Visual Classifier Prediction by Distributional Semantic Embedding of Text Descriptions

Mohamed Elhoseiny and Ahmed Elgammal
Department of Computer Science, Rutgers University
m.elhoseiny@cs.rutgers.edu, elgammal@cs.rutgers.edu

Extended Abstract

One of the main challenges for scaling up object recognition systems is the lack of annotated images for real-world categories. It is estimated that humans can recognize and discriminate among about 30,000 categories (Biederman and others, 1987). Typically there are few images available for training classifiers form most of these categories. This is reflected in the number of images per category available for training in most object categorization datasets, which, as pointed out in (Salakhutdinov et al., 2011), shows a Zipf distribution.

The problem of lack of training images becomes even more sever when we target recognition problems within a general category, i.e., subordinate categorization, for example building classifiers for different bird species or flower types (estimated over 10000 living bird species, similar for flowers).

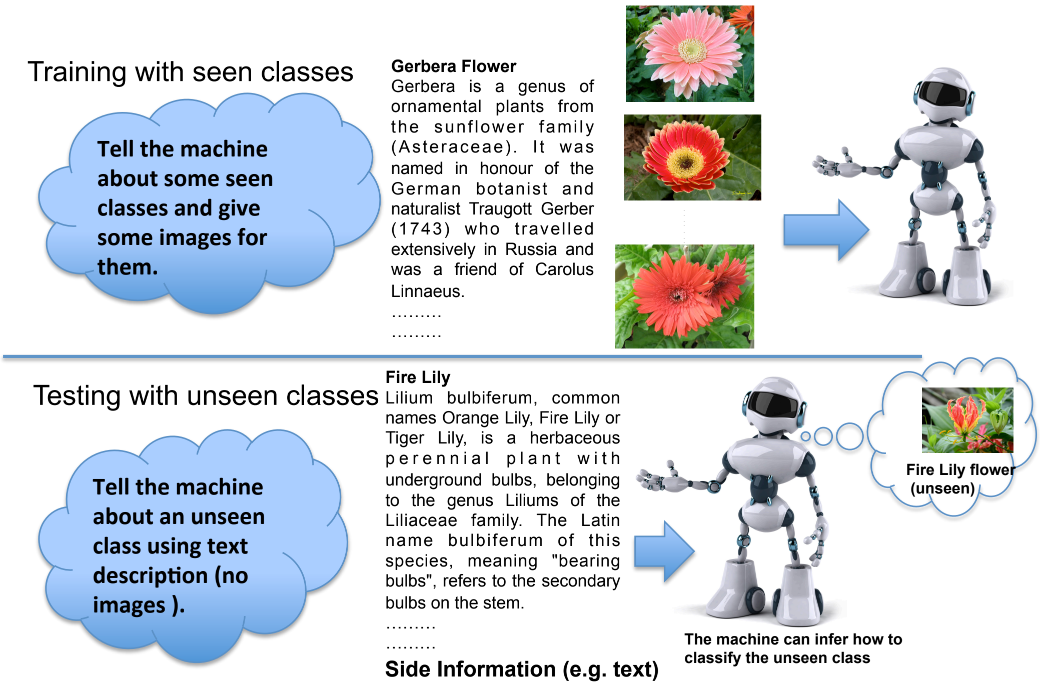
In contrast to the lack of reasonable size training sets for large number of real world categories, there are abundant of textual descriptions of these categories. This comes in the form of dictionary entries, encyclopedia entries, and various online resources. For example, it is possible to find several good descriptions of "Bobolink" in encyclopedias of birds, while there are only few images available for it online.

The main question we address in this paper is how to use purely textual description of categories with no training images to learn a visual classifiers for these categories. In other words, we aim at zero-shot learning of object categories where the description of unseen categories comes in the form of typical text such as an encyclopedia entry; see Fig 1. This is a domain adaptation problem between heterogeneous domain (textual and visual). We explicitly address the question of how to automatically decide which information to transfer between classes without the need of any human intervention. In contrast to most related work, we go beyond simple use of tags and image captions, and apply standard Natural Language Processing techniques to typical text to learn visual classifiers.

