Master of Science in Informatics at Grenoble
Master Informatique
Specialization Data Science

Deep learning with Siamese networks for instance search or identification

Matthias Kohl

June 20 (TODO), 2017

Research project performed at LIG - MRIM

Under the supervision of:
Georges Quénot, Jean-Pierre Chevallet

Defended before a jury composed of:
Head of the jury (TODO)
Jury member 1
Jury member 2

June 2017

Abstract

Your abstract goes here... TODO

Acknowledgement

I would like to express my sincere gratitude to .. for his invaluable assistance and comments in reviewing this report... Good luck :) TODO

Résumé

Your abstract in French goes here... TODO

Contents

— 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.

1.3 Challenges

For all of the problems described above, deep learning approaches based on convolutional neural networks (CNNs) have recently obtained the state-of-the-art results. One of the drawbacks of CNNs is that they require large amounts of data to be trained.

In our research problem in particular, we do not have large enough amounts of data available to fully train a CNN. However, we aim to develop a system that can learn a descriptor specific to the reference images, as the system only needs to recognize and differentiate between instances of that particular dataset. So the system should be tuned to the dataset. Since we usually have less than ten images per instance, this is an important challenge to overcome.

Another challenge comes from the nature of the datasets: we were not able to find similar datasets in the literature, with only a few, clean images, for each instance and many instances. Common datasets in image retrieval like Oxford5k [TODO], Paris6k [TODO] or Holidays [TODO] contain many images per instance, but a lot of noise, as they are used to evaluate whether a system can robustly retrieve the correct images out of a set of random images. This means that the approaches used for image retrieval datasets may not work as well for instance retrieval.

1.4 Contributions

We made the following three major contributions:

1. We analyze existing approaches to image retrieval, classification and determine their shortcomings in our problem setting.

2. Based on these shortcomings, we propose a novel approach to improve results in our problem setting.

3. We evaluate the novel approach and compare with other approaches.

— 2 —
State of the art

2.1 SIFT and bag-of-words

Until recently, the state of the art in image retrieval and matching was based on the idea of bag-of-words [11] [10]. The idea behind the approach is to represent an image as a histogram, or collection of frequencies, of visual words. The visual words should be small, representative patches of the images in the dataset. Usually, these visual words are obtained in multiple steps. First, local features are extracted and encoded. The most common feature used is SIFT [9], a 128-dimensional vector representing scale-invariant features. Then, the features are extracted for all images and clustered into clusters of similar features. The representative for each cluster is the center of the cluster, i.e. the mean of all features falling into that cluster. Each image can then be represented as a histogram of the occurrences of these representative features. This histogram forms the image descriptor, which has as many dimensions as there are representative features and clusters.

Finally, for image classification, a classifier can be learned from the descriptors of all images in the dataset. On the other hand, for image retrieval or matching, the descriptors can be directly matched against each other, based on some similarity measure. In both cases, it can be useful to project the descriptors into a Hilbert space using a different similarity measure than euclidean distance in the original descriptor space. For classification, this is done by employing an SVM classifier with a non-linear kernel [16]. Among the most popular kernels are the radial basis function [14] and the chi-squared function [19].

Until recently, this approach has obtained state of the art results for image retrieval tasks [10].

2.2 Deep learning with CNNs

Starting with the results of AlexNet for image classification in the 2012 ImageNet challenge [7] [13], image classification tasks have been dominated by CNNs, learned using large amounts of data.

A general trend in image related tasks is to move to an end-to-end approach, where the final objective is directly optimized using gradient descent and the gradient is back-propagated to all previous parts of the system. In contrast, the bag-of-words model requires a choice of features (SIFT, ORB (TODO mention ORB before), ...), a choice of the method for clustering features (k-means), a choice of the classifier (AdaBoost [1], SVM, ...), as well as a choice of the kernel if an SVM classifier is used.

