REVIEW ON OBJECT DETECTION WITH CNN

Asad Tayyab[1], Prajwal Rai[2], Pranjal Jain[3], Sarala D V[4]

[1,2,3]7th Semester, Department of Computer Science and Engineering , Dayananda Sagar College of Engineering , Bengaluru
[4]Assistant Professor, Department of Computer Science and Engineering, Dayananda Sagar College of Engineering , Bengaluru

---***---

Abstract - *Object detection is one of the most important achievements of deep learning and image processing since it finds and recognises objects in images. An object detection model may be trained to recognise and detect several objects, making it adaptable. Object detection had a large scope of interest even before deep learning approaches and modern-day image processing capabilities. It has become considerably more widespread in the present generation, thanks to the development of convolutional neural networks (CNNs) and the adaptation of computer vision technology. The latest wave of deep learning approaches to object detection gives up apparently limitless possibilities.*

Key Words : CNN, Object Detection, Neural Networks, Classification, Equilibrium

1. INTRODUCTION

1.1 : Object Detection

Recognition system, which includes detection and recognition in digital files, is a term that applies to a set of interdependent image classification tasks.

The method of guessing the class of a single object in an image is known as picture classification. Entity positioning is the process of finding several or even more items in a visual and drawing a boundary around its range. By detecting and describing any or maybe more elements in a shot, image segmentation merges tasks. Often, whenever a study or user talks about "pattern recognition," they're talking about "object detection."

1.2 : Convolutional Neural Network (CNN)

A CNN is a Neural Supervised learning model which can receive an image and provide priority (fairly basic processing elements) to specific facets in the image, as well as differentiate amongst them. Alternative approaches require significantly more pre-processing than for a CNN. Although simple procedures need hand-engineering of filter, with just enough retraining, CNNs may acquire such filters/characteristics.

2. PROBLEM IDENTIFICATION

Object Detection is a topic which always comes into picture when we talk about Machine Learning and Computer Vision. In Computer Vision, Object Detection and Classification with Machine Learning allows a camera to "see" as humans do, recognising each physical shape as a car, dog, or person. Human identification in real time is becoming increasingly popular among data scientists and in a variety of industries, including smart cities, retail, and surveillance. This demonstrates the need for systems that can detect objects with more precision and accuracy.

In this paper, we review various algorithms that have been built to enhance object detection systems. We discuss various Machine Learning algorithms such as R-CNN and a variety of approaches that various researchers and practitioners have historically employed to solve the task.

3. RELATED WORKS

3.1: K-SVD: Algorithm for Designing Overcomplete Dictionaries for Sparse Representation

Michal Aharon et al [1] suggested an approach for updating dictionaries to achieve sparse signals. The dictionary that leads to the best representation for each member of the set is sought given a set of training signals and rigorous sparsity requirements. The K-SVD technique is a revolutionary generalisation of a K-means clustering algorithm. For tasks including having to fill in vacant pixels and downsizing with both synthetic and real images, the K-dictionary SVD outperforms equivalents namely the non decomposed Haar and overcomplete or unitary DCT. Yet, while functioning with larger picture portions, K-scalability tends becomes a hurdle. SVD's

3.2 : Deep Equilibrium Models

The deep equilibrium model(DEQ) is a new way to modelling sequential data presented by Shaojie Bai et al [2]. DEQs have been shown to function with two cutting-edge deep sequence models, self-attention transformers and trellis networks.DEQ only uses O(1) memory during training, is unaffected by the forward pass's root solver, and is versatile enough to accommodate a broad range of architectural choices. However, it's unclear whether there's any benefit over utilizing gradient checkpointing.

3.3 : Multiscale Deep Equilibrium Models

The multiscale deep equilibrium model (MDEQ) was introduced by Shaojie Bai et al [3], which is a special type of neural system that works well in huge, nonlinear pattern matching environments. Leveraging latent differential to minimize saving phases, the MDEQ resolves for it and downstream radiates via the equilibria of several component resolves at the very same moment. An MDEQ, unlike previous implicit systems such as DEQs as well as Neuro ODEs, settles for and iterates back via coordinated equilibrium of numerous image features at various resolutions.

