Critique Report on the Paper: “Don’t Judge an Object by Its Context: Learning to Overcome Contextual Bias [5]”

Sumaiya Tabassum Nimi

Background Information and Purpose of the Research

While visual context often serves as an important cue for certain vision related tasks like object recognition and scene understanding, it so happens that sometimes relying too much on context induces an inadvertent bias towards recognition in presence of the context, especially when trained on datasets where the labels are available only and not the regions in the images encompassing each label. Hence the performance and robustness of the resulting model is adversely affected in such cases. In order to resolve this issue, in the paper [5], training strategies have been proposed so that the effect resulting due to the aforementioned bias arising from the context is nullified, thereby leading to detection of the biased categories both in presence and absence of the context.

Methods Proposed

Two training strategies proposed in this paper for training classifiers that were not biased by the context of the image are discussed below.

- CAM-based Approach: Since in the work, focus was on bias arising from training on datasets where location annotations were not available, Class Activation Maps, CAM [6], was used as an alternative to the category-relevant location information. In the CAM-based training strategy, novel loss function was designed to minimize overlap between regions in the images representing the co-occurring pairs of class labels. While this training approach lead to better detection of objects in the absence of context, performance dropped when objects and contexts co-occurred, since information coming from the context were not utilized.

- Feature splitting based approach: In order to alleviate the issues with the CAM-based approach, the second approach was proposed that learnt a subspace of the features only on training data where object categories were present without the context that gave rise to bias. Training on the rest of the data proceeded as usual, as well as the inference. Training in this way aimed to train a subset of the features to encode exclusive category-specific information without being biased by the context, while still leaving open the opportunity to learn from the context when available. Since in the dataset, the number of instances where objects were present without their typical context was few, the corresponding loss function for this training approach was designed so that higher weights were assigned to those rare samples.

Strength of the Work

- Simple and intuitive novel strategy proposed for object recognition both in presence and absence of typical context that demonstrated better performance in terms of detecting both biased and unbiased categories of objects, compared to the standard training, baseline approaches and techniques proposed in previous works like class-balancing loss [2] and attribute decorrelation [3] on four object and attribute classification datasets containing large number of co-occurring categories.

- Cross-dataset evaluation was done by testing model trained with COCO-stuff dataset [1] on UnRel dataset [4], which verified that the proposed strategies were not dataset dependent and could be deployed in real-life applications for better results.
- Qualitative results reported through visualizations obtained using class activation maps that established the fact that the proposed training strategies, especially the feature splitting strategy successfully captured regions in the images pertaining to the biased object categories and contexts.

Limitations of the Work

- In the feature splitting strategy proposed, no explanation was provided regarding the choice of layer for splitting and the row-based splitting of the features into exact halves. Some ablation studies should have been conducted valid these choices of values and also to demonstrate the extent to which these choices affect the end performance.
- In Table-3, proposed strategies were compared with the standard training only. Results on State-of-the-Art approaches like class-balancing loss should have been reported.
- Qualitative analyses were done to compare the effectiveness of the proposed approaches with the standard training and with each other, but not with the other State-of-the-Art approaches. For all we knew from Table-2, the performance of the class-balancing loss was reasonably good also. So qualitative analyses should have reported at least for that approach.
- Performance of all the discussed strategies were compared in terms of mAP score on 20 most biased categories in all cases and top-3 recall in one case. Despite this being a fair enough comparison, it would have been even better to report results on some other metrics also for further establishment of the claim of superiority of the proposed methods.
- The choice of the values of the hyperparameters $\lambda1$ and $\lambda2$ were done arbitrarily. Some ablation studies, reported even in the form of a simple graph, could have validated the choice of the values.

Questions Unanswered

- It was not clear by reading the paper why half of the features from one particular layer only was exclusively trained for recognizing the biased object categories in absence of their typical context. It should have been explained intuitively, like it was done for the CAM-based approach, the reason behind not choosing the entire feature space for split instead.
- It was not clear if the split of the weight matrix was just random or there were some heuristics behind.

Suggested Future Studies

- The proposed training approach learned to detect objects away from their contexts in an entirely supervised setting. Had no out-of-context example been provided for a particular scenario during training, the classifier would fail to recognize such cases during inference. Hence the current work could be extended to zero-shot settings to recognize objects in absence of the context, when no such similar training example was there.

- The proposed feature splitting method could be improved with a more intelligent split of features, based on the explainable properties of the features. The classifier models used were pretrained ones which were later fine tuned using the proposed strategies. So initially, features of the pretrained model can be ranked and some selective top-ranked features can be chosen from split.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018.

[2] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[3] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1629–1636, 2014.

[4] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5179–5188, 2017.

[5] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11070–11078, 2020.

[6] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.