CS2001 - Research Topics in Computer Science

Paper Reviews

7

Ph.D. student in Computer Science

1 Spatial Memory for Context Reasoning in Object Detection [1]

In this paper, authors claim that state-of-the-art object detectors suffer from lack of instance-level context modelling and they propose a memory module that preserves spatial layouts of the object instances in an image to learn distinct spatial settings between real world objects. At a high level, what they propose is a 3-D data structure $M^{W \times H \times D}$ in which the first two dimensions are used to mimic the original width and height of an image to preserve its spatial setting and works complementary with any two-stage object detector. Every cell of this $W \times H$ layout holds a D-length vector which is used to encode visual information about the object instances detected at the corresponding area in the original image. Having this spatial memory module in hand, the proposed model iteratively detects object instances by conditioning on previous detections.

Their main contribution is finding an elegant way to use a fixed-size memory module to encode spatial layout. What makes hard to have a fixed-size spatial memory module is that images greatly vary in size so one need to adjust the spatial size of the memory. Since, the prior cannot be dependent on size, they initialize the memory with a fixed spatial size ($20 \times 20 \times 256$) and adjust it depending on the image size using bilinear interpolation.

The problem they attempt to solve is not only novel but also very interesting since it is grounded on psychological evidences of how humans reason visually. The way they formulate the problem of so-called context-guided object detection seems to be precise and references are accurate. Moreover, the experimental setup in which they test their method against the state-of-the-art is mostly appropriate and the ablative analyses show the effectiveness of their model. Yet, the paper has some weaknesses. First of all, it is not clear why they resize *conv5_3* feature maps to 14×14 and how they come up with this number as there is no therotical/experimental ground for that in the paper. Secondly, since their model iteratively detect object instances, one needs to set an iteration limit which basically limits the number of object instances can be detected. Putting such a constraint may seem okay if one knows about the distribution of the test set but it may fail when the test set comes from a different distribution than the training set. Morever, such iterations are GPU-intensive and cannot be easily optimized. The paper has also some gaps that needs to be filled. For example it is not clear that how model is trained. In equation 3, the authors says the spatial memory is optimized jointly with the backbone detector. Though, in the section 5.4, they states they sequentially train the detector and the memory.

As a minor point, some figures they use to explain how their model works are rather confusing, especially figure 3 and figure 4. Other than that, the paper is very coherent and reads very well overall.

2 Iterative Visual Reasoning Beyond Convolutions [2]

In this paper, authors claims that the state-of-the-art object detectors treat object detection as a perception problem, not a reasoning problem. They conclude that object detection must be modeled as a reasoning problem citing psychological works that prove humans greatly reasons over space and semantics while recognizing an object in the wild. The authors propose an iterative reasoning framework that can be stacked on any two-stage object detector and does consist of two core modules: a local module which employs spatial memory and a global graph-reasoning module. At a high level, local module makes predictions based on previous beliefs of the spatial settings of the objects, and the global graph-reasoning module makes its predictions with the help of primitive spatial knowledge (distance between the regions) and external semantic knowledge between objects. These two modules iteratively roll-out and cross-feed predictions to each other to make their estimates consistent.

The paper has two main contributions: employing spatial memory in a way that it can get parallel updates unlike the first paper where it gets updated iteratively, and using a graph-reasoning module that reasons over primitive spatial knowledge and external semantic knowledge to assign regions to object classes.

The problem they bring to our attention is concrete, well-defined and grounded on psychological evidences that show object detection is a reasoning problem to our brain. Their formulation of the problem as so-called context-guided object detection is correct and related work on this topic seems to be correctly cited. Their reasoning for the methods they use to build up their model is mostly concrete and supported by experimental evidences. They carry an extensive ablative analysis to show how each module of their model improves the accuracy. However, the paper has a major weakness. Even though they formulate the problem as object detection in the beginning, they evaluate their model on region classification task which takes great amount of burden off of the model's shoulders as the model does not have to learn about background class and how to regress bounding boxes. Hence, their model evaluation seems irrelevant to the original problem definition. Moreover, there is very little to no explanation for some design-specific choices and this leaves the reader with questions. For example, the authors do not clearly state why logits before softmax function have been chosen as the high-level feature for a region but not the soft-mapping.

