Image and Vision Computing 75 (2018) 21–31

Contents lists available at ScienceDirect

Image and Vision Computing

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

Beyond one-hot encoding: Lower dimensional target embedding ☆

Pau Rodríguez [a,*], Miguel A. Bautista [b], Jordi Gonzàlez [a], Sergio Escalera [a,c]

[a] Computer Vision Center, Universitat Autònoma de Barcelona, Spain
[b] Heidelberg Collaboratory for Image Processing, Heidelberg University, Germany
[c] University of Barcelona, Barcelona, Spain

ARTICLE INFO

Article history:
Received 23 May 2017
Accepted 27 April 2018
Available online 11 May 2018

Keywords:
Error correcting output codes
Output embeddings
Deep learning
Computer vision

ABSTRACT

Target encoding plays a central role when learning Convolutional Neural Networks. In this realm, one-hot encoding is the most prevalent strategy due to its simplicity. However, this so widespread encoding schema assumes a flat label space, thus ignoring rich relationships existing among labels that can be exploited during training. In large-scale datasets, data does not span the full label space, but instead lies in a low-dimensional output manifold. Following this observation, we embed the targets into a low-dimensional space, drastically improving convergence speed while preserving accuracy. Our contribution is two fold: *(i)* We show that random projections of the label space are a valid tool to find such lower dimensional embeddings, boosting dramatically convergence rates at zero computational cost; and *(ii)* we propose a normalized eigenrepresentation of the class manifold that encodes the targets with minimal information loss, improving the accuracy of random projections encoding while enjoying the same convergence rates. Experiments on CIFAR-100, CUB200-2011, Imagenet, and MIT Places demonstrate that the proposed approach drastically improves convergence speed while reaching very competitive accuracy rates.

1. Introduction

Convolutional Neural Networks lie at the core of the latest breakthroughs in large-scale image recognition [1,2], at present even surpassing human performance [3], applied to the classification of objects [4], faces [5], or scenes [6]. Due to its effectiveness and simplicity, one-hot encoding is still the most prevalent procedure for addressing such multi-class classification tasks: in essence, a function $f : \mathbb{R}^p \rightarrow \mathbb{Z}_2^n$ is modeled, that maps image samples to a probability distribution over a discrete set of the n labels of target categories.

Unfortunately, when the output space grows, class labels do not properly span the full label space, mainly due to existing label cross-correlations. Consequently, one-hot encoding might result inadequate for fine-grained classification tasks, since the projection of the outputs into a higher dimensional (orthogonal) space dramatically increases the parameter space of computed models. In addition, for datasets with a large number of labels, the ratio of samples per label is typically reduced. This constitutes an additional challenge for training CNN models in large output spaces, and the reason of slow convergence rates [7].

In order to address the aforementioned limitations, output embeddings have been proposed as an alternative to the one-hot encoding for training in large output spaces [8]: depending on the specific classification task at hand, using different output embeddings captures different aspects of the structure of the output space. Indeed, since embeddings use weight sharing during training for finding simpler (and more natural) partitions of classes, the latent relationships between categories are included in the modeling process.

According to Akata et al. [9], output embeddings can be categorized as:

- Data-independent embeddings, such as drawing rows or columns from a Hadamard matrix [10]: data-independent embeddings produce strong baselines [11], since embedded classes are equidistant due to the lack of prior knowledge;
- Embeddings based on a priori information, like attributes [12], or hierarchies [13]: unfortunately, learning from attributes requires expert knowledge or extra labeling effort and hierarchies require a prior understanding of a taxonomy of classes,

☆ This paper has been recommended for acceptance by Robert Walecki.

* Corresponding author at: Computer Vision Center, Universitat Autònoma de Barcelona, Campus UAB, Edifici O, Cerdanyola del Vallès s/n 08193, Barcelona.
E-mail addresses: pau.rodriguez@cvc.uab.cat, pau.rodriguez@uab.cat (P. Rodríguez), miguel.bautista@iwr.uni-heidelberg.de (M.A. Bautista), jordi.gonzalez@cvc.uab.cat (J. Gonzàlez), sescalera@cvc.uab.cat (S. Escalera).

https://doi.org/10.1016/j.imavis.2018.04.004
0262-8856/

and in addition, approaches that use textual data as prior do not guarantee visual similarity [11]; and
- Learned embeddings, for capturing the semantic structure of word sequences (i.e. annotations) and images jointly [14]. The main drawbacks of learning output embeddings are the need of a high amount of data, and a slow training performance.

Thus, in cases where there exist high quality attributes, methods with prior information are preferred, while in cases of a known equidistant label space, data-independent embeddings are a more suitable alternative. Unfortunately, the architectural design of a model is bound to the particular choice among the above-mentioned embeddings. Thus, once a model is chosen and trained using a specific output embedding, it is hard to reuse it for another tasks requiring a different type of embedding.

In this paper, Error-Correcting Output Codes (ECOCs) are proven to be a better alternative to one-hot encoding for image recognition, since ECOCs are a generalization of the three embedding categories [15], so a change in the ECOC matrix will not constitute a change in the chosen architecture. In addition, ECOCs naturally enable error-correction, low dimensional embedding spaces [16], and bias and variance error reduction [17].

