Critique Report on the Paper: "Localize, Assemble, and Predicate: Contextual Object Proposal Embedding for Visual Relation Detection [6]"

Sumaiya Tabassum Nimi

Background Information and Purpose of the Research

Visual Relation Detection (VRD) is the problem of detecting objects present in an image in addition to finding the relation or interaction between them. The relations are defined as some <subject-predicate-object>tuples, where subjects and objects are related by the relationship defined by predicate. Most of the researches conducted to solve the VRD problem detects all the objects in an image first and then attempts to find relations between all possible pairs of objects. In the paper [6], the following two problems in this traditional pipeline have been addressed.

- Given C classes of objects and R different relations that can exist between each pair, there could be $O(C^2R)$ possible relations. Learning these many relations out of a limited set of data leads to a natural bias towards the more frequent relations, as the frequencies of relations exhibit a long-tailed distribution.
- Inference becomes slower as all detected objects in an image are considered potential subject-object pairs and all possible relations between each pair is considered, whereas there could be only a few valid relations in the image.

Methods Proposed

To address the aforementioned issues with the traditional pipelines for VRD, a three-stage network architecture was proposed that decomposed the problem into three stages namely

- Stage 1: Localizing objects using Region Proposal Network (RPN) in Faster RCNN model [5], with $ResNet - 50$ [2] model as the backbone architecture and selecting no more than 512 object proposals.
- Stage 2: Assembling subject-object pairs out of the object proposals though a novel Contextual Embedding Scheme called Pair Proposal Network (PPN). In PPN, a graph model was constructed where the vertices were the object proposals obtained in Stage 1 and the edges represented the relations between them. To select the edges that represent valid relations, a Conditional Random Field (CRF) [3] model was used for message passing among all proposal embeddings, so that the proposals having larger compatibility (in terms of inter-vector cosine similarity) were drawn closer. The PPN model proposed was the most novel part of the paper and the major contribution, that ensured the desired reduction in inference time by selecting only the most compatible object pairs for categorization and relation prediction.
- Stage 3: For the proposal embedding pairs selected in stage two, classes of the objects were predicted using Multi-Task Loss of Fast RCNN [1] model and the relations between them were predicted using a congregation of subject, object and union bounding boxes interacting with each other in an hour-glass network architecture. The bias towards most frequent

<subject-predicate-object>tuples was mostly alleviated through the proposed relation prediction module as the module was agnostic to the class information of the objects in the image.

Strength of the Work

- Viewed the VRD problem from an angle different from most other related literature. Instead of categorizing the objects first and then finding relations between all the object pairs, some object proposals were generated first. Then pairs of object proposals most compatible with each other were selected and only these pairs were considered both for object classification and relationship prediction. Although we note that the essence is similar to [7], the difference is the proposed novel PPN model, that was claimed to be light and fast.
- Achieved and reported relationship detection performance superior to the prior works in terms of most of the metrics on VRD dataset [4].
- The entire network architecture is end-to-end trainable.
- The first work to propose parallelizable CRF modelling of proposal embeddings to select the most compatible pairs.

Limitations of the Work

- The authors claimed reduction in inference time but did not demonstrate the comparison of inference time with the previous approaches (especially [7], that proposed a very similar architecture) on any experimental result.
- It was claimed in Page-2 of the paper that the first two stages of the work were "class- and predicate-agnostic". Whereas in Page-4 it was stated that the category compatibility function used in Stage-2 captured the compatibility of labels between overlapping or nearby proposals and this statement contradicted the prior claim.
- It was also claimed that the proposed PPN model is lightweight. However, a quantitative comparison of the size of the proposed model with those proposed in the previous works was not reported to validate the claim.

Questions Unanswered

- One important assumption behind the claim of reduced inference time is that the proposed CRF modelling in PPN is parallelizable. What about embedded hardware environment where parallel processing is not feasible? Will there be sufficient speedup in inference time in those cases?
- The proposed architecture consisted of three separate modules. So the model size seemed to be sufficiently large. Is this architecture deployable in memory-constrained embedded devices?

Suggested Future Studies

Stage-1 and the object classification module in Stage-3 of the proposed architecture in [6] are not lightweight enough to be deployed in memory-constrained devices. These modules can be compressed using model pruning or replaced with their more lightweight counterparts to make the architecture more deployable for embedded systems.

References

[1] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[4] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[6] Ruihai Wu, Kehan Xu, Chenchen Liu, Nan Zhuang, and Yadong Mu. Localize, assemble, and predicate: Contextual object proposal embedding for visual relation detection. In *AAAI*, pages 12297–12304, 2020.

[7] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5678–5686, 2017.