The paper generally reads well but not super coherent. Some parts of the paper mentions low-level details without creating a context around them. As a minor flaw, there are a couple of misspellings that needs to be corrected.

3 Structure Inference Net: Object Detection Using Scene-Level Context and Instance-Level Relationships [3]

In this paper, authors claim that object detection should be considered as both recognition and reasoning problem. With this definition of object detection, they propose an inference module which can be plugged into any two-stage object detector and make use of both visual features (recognition) and an object relationship graph (reasoning) to detect object instances in a given image. In this regard, they formulate object detection as a graph structure inference problem where visual features of the proposed regions are nodes and the encoded relationships between regions are edges. Their model learns region relationships based on their visual appearence and geometric settings and pass messages between regions based on the strength of their relations. Having the nodes got messages from all related nodes, the model updates nodes features and use those new region representation for classification and bounding-box regression.

We can divide the paper's contributions into three pieces. First of all, unlike the previous work, their structure inference module does not rely on classification results from the backbone detector hence it can be trained jointly with the backbone detector in an end-to-end manner. Second of all, they extract a scene representation from the feature maps of the last convolutional layer to serve as global context, and allow it to pass message to regions. In that way, a node's new representation is determined based on the messages from the other nodes (object-object relationships) and also based on the message from the scene (scene-level context). Lastly, they employ two gated recurrent units to decide which parts of those messages are meaningful.

The problem they try to solve is well-defined and interesting. If we, humans, reason on object-object relationships over a complete scene while recognizing object instances, why machines could not behave the same way instead of focusing only on visual featues from a region in isolation? Their formulation of the problem as context-guided object detection is well-grounded and precise. The experimental setup in which they evaluate their method against the state-of-the-art seems mostly logical and appropriate for the task. They also perform an extensive ablative analyses to show each component of their module increase the accuracy. Nonetheless, the paper comes with some problems. First of all, the authors do not give a concrete explanation of why they selects specific non-linear activation functions to weight both geometric relationship and visual relationship between region pairs. Secondly, their model varies between the paper and the actual implementation. The actual implementation has some additional parts that never mentioned in the paper. What is more, when they compare their model with the baseline, they used the accuracy reported as in the baseline paper. However, having trained the baseline with the parameter settings, it achives more accuracy.

The paper is very coherent from start to end and reads really well. The graphics they use to visually explain how their model works are designed really well. Apart from the previously mentioned issues and minor grammatical mistakes, I believe it is a good paper that deserves to be in proceedings of a top-tier conference like CVPR.

4 Reasoning-RCNN: Unifying Adaptive Global Reasoning into Large-scale Object Detection [4]

In this paper, authors try to eliminate the requirement of having an unbiased dataset, in which all categories are distributed almost uniformly, to train an accurate object detector. To achieve that, they formulate object detection task as a reasoning problem on a graph where learned weights for categories by a backbone detector serve as initial nodes to a semantic pool and with the help of commonsense knowledge and scene-level attention, they try to close the gaps for categories which resides on the tail of the long-tailed class distribution. Their model exploits object relationships from a well-known dataset, namely Visual Genome [5]. Having had a semantic pool for categories, image-wise attention by a squeeze-and-excitation [6] block and a semantic knowledge graph in hand, their model learns edges between categories and softmap these new category representations to the detected regions. Finally, these new region representations are concatenated with the original region features and passed to classification and bounding-box regression module.