3.4: GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond

The Non-Local Network (NLNet) involves combining query-specific wider perspective to every inquiry region, a groundbreaking strategy for collecting lengthy linkages as described by Yue Cao et al [4]. But at the other end, a robust scientific analysis demonstrated that the worldwide environments given by non-local systems for various search places inside a photograph are nearly identical. This discovery is being used in the investigation to create a simplified approach that relies on a query-independent construction that maintains the NLNet's efficiency while requiring significantly less computing power. By deploying GC units to various layers, GCNet usually beats basic NLNet & SENet on important identification tasks.

3.5 : End-to-End Object Detection with Transformers

Nicolas Carion et al. [5] describe a new method for detecting objects that treats the problem like a direct set prediction problem. This method streamlines the identification process by removing the need for a variety of custom elements that explicitly reflect our previous employment experience, such as from a non-maximum reduction method or anchoring generation. The new structure, named DEtection TRansformer or DETR, includes a collection worldwide deficit which generates distinct forecasts via bipartite matching, and also a transformer encoder-decoder design. Given a fixed restricted amount of taught item inquiries, DETR explains about the links between both the objects and the global visual backdrop to swiftly construct the final session of recommendations in parallel. DETR is easy to set up and operate, thanks to its modular architecture, which can be swiftly expanded to include panoptic segmentation and competitive outcomes. Furthermore, it beats Faster R-CNN on larger objects, owing to the self-processing attention of global information. However, this new detector architecture introduces new challenges, particularly in terms of training, optimization, and small-object performance. Current detectors have had to go through several years of research to be able to deal with similar issues.

3.6 : Scalable Object Detection using Deep Neural Networks

Dumitru Erhan et al [6] introduced a saliency-based recognition system which forecasts a set of class-agnostic embeddings as well as an unique rating for every unit depending on its chances of containing particular object of interest The approach controls a changeable samples for every category at the network's top tiers, enabling for cross-class generalization. The strategy proved to attain superior recognition rate on VOC2007 and ILSVRC2012 using just the topmost few anticipated places within every photo and a tiny proportion of neural network tests. To express the link between distinct projections, solely multilayer or complete layers are used, and a hand-designed NMS post-processing can enhance their effectiveness.

3.7 : Selective Kernel Networks

Xiang Li et al [7] proposed that every neuronal in a CNN can modify the extent of it's own receptive field based on a range of input feature scales using a continuous selection method. Softmax focus guided by the knowledge in such limbs is used to combine multiple forks with different kernel sizes using a design block called Selective Kernel (SK) unit.Deep networks made up of numerous SK units are known as Selective Kernel Networks (SKNs) (SKNets). On a variety of benchmarks, SKNets demonstrate cutting-edge performance, ranging from large to small models.

3.8 : Non-local Neural Networks

Xiaolong Wang et al [8] presented non-local procedures were given as a component of construction pieces for recording lengthy exchanges. In this non-local operation, which is influenced by the conventional non-local means approach in computer vision, the response at a point is computed as a weighted sum of the characteristics at all places. Adding non-local blocks increases performance considerably over baseline on all tests. In order to represent pixel-level pairwise relations, NLNet learns query-independent focus mappings for every search area, wasting computational power.

3.9 : CBAM: Convolutional Block Attention Module

Sanghyun Woo et al [9] presented the Convolutional Block Attention Module (CBAM), which works as a feed-forward deep neural networks, a simple and effective focus component. From an initial convolution layer, the component implies focus maps in two aspects: channel and spatial, and afterwards combines the focus maps to the input feature map for responsive product improvement. On three independent benchmark datasets, CBAM beats all baselines: ImageNet1K, MS COCO, and VOC 2007. To equalize the relevance of various geographic areas and channels, CBAM uses rescaling. This method, on the other side, employs hybrid algorithm by resizing, which really is unsuccessful for world of global competition modeling.

3.10 : Squeeze-and-Excitation Networks

Jie Hu et al [10] proposed the Squeeze-and-Excitation (SE) block as a unique architectural element for increasing the quality of representations produced by a network by explicitly modelling the interdependencies between its convolutional features' channels. They allow the network to learn to selectively emphasise informative properties while suppressing less valuable ones, allowing feature recalibration. SENets have been found to perform well in a variety of tests, producing cutting-edge outcomes across a wide range of datasets and activities. The inadequacy of prior methods to simulate channel-wise feature interactions properly was also revealed by SE blocks.This method, on the other side, employs hybrid algorithm by resizing, which really is unsuccessful for world of global competition modeling.

