LaTeX Learning Robust Visual-Semantic Embeddings

First Author
Yao-Hung Hubert Tsai
School of Computer Science
yaohungt@cs.cmu.edu

Second Author
Liang-Kang Huang
Machine Learning Department
liangkah@andrew.cmu.edu

Abstract

Many of the existing methods for learning joint embedding of images and text use only supervised information from paired images and its textual attributes. Taking advantage of the recent success of unsupervised learning in deep neural networks, we propose an end-to-end learning framework that is able to extract more robust multi-modal representations across domains. The proposed method combines representation learning models (i.e., auto-encoders) together with cross-domain learning criteria (i.e., Maximum Mean Discrepancy loss) to learn joint embeddings for semantic and visual features. A novel technique of unsupervised-data adaptation inference is introduced to construct more comprehensive embeddings for both labeled and unlabeled data. We evaluate our method on Animals with Attributes and Caltech-UCSD Birds 200-2011 dataset with a wide range of applications, including zero and few-shot image recognition and retrieval, from inductive to transductive settings. Empirically, we show that our framework improves over the current state of the art on many of the considered tasks.

1. Introduction

Over the past few years, due to the availability of large amount of data and the advancement of the training techniques, learning effective and robust representations directly from images or text becomes feasible [19, 27, 32]. These learned representations have facilitated a number of high-level tasks, such as image recognition [40], sentence generation [17], and object detection [35]. Despite useful representations being developed for specific domains, learning more comprehensive representations across different data modalities remains challenging. In practice, more complex tasks, such as image captioning [45] and image tagging [23] often involve data from different modalities (i.e., images and text). Additionally, the learning process would be faster, requiring fewer labeled examples, and hence more scalable to handling a large number of categories if we could transfer cross-domain knowledge more effectively [9]. This motivates learning multi-modal embeddings. In this paper, we consider learning robust joint embeddings across visual and textual modalities in an endto-end fashion under zero and few-shot setting. Zero-shot learning aims at performing specific tasks, such as recognition and retrieval of novel classes, when no label information is available during training [16]. On the other hand, few-shot learning enables us to have few labeled examples in our of-interest categories [38]. In order to compensate the missing information under the zero and fewshot setting, the model should learn to associate novel concepts in image examples with textual attributes and transfer knowledge from training to test classes. A common strategy for deriving the visual-semantic embeddings is to make use of images and textual attributes in a supervised way [41, 2, 48, 49, 50, 22, 7]. Specifically, one can learn transformations of images and textual attributes under the objective that the transformed visual and semantic vectors of the same class should be similar in the joint embeddings space. Despite good performance, this common strategy basically boils down to a supervised learning setting, learning from labeled or paired data only. In this paper, we show that to learn better joint embeddings across different data modalities, it is beneficial to combine supervised and unsupervised learning from both labeled and unlabeled data.Our contributions in this work are as follows. First, to extract meaningful feature representations from both labeled and unlabeled data, one possible option is to train an 13571 auto-encoder [33, 5]. In this way, instead of learning representations directly to align the visual and textual inputs, we choose to learn representations in an auto-encoder using reconstruction objective. Second, we impose a crossmodality distribution matching constraint to require the embeddings learned by the visual and textual auto-decoders to have similar distributions. By minimizing the distributional mismatch between visual and textual domain, we show improved performance on recognition and retrieval tasks. Finally, to achieve better adaptation on the unlabeled data, we perform a novel unsupervised-data adaptation inference technique.

We show that by adopting this technique, the accuracy increases significantly not only for our method but also for many of the existing other models. Fig. 1 illustrates our overall end-to-end differentiable model. To summarize, our proposed method successfully combines supervised and unsupervised learning objectives, and learns from both labeled and unlabeled data to construct joint embeddings of visual and textual data. We demonstrate improved performance on Animals with Attributes (AwA) [21] and Caltech-UCSD Birds 200-2011 [47] datasets on both image recognition and image retrieval tasks under zero and few-shot setting.