The paper puts two major contributions on the table. The first one is the semantic pool they create from the learned weights for each category by a backbone detector. With such a semantic pool, they achieve not to lose dataset specific features for classes even though very small amount of instances has seen by the model. Their second contribution is that they make use of squeeze-and-excitation block in a very elegant way such that they learn a latent space to which they can project original image and keep just the necessary information.

The issue they want to bring our attention is not something hypotetical. As number of classes increases in a dataset, it gets harder to train a detector which does not suffer from dataset bias. They cited most of the related work. The mathematical formulation of the problem seems correct and all the components of this formulation are backed by strong arguments. Their use of external knowledge to close semantic gaps between categories is very promising. The experimental setup in which authors test their model against the state-of-the-art is logical and well-designed for the task of large-scale object detection. They carry a great amount of ablation studies to show each component of their model works as they hypotesize. They also focus on model interpretability and project their new category representations into a lower dimensional space with t-SNE to show how semantically related categories lay close to each other in this new feature space. As a negative side of the paper, it does not mention some of the related work hence the model proposed has not been tested against them.

The paper is very coherent and it reads really well. The graphics they use to explain the internal mechanism of their model are designed perfectly such that by just looking at the graphics, one can get a great amount of sense about their model. Their model is proven to improve the state-of-the-art large-scale object detection. I believe this paper well-deserved the attention it got at CVPR ‘19.

5 Spatial-aware Graph Relation Network for Large-scale Object Detection [7]

In this paper, authors claim that the state-of-the-art object detectors treat each region separately and this leads a big performance drop when images comes from a long-tailed data distribution and they propose a model that can be stacked on any two-stage object detector. At a high level, their model consists of two main submodules. The relation learner module takes RoI-pooled region visual features from the backbone detector and returns a sparse adjacency matrix. The spatial graph reasoning module takes learned weights for object categories and soft-mapping scores for proposed regions from the backbone detector and multiplies them to build new visual representations for the regions. Having built new region representations, their graph reasoning module employs a graph convolutional network for message passing between regions through edges weighet by both the adjacency matrix from relation learner module and a couple gaussian kernels learned on spatial relationship between the regions. Finally, these new region representations are concatenated with the original region features and sent to a fully-connected layer series that performs classification and bounding-box regression.

There are two main scientific contributions brought to context-guided object detection topic by this paper. First, they empirically show that even without using an external knowledge source, we can train detectors which are robust to imbalanced class distributions by modeling object-object relationships exploiting the features that already reside in the dataset. Second, their use of polar coordinates to encode spatial settings of object pairs is the first-time in the literature and seems to be very effective.

The problem they attempt to solve is very challenging as it is too hard to learn accurate semantic relationships for classes reside at the end of a long-tailed data distribution. Their formulation of the problem looks very precise and their design-specific decisions are mostly backed up by strong arguments. Specifically, the use of gaussian filters to learn spatial settings of object pairs over polar coordinates is very intuitive and proved emprically to be effective. They mostly cite the related work and summarized their pros and cons elegantly. The experiments they perform to test their model against the state-of-the-art looks appropriate for the task of large-scale object detection. Moreover, they carry out an well-designed ablation study to prove all submodules which their model is built upon work as hypotesized and contributes to the overall model performance. As a negative side, they neither mention nor emprically show how they conclude some parameters values. For example, building the adjacency matrix in their relation learner module, they only retain top k relations to make the matrix sparse. However, they does not give any information how they choose $k = 32$ for their experimental setup.

Overall, this paper is very coherent from start to end, and reads very well. The graphs they use to convey information about how their model actually works are well-designed. Their model is emprically proven to be effective and improve the state-of-the-art.

References

[1] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4086–4096, 2017.

[2] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7239–7248, 2018.

[3] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6985–6994, 2018.

[4] Hang Xu, Chenhan Jiang, Xiaodan Liang, Liang Lin, and Zhenguo Li. Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[7] Hang Xu, Chenhan Jiang, Xiaodan Liang, and Zhenguo Li. Spatial-aware graph relation network for large-scale object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9298–9307, 2019.