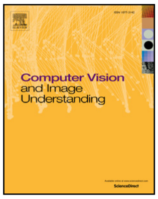
Computer Vision and Image Understanding 000 (2017) 1–14

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Harnessing noisy Web images for deep representation

Phong D. Vo*, Alexandru Ginsca, Hervé Le Borgne, Adrian Popescu

Vision & Content Engineering Laboratory, CEA LIST, France

ARTICLE INFO

Article history:
Received 7 June 2016
Revised 4 January 2017
Accepted 27 January 2017
Available online xxx

Keywords:
Representation learning
Deep learning
Convolutional networks
Semi-supervised learning
Domain adaptation
Noisy data

ABSTRACT

The keep-growing content of Web images is probably the next important data source to scale up deep neural networks which recently surpass human in image classification tasks. The fact that deep networks are hungry for labelled data limits themselves from extracting valuable information of Web images which are abundant and cheap. There have been efforts to train neural networks such as autoencoders with respect to either unsupervised or semi-supervised settings. Nonetheless they are less performant than supervised methods partly because the loss function used in unsupervised methods, for instance Euclidean loss, failed to guide the network to learn discriminative features and ignore unnecessary details. We instead train convolutional networks in a supervised setting but use weakly labelled data which are large amounts of unannotated Web images downloaded from Flickr and Bing. Our experiments are conducted at several data scales, with different choices of network architecture, and alternating between different data preprocessing techniques. The effectiveness of our approach is shown by the good generalization of the learned representations with new six public datasets.

1. Introduction

For a long time the vision community has been striving for the quest of creating human-like intelligent systems. Recently the resurgence of neural networks (Bengio et al., 2007; Hinton, 2005; Hinton et al., 2006) has first led to a revolution in computer vision (Ciresan et al., 2012; Krizhevsky et al., 2012; Razavian et al., 2014a; Simonyan and Zisserman, 2014; Szegedy et al., 2015), as well as in other areas including reinforcement learning (Mnih et al., 2013), speech recognition (Graves et al., 2013), and natural language processing (Mikolov et al., 2013). For the most part, those neural network models are supervised, which require lots of labelled training data hence pose scalability challenges. This paper proposes an alternative to train deep neural networks using massive amount of unannotated Web images. Such an approach is sometimes named *webly supervised learning* (Chen and Gupta, 2015; Joulin et al., 2016; Zhang et al., 2015).

Convnets became the *de facto* representation learning method in image classification thanks to its excellent generalization ability. A convnet can be seen as an end-to-end feature mapping, starting from raw pixel intensities then learn a robust representation through many hidden layers of different types. At the top of the network, the last layer is usually a loss function which is specific to each problem. Giryes et al. (2015) proved that under random Gaussian weights, deep neural networks are distance-preserving mappings with a special treatment for intra- and inter-class data.

An important property of convnets is that they learn distributed representation, meaning that one concept is represented by multiple neurons and each neuron participates in the representation of more than one concept. Distributed representation is much more expressive than a compact representation, in the sense that it has fewer hidden units (Delalleau and Bengio, 2011) and much more regions of linearity (Montúfar et al., 2014). Theoretical justifications furthermore show deep networks as a class of universal approximates (Cybenko, 1989; Hornik et al., 1989). A more recent work (An et al., 2015) proved that a two-layer rectifier network can make any disjoint data linearly separable.

While distributed representation is commonly present in many deep networks, convolutional and pooling layers, which are exclusive in convnet, are known to provide shift-invariance (Bruna et al., 2013) and local context preservation. In fact convolutional layers are crucial to obtain better representation than other deep networks such as stacked auto-encoders, as testified by comparing reported results in Le et al. (2012) and Krizhevsky et al. (2012). Importantly, convnets go beyond the I.I.D (independent identical distribution) assumption where their inner representation is highly transferable to related tasks. For example the convnet model trained for image classification (Krizhevsky et al., 2012) can be used as a feature detector (Razavian et al., 2014b) for object

* Corresponding author.
E-mail address: phong.vodinh@gmail.com (P.D. Vo).

http://dx.doi.org/10.1016/j.cviu.2017.01.009
1077-3142/© 2017 Elsevier Inc. All rights reserved.

detection (Erhan et al., 2014), image segmentation (Long et al., 2015), and image retrieval (Babenko et al., 2014; Wang et al., 2014).

Representation learning has been continually pursued by unsupervised methods such as auto-encoders (Hinton and Salakhutdinov, 2006), deep belief nets (Hinton et al., 2006), and sparse encoding (Ranzato et al., 2006). However it is desirable that representation learning combines the advantages of both supervised and unsupervised schemes, in particular to add a strong prior on data. Since label information of training data is unavailable in an unsupervised setting, the objective function of an unsupervised network uses reconstruction loss (Hinton and Salakhutdinov, 2006). This loss focuses too much on redundant image details, trying to reconstruct as well as possible input images at pixel level. Hence, focusing on reconstruction, it ignores the data feature that are interesting in a discriminative context, in particular in order to obtain an acceptable generalisation during the testing step. on the contrary, supervised learning has access to the labels of training data, and thus is better guided to this goal. By minimizing the classification loss, supervised training prunes unnecessary details (linked to reconstruction purpose) and promotes discriminative features.