Inspired by the latest advances on ECOCs, we circumvent one-hot encoding by integrating the Error-Correcting Output Codes into CNNs, as a generalization of output embedding. As a result, a best-of-both-worlds approach is indeed proposed: compact outputs, data-based hierarchies, and error correction. Using our approach, training models in low-dimensional spaces drastically improves convergence speed in comparison to one-hot encoding. Fig. 1 shows an overview of the proposed model.

The rest of the paper is organized as follows: Section 2 reviews the existing work most closely related to this paper. Section 3 presents the contribution of the proposed embedding technique, which is two fold: *(i)* we show that random projections of the label space are suitable for finding useful lower dimensional embeddings, while boosting dramatically convergence rates at zero computational cost; and *(ii)* In order to generate partitions of the label space that are more discriminative than the random encoding (which generates random partitions of the label space), we also propose a normalized eigenrepresentation of the class manifold to encode the targets with minimal information loss, thus improving the accuracy of random projections encoding while enjoying the same convergence rates. Subsequently, the experimental results on CIFAR-100 [18], CUB200-2011 [19], MIT Places [6], and ImageNet [1] presented in Section 4 show that our approach drastically improves convergence speed while maintaining a competitive accuracy. Lastly, Section 5 concludes the paper discussing how, when gradient sparsity on the output neurons is highly reduced, more robust gradient estimates and better representations can be found.

2. Related work

This section reviews those works on output embeddings most related to ours, in particular those using ECOC.

2.1. Output embeddings

Most of the related literature addresses the challenge of zero-shot learning, i.e. training a classifier in the absence of labels. Often, the proposed approaches take into account the attributes of objects [9,20-22] related to the different classes through well-known, shared object features.

Due to their computing efficiency based on a divide-and-conquer strategy, output embeddings have been also proven useful for those multi-class classification problems in which testing all possible class labels and hierarchical structures is not feasible [23,19,14,8]. Given a large output space, most labels are usually considered instances of a superior category e.g., sunflower and violet are flower plants. In this sense, the inherent hierarchical structure of the data makes divide-and-conquer hierarchical output spaces a suitable alternative to the traditionally flat 1-of-N classifiers. Likewise in the context of language processing, Mikolov et al. combine Huffman binary codes and hierarchical soft-max in order to map the most frequent codes to shorter paths in a tree [24].

Because output embeddings enforce weight sharing, they have been also used when the number of classes is rather large, with no clear inter-class boundaries, and a decaying ratio of the number of

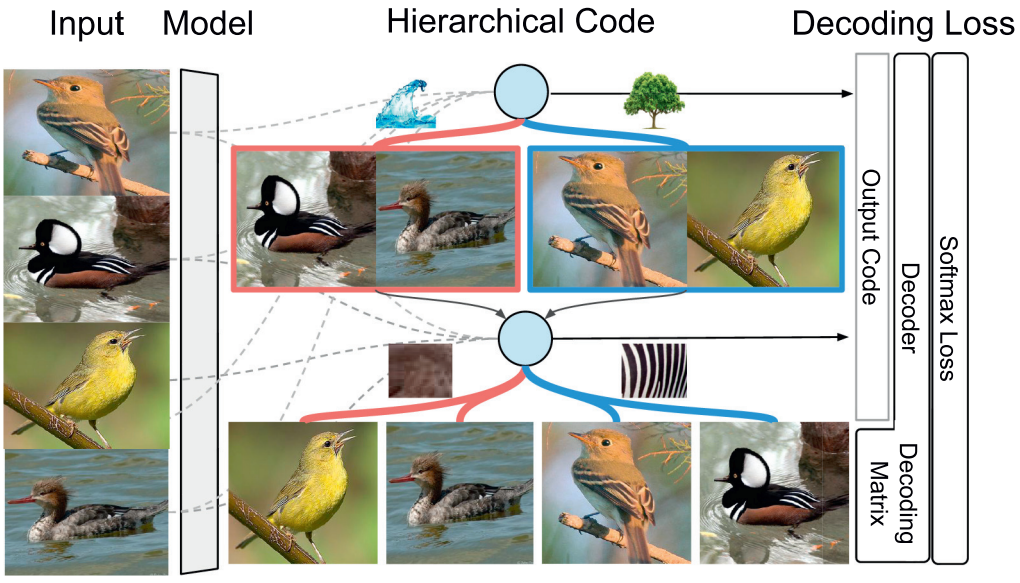

Fig. 1. This paper proposes to replace the traditional one-hot output scheme of CNNs with a reduced scheme with at least $\log_2(k)$ outputs. In addition, when using a hierarchical representation of the data labels, outputs show that the most discriminative attributes to split the target classes have been learned. In essence, a decoder computes the similarities of the "predicted code" in a "code-matrix", and subsequently the output label is then obtained through a softmax layer. The internal code representation is depicted in a tree structure, where each bit of the code corresponds to the actual learned partition from the data, from lower partition cost (aquatic) to higher (stripped).