A study on zero-shot learning based on generative model in image classification and object detection

48-186630 Kuanchao Chu

Abstract—**For a deep neural network-based perceptual system, generative zero-shot learning is a solution to detect objects from the unseen classes which are omitted from the training set. It utilizes existing knowledge inside auxiliary semantic representations to compensate for the absence of corresponding visual data. In this work, we disassemble the problem into two parts, a core zero-shot image classification task and the following challenge to combine it with general detection models. Firstly, we propose an enhanced model based on a conditional generative adversarial nets, with an additional loss that regularizes the generator by interpolated class semantics. Such a generative model is used to synthesize unseen object features in a data augmentation fashion. Besides, an ensemble classifier is designed to mitigate the classification performance trade-off between seen and unseen objects. In the second part, we build a framework that is capable of integrating the generative model to a one-stage object detector and achieve detecting the object from novel classes. The models are evaluated on several benchmark datasets, including AWA2, CUB, and SUN for the classification task, and Microsoft COCO for the object detection task.**

I. INTRODUCTION

Intelligent robots and agents are gradually changing the human world. We benefit from their safeness, productivity, and autonomy. In the past decade, the deep neural networks wildly boost their problem-solving capacity, with the superior performance on generalization and feature extraction.

However, such algorithms heavily rely on high-quality training datasets. Under the extreme condition that zero training data is available for some object classes, the network is even unable to learn from these classes. Since the nature of the long-tail distribution of object categories and the considerable manual labor and cost for data collection and annotation, the insufficient data issue is quite common. It hence builds a barrier for many real-world applications. As to mitigate this issue, zero-shot learning is a compelling approach that deals with the problem setting of zero training data obtained for part of the classes.

In zero-shot learning, semantic or textual descriptions are usually used as side information to compensate for the absence of visual data of some classes, which are called the ”unseen.” This benefits from that the cost of acquiring the semantic knowledge is much lower than preparing corresponding unseen object images, and less supervision is required as well. The majority approaches tend to build a robust mapping relationship between the visual and semantic spaces[1].

Recently, some[2] utilize generative models to synthesize visual representations of the unseen classes. It turns out a significant improvement in the classification accuracy of unseen classes, along with the flexibility as capable of generating individual samples explicitly. Therefore, we put the focus on extending and exploring the availability of such models.

In this research, our contributions are as follows. (1) We propose novel strategies including class-semantic interpolation on generator regularization and the ensemble classifier to enhance task performance of the baseline generative model, according to its model behaviors in our case analysis. (2) We design an extensible framework that combines the generative model alongside a regularization on object confidence for image classification with the one-stage YoloV3-like object detector to achieve zero-shot detection. Its efficacy is verified and concluded with design guidelines and current limitations. (3) The applicable real-world scenario and the gaps in between are discussed and analyzed.

II. BACKGROUND

Generalized zero-shot learning: Let $y^{seen} = \{y^{s_1}, ..., y^{s_K}\}$ consisting of K discrete seen object class labels, and $y^{unseen} = \{y^{u_1}, ..., y^{u_L}\}$ is comprised of L discrete unseen object class labels. y^{seen} and y^{unseen} are disjoint. During the training phase, only images of y^{seen} are provided, but the corresponding semantic knowledge is available for both. The goal is to learn $f : x \rightarrow y^{seen} \cup y^{unseen}$ for the test sample x which could belongs to any class of the seen or unseen.

Semantic knowledge base: It plays an enormous role in zero-shot learning and can be collected from a few resources. (1) Supervised knowledge: A series of attributes that can describe the properties of a class, including the common appearance or nature, are labeled by human experts with discrete scores. (2) Unsupervised knowledge: the word representations obtained from a trained language model can be used as the class-wise semantic vectors.

III. GENERATIVE ZERO-SHOT IMAGE CLASSIFICATION

We train a feature generator to generate object features of the unseen classes explicitly. Hence, the training set of image classification can be synthesized by the real image features extracted by ResNet101 and the generated features. The baseline feature generator is a conditional generative adversarial nets with WGAN-GP loss, plus a cross-entropy classification loss which encourages the generator to output feature that can be easily classified on the pre-trained seen objects classifier.

To alleviate the bias toward seen classes, which is caused by the unreal generated features, we introduce

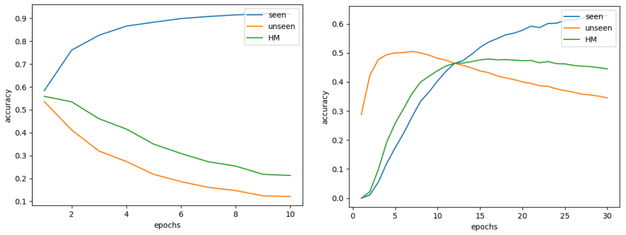

Fig. 1: The early climax behavior on AWA2 (left) and CUB (right). Colored lines represent the classification accuracy of trained epochs.

a regularization loss term reg_{intp} that is targeted to fill in spaces on the manifold structure of features between semantically similar classes. It guides the generated features that conditioned on interpolated semantics could be considered as real ones for the discriminator.