Training Convnets for representation learning has already been investigated (Dosovitskiy et al., 2014; Rasmus et al., 2015; Valpola, 2015). In particular, the pioneer work of Dosovitskiy et al. (2014) consists in training general feature detectors using supervised convnets combined with artificially generated training data labelling information. The learned representation is therefore quite robust and outperforms other unsupervised representation learning methods. However, the method proposed in Dosovitskiy et al. (2014) is limited to small images of dimension 32 × 32. The seminal work of Hinton et al. (2006) also combines unsupervised and supervised learning. However, the perspective is totally different since the unsupervised part initializes weights, then a supervised step performs fine-tuning. This work had a great influence in the third wave of artificial neural nets research that started in the mid 2000's, leading to the nowadays "deep learning based" successful approaches.

In the same vein as Dosovitskiy et al. (2014), we explore the approach of training large-scale convnets under supervision using weakly labelled data. Prior to deep learning, there have been some works that uses images harvested from Internet and photo sharing sites such as Flickr to train scalable image classifiers (Ulges et al., 2011; Wang et al., 2012). Working with Web images comes with both pros and cons: images are abundant and cheap to collect but very noisy with regard to their annotations. Lately, it has been shown that convnets are surprisingly robust to noise. In Sukhbaatar et al. (2014) and soon followed by Xiao et al. (2015) several solutions to train deep convolutional networks as classifiers under noisy condition are studied. In their works, training data are assumed to contain mislabelled images so that probabilistic frameworks are proposed to estimate conditional mislabelling probabilities. Finally those probabilities are integrated into extra label noise layers placed at the top of convnet in order to improve posterior predictions. Different from Sukhbaatar et al. (2014) and Xiao et al. (2015) we are rather interested in building a robust representation for general purposes from noisy data. Our experiments shown that even without any of special treatment of noisy images, convnet already performs very well. We aim to improve further this performance, not just limited in few specific cases but across a variety of domains.

In Divvala et al. (2014) the term "webly supervised learning" is introduced to qualify the method consisting in collecting large collection of image on the Web to further learn visual concepts (however, their work dealt with training Deformable Part Models and not convnets). The approach of Chen and Gupta (2015) is inspired by curriculum learning (Bengio et al., 2009a) applied to webly supervised learning of CNN. The idea is to first to learn with easy images then gradually adapt the model with harder ones. They consider that images coming from Google image are easy because biased toward a simpler and cleaner presentation, that is supported by previous experiments (Mezuman and Weiss, 2012). FlickR images are, on their side, considered as more realistic. Hence, after an initial learning with Google images, they compute how good is the resulting representation to classify FlickR ones. This analysis gives insights on the confusion between classes. These last are further injected into a weighted soft-max loss function to perform fine-tuning on the FlickR images. While interesting, this approach is different from ours, since we rely on the intrinsic coherency of classes, independently of their performances on a preliminary classification task. Moreover, our paper includes an in-depth study of the impact of the noisy input data on the quality of performances of the Convnet, both in an end-to-end scheme and a transfer learning context.

An other recent work dealt with webly supervised learning of CNN, but explicitly modelled the level of noise in the training dataset (Kakar and Chia, 2015). Given such a knowledge, it is possible to learn a two-layer feedforward neural network to maximize the log likelihood of the observed data. However, contrary to our approach and other works, Kakar and Chia (2015) requires an explicit knowledge of the "label noise" level, both for false positive and false negative. As in their work, it can be estimated manually for a couple of classes, but it becomes intractable when the number of classes becomes large.

In parallel to these works, Zhang et al. (2015) proposed to apply DeepWalk (Perozzi et al., 2014) to a collection of captioned photos to obtain an embedding that model the image-word cooccurrences. The hypothesis is that this collection is a way to model the "collective intelligence" and thus to reduce the noise. The visual features obtained with a Convnet are then mapped to this image-tags embedding with a simple l_1-normed regression. Strictly speaking, the embedding process rather reflects the "collective belief": if a annotation error is commonly present in the collection (*e.g.* associating *whale* images to *fish* tags rather than *mammal*) it will be incorporated to the embedding. However, it this error is also present in the testing dataset, it may be beneficial to the system.

Recently Joulin et al. (2016) conducted a significant experimental work, showing that Convnets trained on noisy data can obtained remarkable results on several tasks, including transfer learning. Such experimental work is also reported in our work in Section 6 with AlexNet (Jia et al., 2014) and GoogleNet (Szegedy et al., 2015) as in Joulin et al. (2016), but we also tested with VGG (Simonyan and Zisserman, 2014). Moreover, an additive contribution our paper is to propose several methods to clean the training datasets and further improve the performances.

Our contribution is twofold. First, we train convnets using noisy and unannotated Web images retrieved from the image search engine Bing and the photo sharing network Flickr. Experiments are scaled from a small image collection of hundred concepts and 400K images to a larger collection with a thousand concepts with 3.14 million images. In both scales the learned representations provide very generalized features that lead to promising accuracies on many classification datasets. Second, we convey image reranking techniques to remove noises from training data and train convnets of deeper architectures. Results show that the proposed techniques help improving classification results significantly. The best configuration outperforms CaffeNet and closes the gap with Vgg-16 (Simonyan and Zisserman, 2014).

In the remainder of this paper, we present data collection procedures in Section 2 and methods in Section 3. Section 5 presents experiment results and Section 9 concludes the paper.

Please cite this article as: P.D. Vo et al., Harnessing noisy Web images for deep representation, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.01.009