3.11 : Implicit neural representations with periodic activation functions

Vincent Sitzmann et al. [11] suggested using periodic activation functions for implicit neural representations and shown that these networks, dubbed SIRENs, are capable of replicating complicated natural signals and their derivatives. A deep learning system's representation provides accurate representations of natural signals such as pictures, audio, and video. However, because the sine function is periodic, rising up the hill or taking a significant stride during training may cause you to fall back down.

3.12 : End-to-end memory networks

Over a potentially enormous external memory, Sainbayar Sukhbaatar et al [12] suggesteda recurrent attention model in a neural network. Because the network is taught from start to finish, it requires far less monitoring during training, making it more practical in real-world applications. On a variety of tasks ranging from question answering to language modelling, it demonstrates how backpropagation may be used to train a neural network with an explicit memory and a recurrent attention mechanism for reading the memory. The model, on the other hand, falls short of memory networks trained with strong supervision, failing a number of the 1k QA tests.

3.13 : Attention Is All You Need

The Transformer is a novel fundamental network architecture designed by Ashish Vaswani et al [13] that is exclusively dependent on attention processes, with no repetition or convolutions. In comparison to systems using recurrent or convolutional layers, the Transformer is substantially faster to train. The Transformer will, however, be expanded to issues that require input and output modalities other than text, and local, constrained attention approaches will be examined to efficiently manage massive inputs and outputs including graphics, audio, and video.

3.14 : Implicit Feature Pyramid Network for Object Detection

Tiancai Wang et al [14] proposed modeling the transformation of FPN utilizing an implicit function newly constructed in the deep equilibrium model (DEQ) and constructing a residual-like iteration to update the hidden states efficiently. The proposed i-FPN considerably increases the performance of object detectors. Unrolling solvers, on the other hand, result in a significant memory expense as the number of iterations grows, even though they achieve weight-sharing of all unrolled blocks.

3.15 : Pyramid vision transformer: A versatile backbone for dense prediction without convolutions

Wenhai Wang et al [15] proposed the Pyramid Vision Transformer (PVT), which employs a progressive decreasing pyramid to reduce the number of calculations required for large feature maps and overcomes the limits of transferring Transformer to other dense prediction tasks. The proposed PVT outperforms well-built CNN backbones under equal amounts of parameters; nonetheless, there are still some specialized modules and processes established for CNNs that were not studied in this work.

3.16 : SAL: Sign Agnostic Learning of Shapes from Raw Data

Sign Agnostic Learning, a deep learning technique for learning implicit; Matan Atzmon et al [16] proposed interpretations using unregistered geographic information including point clouds and triangle soups. SAL is capable of reconstructing elevated forms from unprocessed point clouds and integrates effortlessly into common models for acquiring shape regions from basic geometric features. It does, however, have certain drawbacks, the most significant of which being its inability to capture thin structures.

3.17 : Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling

Inside the framework of sequence modeling, Junyoung Chung et al [17] assessed the long short-term memory (LSTM) unit, the gated recurrent unit (GRU), and a much more traditional hyperbolic tangent unit. The RNNs with gated subunits (GRU-RNN and LSTM-RNN) clearly outperformed the more traditional tanh-RNN from both Ubisoft datasets. The benefits of latching units over traditional recurrent units were clearly illustrated in these tests.

3.18 : Convolutional Sequence to Sequence Learning

The usage of a totally convolutional neural network-based architecture was proposed by Jonas Gehring et al [18].It implies that, unlike recurrent models, computations across all components may be entirely parallelized during training to make use of GPU power, and that refinement is easier because the number of non-linearities is independent of the input length. This approach beats sophisticated recurrent models by an order of magnitude on very large benchmark datasets.

3.19 : Cascade R-CNN: Delving into High Quality Object Detection

Zhaowei Cai et al [19] introduced the Cascade R-CNN, a multi-stage object recognition architecture to solve the problems that tend to hamper detection performance as IoU thresholds are increased. The detectors are trained one by one, with the output of each detector being utilised to train the next higher grade detector if the output is a decent distribution. The architecture prevents overfitting during training and quality mismatch during inference. The Cascade R-CNN has been shown to enhance a variety of object identification designs.