1.1. Related Work

In this section, we provide an overview of learning multimodal embeddings across visual and textual domain. Zero and Few-Shot Learning Zero-shot [7, 1, 2] and few-shot learning [8, 39, 20] are related problems, but somewhat different in the setting of the training data. While few-shot learning aims to learn specific classes through one or few examples, zero-shot learning aims to learn even when no examples of the classes are presented. In this setting, zero-shot learning should rely on the side information provided by other domains. In the case of image recognition, this often comes in the form of textual descriptions. Thus, the focus of zero-shot image recognition is to derive joint embeddings of visual and textual data, so that the missing information of specific classes could be transferred from the textual domain. Since the relation between raw pixels and text descriptions is non-trivial, most of the previous work relied on learning the embeddings through a large amount of data. Witnessing the success of deep learning in extracting useful representations, much of the existing work mostly applies deep neural networks to first transform raw pixels and text into more informative representations, followed by using various techniques to further identify the relation between them. For example, Socher et al. [41] used deep architectures [13] to learn representations for both images and text, and then used a Bayesian framework to perform classification. Norouzi et al. [29] introduced a simple idea that treated classification scores output by the deep network [19] as weights in convex combination of word vectors. Fu et al. [10] proposed a method that learns projections from low-level visual and textual features to form a hypergraph in the embedding space and performed label propagation for recognition. A number of similar methods learn transformations from input image representations to the semantic space for the recognition or retrieval purposes [2, 49, 1, 48, 7, 50, 51, 30, 37, 6, 12]. A number of recent approaches also attempt to learn the entire task with deep models in an end-to-end fashion. Frome et al. [9] constructed a deep model that took visual embeddings extracted by CNN [19] and word embeddings as input, and trained the model with the objective that the visual and word embeddings of the same class should be well aligned under linear transformations. Ba et al. [22] predicted the output weights of both the convolutional and fully connected layers in a deep convolutional neural network. Instead of using textual attributes or word embeddings model, Reed et al. [34] proposed to train a neural language model directly from raw text with the goal of encoding only the relevant visual concepts for various categories. Visual and Semantic Knowledge Transfer Liu et al. [24] developed multi-task deep visualsemantic embeddings model for selecting video thumbnails based on side semantic information (i.e., title, description, and query). By incorporating knowledge about objects similarities between visual and semantic domains, Tang et al. [44] improved object detection in a semisupervised fashion. Kottur et al. [18] proposed to learn visually grounded word embeddings (visw2v) and showed improvements over text only word embeddings (word2vec) on various challenging tasks. Reed et al. designed a text-conditional convolutional GAN architecture to synthesize an image from text. Recently, Wang et al. [46] introduced structure-preserving constraints in learning joint embeddings of images and text for image-to-sentence and sentence-to-image retrieval tasks. Unsupervised Multimodal Representations Learning One of our key contributions is to effectively combine supervised and unsupervised learning tasks for learning multi-modal embeddings. This is inspired and supported by several previous works that provided evidence of how unsupervised learning tasks could benefit cross-modal feature learning. Ngiam et al. [28] proposed various models based on Restricted Boltzmann Machine, Deep Belief Network, and Deep Auto-encoder to perform feature learning over multiple modalities. The derived multi-modal features demonstrated an improved performance over single-modal features on the audio-visual speech classification tasks. Srivastava and Salakhutdinov [42] developed a Multimodal Deep 3572 Boltzmann Machine for fusing together multiple diverse modalities even when some of them are absent. Providing inputs of images and text, their generative model manifested noticeable performance improvement on classification and retrieval tasks.

1.2. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

1.3. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When ref-

[1]This is what a footnote looks like. It often distracts the reader from the main flow of the argument.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 1. Results. Ours is better.

erenced in the text, enclose the citation number in square brackets, for example [1]. Where appropriate, include the name(s) of editors of referenced books.

2. Final copy

You must include your signed IEEE copyright release form when you submit your finished paper. We MUST have this form before your paper can be published in the proceedings.

Please direct any questions to the production editor in charge of these proceedings at the IEEE Computer Society Press: Phone (714) 821-8380, or Fax (714) 761-1784.

References

[1] Authors. The frobnicatable foo filter, 2014. Face and Gesture submission ID 324. Supplied as additional material `fg324.pdf`.