Besides, early climax behavior shown in fig. 1 is studied. As one of the evaluation metrics, top-1 accuracy for unseen classes, one can observe its peak in the early training stage. This causes the full potential on classifying seen and unseen objects cannot be achieved simultaneously, since the major performance metric is the harmonic mean (HM). We modify the original softmax classifier to an ensemble classifier, which has a decision module ahead of two softmax classifiers in parallel. The decision module predicts seen or unseen classes for the incoming image samples, then sends it to one of the classifiers with weights from different epochs.

The experiments are carried out on the benchmark datasets, with 10˜20% of the classes are selected as unseen, and 6˜12 times the averaged features per seen classes are generated for each unseen classes. As a result, a 1˜4% improvement over the baseline model in the HM of image classification is achieved.

IV. The Framework of Generative Zero-Shot Object Detection

In an object detection task, both recognition and localization problems should be dealt with. The proposed system is illustrated as in fig. 2. With the generative model, one can synthesize feature maps that contain unseen objects. We train the classifier with m epochs of real feature maps as a warm-up, then with another n epochs of synthesized feature maps.

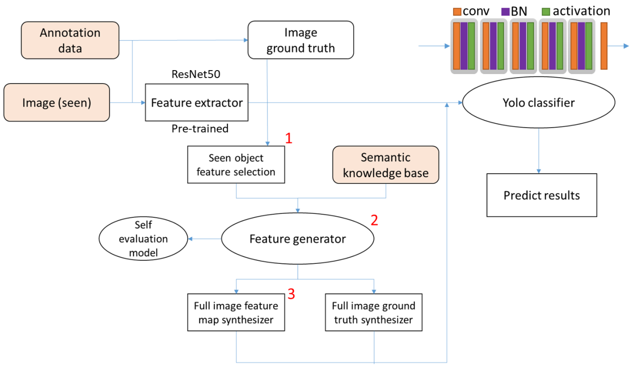

Fig. 2: System overview of the generative zero-shot object detector.

We add a regularization loss term reg_{obj} on the generator, to encourage it outputs features that can receive high object confidence on the classifier pre-trained on real images. Fig. 3 shows the difference: the loss of object confidence (green line) converges rapidly with reg_{obj} applied during the generator training.

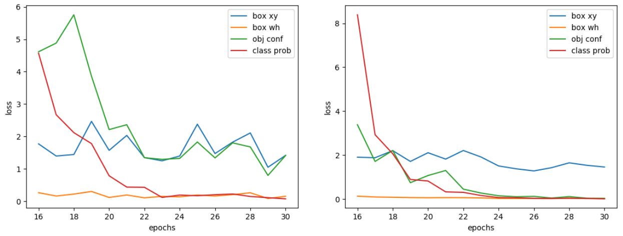

Fig. 3: The loss components while training with the synthesized set. The right image is the result with reg_{obj} applied while the left is not.

We evaluate the detector with the COCO dataset (2014). 20% of the classes are selected as unseen, and the contextual word representations for class labels are retrieved from a pre-trained BERT model. The system is implemented with the feature maps of layer 0 FPN. We follow the standard evaluation protocol on COCO to calculate the $AP_{.50}$ and AP_{unseen} on the minival subset, where $AP_{.50}$ is the score averaged for all the classes, and AP_{unseen} is for the unseen classes only. Comparing the vanilla YoloV3-like detector and our zero-shot detector, $AP_{.50}$ is increased from 12.52 to 15.26, while AP_{unseen} from 0 to 1.65. Fig. 4 shows a pair of results from vanilla and zero-shot detectors.

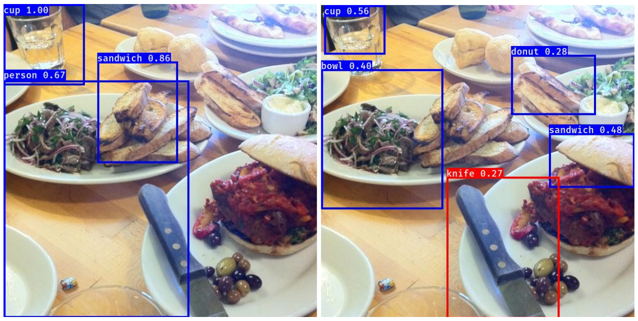

Fig. 4: The samples of detection results. Red box represents object from the unseen classes.

V. Conclusions

The model enhancements that we proposed mitigate the existing shortcomings in the generative model for the zero-shot classification. We further build an extensive framework for zero-shot detection. The results can be further applied to solve insufficient data problems faced by various deep learning tasks.

References

[1] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, “Zero-shot learning – a comprehensive evaluation of the good, the bad and the ugly,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2019.

[2] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, “Feature generating networks for zero-shot learning,” in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.