Using a CNN, features are extracted at a low abstraction level by the first convolutional layers, then higher level features are formed by combining low level features from the previous layers. Finally, high-level features are combined into a classifier by linear layers. The advantage of this approach is that the features are learned at all abstraction levels. Furthermore, the modularity of the approach allows us to easily transfer the lower level features learned from a large dataset to a smaller dataset, where there may not be enough data to efficiently learn lower level features.

The following sections describe the high-level architecture of the most widely used CNNs in classification and image retrieval.

2.2.1 AlexNet

AlexNet [7] contains five convolutional layers and three linear layers. The first convolutional layer has a large kernel and stride to quickly increase the receptive field of the consequent filters. The first two convolutional layers are followed by a pooling layer to further increase the receptive field.

Finally, an important improvement for AlexNet to avoid over-fit is the addition of dropout layers [5] before the two first linear layers.

Apart from that, AlexNet is a simple adaptation of LeNet by LeCun et al [8] for the ImageNet challenge.

2.2.2 VGG

The VGG architecture, introduced by Simonyan et al [17] is the first to use exclusively 3×3 convolutional kernels. This means that the receptive field of the first filters is much smaller, and thus many more layers are needed, as well as more pooling layers. The most popular VGG architectures are VGG-16 and VGG-19, having 16 and 19 layers in total respectively.

Using smaller kernels in convolutions and consequently more layers allows to reduce the number of trainable parameters used by the network, while increasing the number of non-linear layers. This was shown to allow better generalization properties, and thus less over-fit [17].

2.2.3 ResNet, Inception, DenseNet

The idea of increasing the number of layers is further developed by the ResNet, DenseNet and Inception architectures. However, naively increasing the number of layers leads to the problem of vanishing gradient [4].

All of these networks overcome the vanishing gradient problem by introducing skip connections, where the output of a previous layer is added to the output of a layer, skipping some number of layers in between.

The ResNet architecture by He et al [3] always uses blocks of 3 layers along with a skip of those 3 layers. The smaller types of ResNet use blocks of 2 layers.

The Inception architecture by Szegedy et al [18] uses a combination of skip connections of different length and different types.

Finally, DenseNet by Huang et al [6] takes the skip connection idea one step further: the input of each layer is dependent on the output of all n previous layers to some extent.