3.20 : Mask R-CNN

Kaiming He et al [20] presented a conceptually simple, versatile, and extensive object instance segmentation system that recognises objects in an image while also providing a high-quality segmentation mask for each instance. An object mask prediction branch is added to the current bounding box recognition branch, resulting in a quicker R-CNN. The proposed strategy exceeds all existing single-model entries on every challenge and is straightforward to learn. However, because it only works with static photos, it is unable to investigate temporal information about the object of interest, such as dynamic hand gestures.

3.21 : Memory Networks

Memory networks are a novel type of learning model described by Jason Weston et al [21]. Memory networks require inference components and a long-term memory component in order to reason, and they've figured out how to combine the two. Reading and writing to long-term memory with the objective of predicting anything is possible. Words that had never been seen before could be modeled. Models could also be improved to account for when data was written to a memory slot. Experiments with just a single dictionary and linear embeddings, on the other hand, produced lower results.

3.22 : Robust Face Recognition via Sparse

Representation

John Wright et al [22] studied occlusion and disguise, as well as the challenges of automatically identifying human faces from frontal viewpoints with variable expression and illumination. Many linear regression models were recast as classification problems, with the assertion that a novel theory derived from sparse signal representation contains the key to solving it. For image-based object recognition, a general classification method based on a sparse representation derived via '1-minimization' was developed. Traditional features like Eigenfaces and Laplacianfaces, for example, perform equally well when down sampled pictures and random projections are used. The framework, on the other hand, is concerned about the number of features being large enough and the sparse representation being generated appropriately.

3.23 : Disentangled Non-Local Neural Networks

The non-local component is a common component for enhancing a conventional neural network's environment capacity, according to Minghao Yin et al [23]. The attention calculation of the non-local block may be separated into two terms, according to this article: a whitened pairwise term accounting for the connection between two pixels and a unary term representing the saliency of each pixel. It's worth noting that the two words that were taught separately represent two different types of visual input. The DNL block is introduced, which makes learning whitened pairwise and unary phrases more cooperative. When both the whitened pairwise term and the unary term are present within a non-local block, they do not learn as powerful visual cues.

3.24 : ResNeSt: Split-Attention Networks

Hang Zhang et al [24] propose a Split-Attention block that allows attention to be shared between feature-map groupings. By stacking these Split-Attention blocks ResNet-style, a new ResNet version dubbed ResNeSt is created. The ResNet structure is preserved by the network, allowing it to be used in downstream processes without incurring additional computational costs. The proposed ResNeSt outperforms all current ResNet variants, improving speed-accuracy trade-offs while maintaining computational economy. However, "network surgery" is required to increase performance on a given computer vision task in order to tweak the ResNet to make it more effective for that task.

3.25 : A non-local algorithm for image denoising

Antoni Buades et al [25] created a new metric, method noise, to assess and compare the performance of digital picture denoising algorithms. For a major class of denoising techniques, especially the local smoothing filters, this is how noise is computed and investigated. The nonlocal means (NL-means) approach and various comparison tests between the NL-means algorithm and the local smoothing filters are discussed.

3.26 : Neural Machine Translation by Jointly Learning to Align And Translate

Dzmitry Bahdanau et al [26] hypothesize using a fixed-length variable is a major limitation in boosting the effectiveness of this primitive encoder–decoder design, and we propose that this be overcome by enabling a framework to instantaneously (soft-)search for aspects of a text line that are suitable for determining a word correctly without possessing to form such parts as an amorphous phase. The suggested technique was comparable to present phrase-based statistical machine translation in terms of performance. Better handling of unfamiliar or unusual words, on the other hand, will have to wait till the future.

3.27 : 3D Face Recognition under Occlusion using Masked Projection

Nese Alyuz et al. [27] developed a complete automatic 3D face recognition system that is occlusion-resistant. They primarily deal with two issues: Missing data treatment for classification based on subspace analysis, as well as occlusion handling for surface registration. Even when there are a lot of occlusions, expressions, and tiny position variations, the proposed technique can function well. When the face is turned more than 30 degrees, however, achieving the initial alignment becomes more difficult.