Similar to the setting of zero-shot learning, we use classes with training data ("seen classes") to predict classifiers for classes with no training data ("unseen classes"). Recent works on zero-shot learning of object categories focused on leveraging knowledge about common attributes and shared parts (Lampert et al., 2009; Farhadi et al., 2009). Typically, attributes are manually defined by humans and are used to transfer knowledge between seen and unseen classes. In contrast, in our work, we do not use any explicit attributes. The description of a new category is purely textual, and the process is totally automatic without human annotation beyond the category labels.

In general, knowledge transfer aims at enhancing recognition by exploiting shared knowledge between classes. This can come in different ways. Sharing knowledge can by achieved by enforcing a hierarchical structure on the classes, general to specific. Such hierarchy is used to impose constraints on the classifier parameters. Such hierarchies can be exported from text domain, e.g., WordNet, or learned from visual features. Our work can be seen in this context, where, we use learned visual classifiers and textual information to learn across-domain

Proceedings of the 2015 Workshop on Vision and Language (VL'15), pages 48–50,
Lisbon, Portugal, 18 September 2015. ©2015 Association for Computational Linguistics.


Training with seen classes
Tell the machine about some seen classes and give some images for them.
Testing with unseen classes
Tell the machine about an unseen class using text description (no images).
Fire Lily flower (unseen)
The machine can infer how to classify the unseen class
Side Information (e.g. text)

Gerbera Flower
Gerbera is a genus of ornamental plants from the sunflower family (Asteraceae). It was named in honour of the German botanist and naturalist Traugott Gerber (1743) who travelled extensively in Russia and was a friend of Carolus Linnaeus.
....
....

Fire Lily
Lilium bulbiferum, common names Orange Lily, Fire Lily or Tiger Lily, is a herbaceous perennial plant with underground bulbs, belonging to the genus Liliums of the Liliaceae family. The Latin name bulbiferum of this species, meaning "bearing bulbs", refers to the secondary bulbs on the stem.
....
....

Figure 1: Zero Shot Learning from Side Information (e.g., text)

correlation that facilitates the prediction of visual classifiers for unseen classes.

Scope of the presentation

In this talk, we will present an on-going research on the task of learning visual classifiers from purely textual description with zero or very few visual examples. In an ICCV13 (Elhoseiny et al., 2013), we investigated this new problem, we proposed two baseline formulations based on regression and domain adaptation. Then, we proposed a new constrained optimization formulation that combines a regression function and a knowledge transfer function with additional constraints to solve the problem. In this talk/presentation, we will present our new zero-shot learning framework for predicting kernelized classifiers in the visual domain for categories with no training images where the knowledge comes from textual description about these categories. Through our new optimization framework, the proposed approach is capable of embedding the class-level knowledge from the text domain as kernel classifiers in the visual domain. We also proposed a distributional semantic kernel between text descriptions which is shown to be effective in our setting. The proposed framework is not restricted to textual descriptions, and can also be applied to other forms knowledge representations. Our approach was applied for the challenging task of zero-shot learning of fine-grained categories from text descriptions of these categories. The results surpasses the results in (Elhoseiny et al., 2013) under the same setting, and also other baselines including (Norouzi et al., 2014). We also show the value of our proposed distributional semantic kernel under this setting. We also show that our framework is applicable to other form of side information including weak attributes in addition to text.

References

Irving Biederman et al. 1987. Recognition-by-components: A theory of human image understanding. *Psychological review*.

Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. 2013. Write a classifier: Zero shot learning using purely textual descriptions. In *ICCV*.

Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. 2009. Describing objects by their attributes. In *CVPR*.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by betweenclass attribute transfer. In *CVPR*.

Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. 2014. Zero-shot learning by convex combination of semantic embeddings. *ICLR*.

Ruslan Salakhutdinov, Antonio Torralba, and Joshua B. Tenenbaum. 2011. Learning to share visual appearance for multiclass object detection. In *CVPR*.