Critique Report on the Paper: "Context-Aware Zero-Shot Learning for Object Recognition [6]"

Sumaiya Tabassum Nimi

Background Information and Purpose of the Research

In zero-shot learning (ZSL), those objects, on which no data was available during supervised training of the model, are attempted to be classified based on pertinent information available from secondary sources, like the semantic representation of the object. Although in an image, there are usually several objects, constituting the background or context of the target object to be categorized, none of the previous works on ZSL considered accumulating information from this background or context towards classification of the target. In the paper [6], the task of ZSL was formulated as a Bayesian inference that found conditional probability of an object belonging to a class incorporating information about the aforementioned context of the object, visual cues of the object and a precalculated prior component. Basically, this work was an extension of the previously proposed DeViSe [1] model, the novelty being introduction of probabilistic modelling that incorporated visual context and the handling of the class imbalance issues.

Methods Proposed

The probability that an object belongs to a class i was conditioned on the following two components that were hypothesized independent of each other.

- Context Component: In this module, compatibility between the context or background of the object and semantic representation of the classes was computed using a 2-layer neural network. Information about the context was calculated using weighted combinations of the following two types of representations of the source (whose class labels available during training) and target (whose class labels not available during training) background or context objects.
  - Visual representation of the object, calculated by projecting output from pretrained Inception-v3 CNN [5] model
  - Semantic or textual representations of the class labels (only known for source objects), obtained using $Word2Vec$ [4].
- Visual Component: In this module, compatibility between the visual cues of the object (calculated by projecting output from the penultimate layer of pretrained Inception-v3 CNN model) and semantic representation of the classes (obtained using $Word2Vec$ [4]) was computed.

In addition, a prior term was also incorporated in the probabilistic formulation that encompassed information about the distribution of the data known a priori, calculated from semantic representation of the classes using a 2-layer neural network. All these three components were trained using parameterized energy-based learning [3].

Strength of the Work

- Novel approach for ZSL incorporating visual background of image was proposed in this paper.
- The proposed approach has taken the class imbalance of the dataset into consideration and has an associated way around.
- Novel probabilistic formulation of the ZSL problem was proposed.

Limitations of the Work

- It seemed from the experimental results that the only factor in the proposed model that is responsible for better results is the context component. If this component is not included, results obtained were no better than the DeviSe model, which was evident from Table-2 of the paper. In other words, the proposed model heavily relied on the information obtained from the background objects to understand the context. So in cases, when the image does not contain many such objects, the proposed framework will not lead to improved performance. This fact was also noted by the authors in Section-4.1. This is the reason that the proposed approach was only tested on Visual Genome [2] database that contains a large number of annotations for the background objects. The proposed approach was neither quantitatively nor qualitatively validated on datasets or cases where these many background information are not available.
- The performance of the proposed approach was not quantitatively compared with previously proposed works on ZSL, other than DeviSe. So it was not clear, whether this work pushed the boundary of the corresponding research domain. Although the proposed probabilistic framework seemed theoretically sound, there could be a simpler model available that outperforms this approach. A comprehensive comparison in terms of the experimental results should have been reported to confirm that this was not the case.
- The proposed model seemed memory inefficient, since three modules need to be loaded into memory at a time, each of which work on high-dimensional feature maps obtained from pretrained CNN models and also semantic representations of all the class labels.

Questions Unanswered

- In both the context component and the prior component, 2-layer perceptrons were used, but the exact architecture, like whether the network was fully connected or convolutional or recurrent, number of hidden neurons or filters, were not mentioned. These things should have been discussed for reproducibility of the reported results.
- The exact values of the hyperparameters α_C, α_V and α_P, used for obtaining the reported results, were not reported.

- The significance of the red and blue points used in Figure-2 should have been discussed.
- The proposed model heavily relied on correct labelling of the background objects for constructing the context component for proper understanding of the context surrounding the target object. What if the labelling of the these background objects were incorrect? How much robust is the final classification of the target object to these incorrect labels?

Suggested Future Studies

- All the three components used some common representations, like feature map obtained from penultimate layer of pretrained Inception-v3 CNN model. A possible modification of the proposed architecture could explore fusing these three separate modules into one and calculating each of the three components used in the probabilistic formulation out of this one single model. This will lead to reduced load on memory and speedup in inference.
- Instead of using information for all the background objects in the context component, there could be some filtering so that out of all the background information, only the significant ones are used for inference.
- Instead of using separate MLPs in the individual components and then computing the final results using scaling of the obtained probabilities, the Bayesian framework could be redefined using deep probabilistic models like Boltzmann machines.

References

[1] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[2] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[3] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016.

[6] Eloi Zablocki, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. Context-aware zero-shot learning for object recognition. In *International Conference on Machine Learning*, pages 7292–7303. PMLR, 2019.