Multi-Cue Zero-Shot Learning

Lina Zhu

June 4 ,2018

Abstract

The course of extending visual category recognition to large numbers is still challenging. A promising research direction is zero-shot learning, which does not require any training data to identify new classes, but instead relies on some form of ancillary information describing new classes. Our goal is to circumvent this bottleneck by extracting multiples instead of these. Annotation Sources of information from multiple unstructured texts are easily available online. In order to make up for the weaker form in which we have integrated the auxiliary information, we have strengthened the supervision in the form of semantic part annotations in the class we impart knowledge to. We achieve our goal through a joint embedded framework to map multiple text parts and multiple semantic parts into a common space.

1. Introduction

Getting visual concepts in humans and machines is still very different. This is assuming that the concept of early childhood is mainly learning through visual concepts is a straightforward example method based on sensory information and associated in various ways [3]. However, this simply does not explain the visual knowledge of diverse adults. Many of our knowledge is preserved and transmitted through text and current online resources. This eliminates the need for humans to recognize that an object has seen a single instance of this object. Our goal is to make up for the use of weaker performance losses but to make it more widely available - more visual supervision of our curriculum is shifting with helper language information. Following multimodal embedded paradigm zero learning [1], we have constructed a new framework that uses powerful visual supervising formulae in embedding to be flexible enough to accommodate a wide range of textual sources. Our contributions are as follows: (1) We propose to adjust deep fragmentation embedding[2] for language generation to achieve zero-learning to facilitate the integration of multilingual cues and visually embedded information into the joint space. Our framework supports and integrates a wide range of textual and visual resources. (2) We propose a novel language embedding method for unstructured texts and human annotations that do not require any attributes. (3) We use strong supervised semantic part annotations to compensate for weaker but more widely available auxiliary language information. Our latest technology for improving fine-grained zero learning uses unsupervised text sources as supporting letters and supervising attribute annotations. (4) We show that using more powerful visual annotation training allows for the same powerful supervision during the accreditation process without the need to improve zero-shoot performance.

2. experiment

In our experimental evaluation, we used the fine-grained California Institute of Technology UCSD Birds dataset containing approximately 60 images of 200 different North American bird populations each. Each class is also annotated with 312 visual attributes. In the zero setting, 150 classes are used for training and 50 other classes are used for training tests. For parametric verification, we also use the zero shot to set 100 courses for the training within the 150 levels of the training set we use, and the rest for validation. We extract the fully connected layer of the picture feature depth CNN from the activation of the picture. We rescaled the image to 224 x 224 and fed it to the network to pre-train the VGG network as a plurality of visual parts in accordance with the 16-layer model architecture. We used the feature image extracted from the annotation part position of the image. To do this, we crop the overlapped image size to a 50x50 wrapper. We draw this particular part position, adjust the size of each border to 224x224 and follow the rest of the pipe. As a supervised language part, we use a manual annotation for each class of attributes with continuous values to measure the strength of each class's attributes.

2.1. Partial annotation of strong supervision

In addition to using non-linear embedded targets, our joint part is embedded from using multiple visual or linguistic parts. We extracted 19 parts from each image corresponding to the entire image, head, body and the complete bounding box[4], and bounded boxes drawn around 15 parts. We evaluated the impact of the parts in the following manner: (1) Training And use a single part for testing, (2) with multiple training parts and one-piece testing, and (3) training and testing with multiple parts.Zero image classification. For zero image classification, we calculate the average Top-1 accuracy per class in the invisible course. In other words, we consider that only when the predicted category label is a prediction is the correct category label that is matching the image. We average the forecast based on each class. The results are shown in Table 1. For attributes, using multiple visualizations has improved accuracy from 43.3% to 47.0% at training time, improving the latest technology. On the other hand, the use of multiple visual components at the test time achieves 56.5% accuracy, further improving the latest technology of supervision on this data set. For word bags, the use of multiple visual effects parts increased accuracy by 26.0%. The multiple visual parts reached an impressive 32.1% accuracy as the new technology gained without using humans is supervised in language. These results support our intuition that the use of powerful semantically supervised visual parts leads to more discernible image tables and thus helps to classify fine-grained images taken from zero point shots.

Table 1. Multiple visual parts (VP) for classification. VP are extracted from the annotations that are provided with the dataset.

Train VP	Test VP	Attributes	word2vec	BoW
1	1	43.3	25.0	21.8
19	1	47.0	26.8	22.6
19	19	56.5	32.1	26.0

3. Conclusion

We learned experimental conditions that allow the integration of different categories of descriptions and detailed part annotations, and thus significantly improve the state-of-the-art of the two tasks within the scope. In particular, we have shown how to compensate for the loss of precision used with weaker auxiliary information level annotations with detailed visual parts. Our approach helps to jointly embed multilingual parts and visual information in a joint space. With strong visual supervision and user-friendly attention attributes we improved the latest technology CUB datasets to supervised settings in 56.5% (from 50.2%), we combined different unsupervised texts to embed and further improved the results of unsupervised settings to 34.7 %. As a conclusion, we propose several fine-grained extended zero-point learning. First, using multiple visual parts, if available, ie training or testing time, instead of using a visual part leads to a significant increase in performance. Second, it can support the further improvement of these multi-language parts of multiple visual parts. Third, the space does contain some information and distances between potential class and attribute names that can eliminate expensive human annotation associations for class attributes. Following these practices, we have improved the fine-grained zero-shooting state and unsupervised text embeddings under supervision.

References

[1] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classi- fication. In *CVPR*, 2015. 1

[2] A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1

[3] D. K. Roy and A. P. Pentland. Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146, 2002. 1

[4] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Partbased r-cnns for fine-grained category detection. In *ECCV*, 2014. 2