Deep learning with Siamese networks for instance search or identification

Matthias Kohl

May 10, 2017

Abstract

TODO

1 Introduction

1.1 Motivation

The research presented here is motivated by the GUIMUTEIC project, a collaborative project between industry and LIG. The aim is to develop a smart audio-guide for touristic or cultural sites. TODO cite project/LIG/etc

In practice, the final product should offer an augmented reality interface to the user, with information about the object or objects the user is looking at.

One part of the development consists in finding ways of identifying objects the user is looking at. There are multiple possibilities, for example based on the geo-localization of the user and other sensors. In our research, we focus on the recognition of objects based only on visual clues.

1.2 Research problem

More specifically, we are interested in the following problem: We are given a collection with reference images for each object, or instance, to be recognized. The task is to develop a system that, given an image of one of the instances, can decide which instance the image represents. We assume that, on average, less than ten images are available for each instance. We will refer to this problem as instance retrieval in the following.

There are a few problems similar to instance retrieval. For one, there is the image classification problem: we are given a collection of images where each image is assigned a class, like dog or fridge. The task is to develop a system that, given an image, decides which class the image represents. This problem is of course similar if we simply consider each instance to be a separate class. However, in classification, we usually assume to have many images per class: hundreds or even thousands. This means our problem is closer to few-shots classification, where only few images are available for each class.

Another similar problem is image retrieval. In image retrieval, we are given a collection of reference images and a query image. We aim to rank the reference images by similarity to the query image. Usually, in image retrieval, reference images are not labeled or loosely labeled and many images are irrelevant to the query image. The challenge is to rank the most similar images on top. Instance retrieval is related to image retrieval in that we aim to develop a notion of similarity between images. However, in instance retrieval, we do not care about the rank of the returned images, since we only consider the highest ranked image, which should represent the instance to retrieve.

TODO possibly elaborate here

1.3 Challenges

2 State of the art

Previously, most systems were based on a bag-of-words approach in order to match images [1]. Combined with better techniques of matching the bag-of-words descriptors, this approach was previously the state-of-the-art [2].

Recently, the state-of-the-art has drastically improved due to the addition of learned features, based on convolutional neural networks (CNNs). This new approach greatly improved mean average precision scores [3] of the system on the same datasets.

The general trend is to move from an approach based on a combination of engineered features (bag-of-words combined with support vector machines or SVMs) to a more end-to-end learning of the matching of two images.

For this, the Siamese architecture is used, where the convolutional features of two or more CNNs with shared weights are combined and a multi-way loss is optimized, which discriminates between the features of two or more images. This concept was first introduced by Chopra et al [4] to learn a dissimilarity metric between two images.

For a learning-to-rank setting, this idea was extended to a triplet loss by Weinberger et al [5]. This loss is better suited when learning to rank, rather than instance search, but it also allows for a more stable convergence, which is beneficial in both cases.

Using this triplet loss has achieved state-of-the-art results in both face recognition [6] and image retrieval [3]. This is mostly due to the usage of far deeper CNNs, moving from architectures such as AlexNet [7] to VGG [8] and finally Inception [9] and ResNet [10].

The higher depth of networks like ResNet and Inception as compared to AlexNet allows for higher regularization and thus less over-fitting to a specific dataset for these architectures. Batch normalization increases this effect even more. This is a desirable trait for image retrieval, as the variability within each instance is too small to learn classification directly. So the better we can generalize features learned from a bigger dataset to the reference dataset, the better performance we can expect.

3 Evaluation

We are using the following datasets in our experiments:

1. CLICIDE: dataset of photographs taken in an art museum, consisting exclusively of paintings. The dataset is characteristic because the different images for each instance consist of one global view of the instance and multiple sub-parts.

   Number of reference images: 3245 (the full dataset contains 3452 images with 207 images of walls and no meaningful instance) Number of test images: 165 (the full dataset contains 177 test images, out of which 165 share their instance with at least one image of the reference set) Number of instances: 464

2. GaRoFou_I: dataset of photographs of an art museum, consisting of display cabinets, which contain sculptures, rocks and various types of objects.

   Number of reference images: 1068 Number of test images: 184 (all are used for evaluation) Number of instances: 311

3.1 Evaluation metrics

3.1.1 Metrics definitions

Precision@*k* For a test image and m reference images and a ranking of the reference images $I_1, \dots, I_m$, we define the number of relevant images at k with $k \leq m$: N_k^{rel} is the number of images in the sub-ranking $I_1, \dots, I_k$ from the same instance than the test image.

We then define Precision at k as:

$$Precision@k = \frac{N_k^{rel}}{k} \tag{1}$$

MAP As above, for a test image, m reference images and a ranking of the reference images, we define the average precision:

$$AP = \frac{1}{N_m^{rel}} \sum_{k=1}^{m} Precision@k \tag{2}$$

The MAP score is defined as the mean of the AP score over all test images.

3.1.2 Used metrics

As of now, we use the mean $Precision@1$ to evaluate the system. It is the mean of the $Precision@1$ for all test images. This is what we are most interested in: in the context of a search system for retrieving art in a museum, we are only interested in one result, as the user should not make a choice out of several results.

To be comparable with other papers in the field, we will implement the MAP score as well.

References

[1] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007.

[2] A. Mikulik, M. Perdoch, O. Chum, and J. Matas, "Learning Vocabularies over a Fine Quantization," *International Journal of Computer Vision*, vol. 103, pp. 163–175, May 2013.

[3] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep Image Retrieval: Learning Global Representations for Image Search," in *Computer Vision – ECCV 2016*, pp. 241–257, Springer, Cham, Oct. 2016.

[4] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 539–546 vol. 1, June 2005.

[5] K. Q. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *Advances in neural information processing systems*, vol. 18, p. 1473, 2006.

[6] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

[8] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Sept. 2014. arXiv: 1409.1556.

[9] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *arXiv:1602.07261 [cs]*, Feb. 2016. arXiv: 1602.07261.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015. arXiv: 1512.03385.