3.28 : RMPE: Regional Multi-Person Pose Estimation

Hao-Shu Fang et al [28] provide a novel regional multi-person pose estimation framework to aid position estimation in the presence of faulty human bounding boxes. On the multi person dataset, it achieves 76.7 mAP by handling duplication detections. Due to its reliance on second-order body components, the part-based framework loses its capacity to identify body parts from a global position viewpoint.

3.29 : Memory Enhanced Global-Local Aggregation for Video Object Detection

According to Yihong Chen et al [29], people recognise items in movies using two essential hints: global semantic information and local localisation information. The memory enhanced global-local aggregation network is introduced in this paper, and it is one of the first to consider both global and local information. Furthermore, their proposed MEGA might allow the key frame to access substantially more material than prior techniques thanks to a new and properly developed Long Range Memory module. On the ImageNet VID dataset, our approach delivers state-of-the-art performance. When it comes to localization, though, a lack of geographical information would constitute a problem.

4. CONCLUSIONS

Object detection is used widely in most computer and robot vision systems. Although remarkable progress has been made in recent times, as well as some scoring systems have indeed been integrated into a range of consumer electronics and supported mobility technologies, we are still a considerable distance from human-level performance, particularly in open-world learning. It's worth noting that object detection isn't widely employed in many areas where it may be quite useful.

Object detection systems have become increasingly important as robotic systems and abstract machine in general become more ubiquitous. Finally, object monitoring systems will be necessary for nano-robots or robots exploring locations heretofore unrecognized to humans, such as deep water depths or even other planets, and all these detection techniques will be expected to learn new item classes as they are found. In such cases, the ability to learn in a real-time open-world setting would be crucial.

5. ACKNOWLEDGEMENTS

We would like to express our sincere thanks to Dr. Vindhya P Malagi, Dr. Ramya R S and Prof. Sarala D V for their cooperation and guidance. We would also like to thank our computer department HOD, Dr. Ramesh Babu D R and all the staff who have directed us to grow this project concept throughout.

6. REFERENCES

[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on signal processing, 54(11):4311–4322, 2006.

[2] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In Advances in Neural Information Processing Systems (NeurIPS), 2019.

[3] Shaojie Bai, Vladlen Koltun, and J Zico Kolter. Multiscale deep equilibrium models. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

[4] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCV Workshop), 2019.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision (ECCV), pages 213–229, 2020.

[6] Dumitru Erhan, Christian Szegedy, Alexander Toshev, Dragomir Anguelov. Scalable Object Detection Using Deep Neural Networks. arXiv preprint arXiv:1312.2249v1, 2013.

[7] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 510–519, 2019.

[8] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, 2018.

[9] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), pages 3–19, 2018

[10] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[11] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

[12] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In Advances in Neural Information Processing Systems (NeurIPS), 2015.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko reit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), 2017.

[14] Tiancai Wang, Xiangyu Zhang, and Jian Sun. Implicit feature pyramid network for object detection. arXiv preprint arXiv:2012.13563, 2020.

[15] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, KaitaoSong, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. arXiv preprint arXiv:2102.12122, 2021.

[16] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In Proc. CVPR, 2020.

[17] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint: 1412.3555, 2014.

[18] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. arXiv preprint arXiv:1705.03122v2, 2017.

[19] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In CVPR, 2018.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In Proc. IEEE Int. Conf. Comp. Vis., 2017.

[21] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In International Conference on Learning Representations (ICLR), 2015.

[22] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse

representation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 31(2):210–227, 2008.

[23] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 191–207, 2020.

[24] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. ResNeSt: Split-attention networks. arXiv preprint arXiv:2004.08955, 2020.

[25] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A non-local algorithm for image denoising. In Computer Vision and Pattern Recognition (CVPR), 2005.

[26] Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[27] Nese Alyuz, Berk Gokberk and Lale Akarun. 3-D Face Recognition Under Occlusion Using Masked Projection. In IEEE Transactions on Information Forensics and Security (Volume: 8, Issue: 5, May 2013), 2013

[28] Hao-Shu Fang , Shuqin Xie , Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2334–2343 (2017)

[29] Yihong Chen, Yue Cao, Han Hu, and Liwei Wan. Memory enhanced global-local aggregation for video object detection. In: The Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)