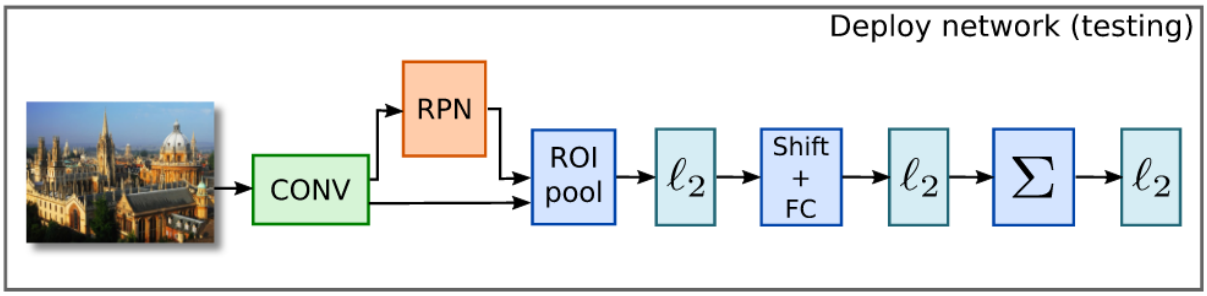

Figure 2.1: Architecture of a CNN network for image retrieval

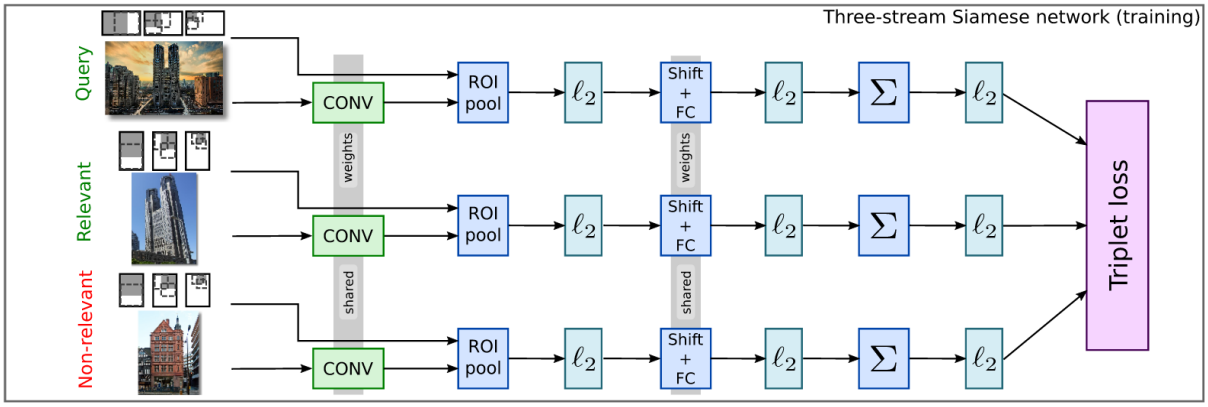

Figure 2.2: Siamese architecture of a CNN for image retrieval used in training

Finally, all of the very deep architectures use batch norm layers to further reduce over-fit: a batch norm layer simply normalizes the features over each batch, and then applies a learnable scaling and shifting.

2.3 Image retrieval using CNNs

For image retrieval, the current state of the art is set by Gordo et al [2]. It is based on an end-to-end approach. The goal is to learn a global descriptor for images that is well suited for comparing images.

Figure 2.1 shows the detailed architecture used. It first extracts the convolutional features of a pre-trained CNN. Then, a Region Proposal Network (RPN) [12] is used to extract the regions of interest. For each region of interest, a shifting and linear layer are used to reduce the dimensionality of the descriptor. The final descriptor is simply a normalized sum of the region-wise descriptors. This network can be learned end-to-end and the obtained descriptor achieves state of the art results, which can be even further improved by using query expansion and database side feature augmentation (TODO either remove or mention before).

Gordo et al [2] found that using very deep networks outperforms the shallower networks. The state of the art results are thus set by a very deep ResNet architecture.

2.3.1 Siamese networks and triplet loss

Figure 2.2 shows the architecture used to train the CNN that produces a descriptor for images, as can be seen in Figure 2.1.

The major difference is that during training, a Siamese architecture is used: the CNN is evaluated with multiple images, using the same weights for each. Then, a combined loss is obtained from the descriptors. Finally, the loss is back-propagated once through all streams of the network, since shared weights are used for all images.

In the case of this particular architecture, three images are evaluated and a triplet loss is used. This triplet loss has previously been used in face recognition tasks by Schroff et al [15]. The triplet loss was first introduced by Weinberger et al [20] as a way of learning the best suited distance metric in a k-nearest neighbor classification problem.

Bibliography

[1] Yoav Freund and Robert E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[2] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep Image Retrieval: Learning Global Representations for Image Search. In *Computer Vision – ECCV 2016*, pages 241–257. Springer, Cham, October 2016.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. arXiv: 1512.03385.

[4] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[5] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[6] Gao Huang, Zhuang Liu, Kilian Q. Weinberger, and Laurens van der Maaten. Densely Connected Convolutional Networks. *arXiv:1608.06993 [cs]*, August 2016. arXiv: 1608.06993.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[9] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[10] Andrej Mikulik, Michal Perdoch, Ondřej Chum, and Jiří Matas. Learning Vocabularies over a Fine Quantization. *International Journal of Computer Vision*, 103(1):163–175, May 2013.

[11] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015.

[14] Bernhard Scholkopf, Kah-Kay Sung, Christopher JC Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio, and Vladimir Vapnik. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing*, 45(11):2758–2765, 1997.

[15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[16] John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[17] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, September 2014. arXiv: 1409.1556.

[18] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv:1602.07261 [cs]*, February 2016. arXiv: 1602.07261.

[19] Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *IEEE transactions on pattern analysis and machine intelligence*, 34(3):480–492, 2012.

[20] Kilian Q. Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems*, 18:1473, 2006.