-TC12 dataset. Snapshot ensembles approach [16] is used to combine multiple neural networks predictions, and, hence to improve the accuracy. The system architecture that we develop is as follows: one layer of VGG-16 Keras model, then, two full connected Dense layers. Feature vectors are evaluated by VGG-16 and feeded to Dense layers for final classification.

Snapshot Ensembling produces an ensemble of accurate and diverse models from a single training process. At the heart of Snapshot Ensembling is an optimization process which visits several local minima before converging to a final solution. We take model snapshots at these various minima and average their predictions at test time.

5 Classification Experiments

The choice of correct metrics and loss functions is the significant stage in the multi-label classification. In this stage we use binary crossentropy as loss function and binary accuracy as accuracy metric.

Initial classification experiments (see Fig. 1) show adequate accuracy and loss rate.

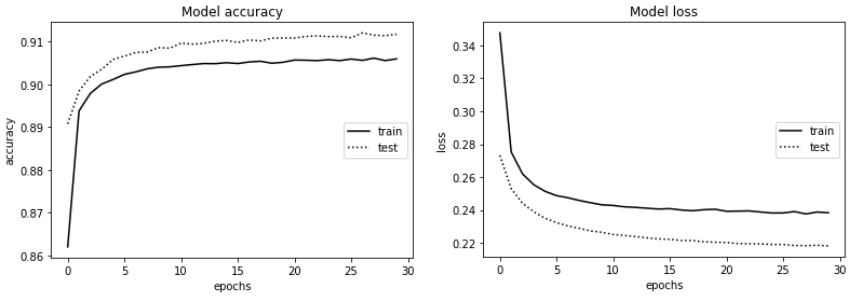

Fig. 1. Accuracy and loss rate

In [15], authors present classification experiments and results using several classifiers on the IAPR-TC12 dataset. The highest percentage of correct annotations approximately is 85% accuracy obtained with the random forest classifier on the dataset with 5 classes. Therefore, we can say that the accuracy of our system is sufficient compared to other publications.

The multi-label confusion matrix allows to get the F_1 score for each label (see Table 1).

Table 1. F_1 score (%) on the IAPR-TC12 dataset for each label

Method	IAPR-TC12
Single neural net	[44%, 89%, 20%, 5%, 0%, 15%]
Snapshot Ensemble	[52%, 89%, 23%, 17%, 0%, 20%]

An unbalanced data set caused low scores at the Table 1. Thus further research may be associated with building a balanced data set and developing a metric for the multi-label classification problem.

6 Conclusion

In this paper has been suggested a concept for the development of automatic image annotation system. Transfer learning has been used to train the VGG-16 model for the image annotation problem. Snapshot ensemble has been employed as approach to build the neural networks ensemble. Initial classification experiments show appropriate accuracy and loss rate also multi-label confusion matrix had been evaluated.

The prospect of future researches is in using the genetic algorithm for hyperparameter optimization, e.g. number of layers, types activations, learning rate etc.

References

1. K. Simonyan and A. Zisserman: Very Deep Convolutional Networks for Large-Scale Image Recognition. In ICLR, 2015.
2. Qimin Cheng, Qian Zhang, Peng Fu, Conghuan Tu, Sen Li.: A survey and analysis on automatic image annotation. Pattern Recognition 79, 242–259, (2018).
3. Cong Jin, Qing-Mei Sun, Shu-Wei Jin: A hybrid automatic image annotation approach. Multimedia Tools and Application, https://doi.org/10.1007/s11042-018-6742-6, (2018)
4. Guanglei Chu ,Kai Niu, Baoyu Tian: Automatic image annotation combining SVM and KNN algorithm. IEEE 3rd International Conference on Cloud Computing and Intelligence Systems 27-29 Nov. (2014).
5. Ameesh Makadia, Vladimir Pavlovic, Sanjiv Kumar: Baselines for Image Annotation. Article in International Journal of Computer Vision. (2010).
6. S. L. Feng, R. Manmatha and V. Lavrenko: Multiple Bernoulli Relevance Models for Image and Video Annotation. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR, (2004).
7. Venkatesh N. Murthy, Ethem F. Can: A Hybrid Model for Automatic Image Annotation. Proceedings of International Conference on Multimedia Retrieval, (2014).
8. Guo Qiaojin, Li Ning, Yang Yubin and Wu Gangshan: Supervised LDA for Image Annotation 2011 IEEE International Conference on Systems, Man, and Cybernetics, (2011).
9. Image Annotation Using Deep Learning: A Review 2017 International Conference on Intelligent Computing and Control, (2017)
10. Ayushi Dutta·Yashaswi Verma·C. V. Jawahar: Automatic image annotation: the quirks and what works. Multimedia Tools and Applications, Vol. 77, Issue 24, 31991–32011, (2018).
11. Yanchun Ma Yongjian, Liu Qing, Xie Lin Li: CNN-feature based automatic image annotation method. Multimedia Tools and Applications, Vol. 78, Issue 3, 3767–3780, (2019).
12. Sonu Pratap, Singh Gurjar, Shivam Gupta: Automatic Image Annotation Model Using LSTM Approach, Signal & Image Processing, An International Journal (SIPIJ) Vol.8, No.4, (2017).
13. An Adaptive Recognition Model for Image Annotation, Article in IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews), (2012).
14. Jiwei Hu, Kin-Man Lam,Ping Lou and Quan Liu: Constructing a Hierarchical Tree For Image Annotation, The Hong Kong Polytechnic University, (2017).
15. Hugo Jair Escalante, Carlos A. Hernandez, Jesus A. Gonzalez,A. Lopez-Lopez, Manuel Montes, Eduardo. Morales, L. Enrique Sucar, and Luis Villasenor: Segmented and Annotated IAPR-TC12 benchmark, (2009).
16. Huang G., Li Y., Pleiss G., Liu Z., Hopcroft J.E., Weinberger K.Q. Snapshot Ensembles: Train 1, Get M for Free. Retrieved from http:// arxiv.org/abs/1704.00109, (2017).