Yash Sanjay Bhalgat

Statement of Purpose

Senior ML Researcher, Qualcomm AI Research Webpage: `https://yashbhalgat.github.io`

My primary research interest lies in Geometric Deep Learning and its applications to 3D Computer Vision, particularly focusing on the inference efficiency characteristics of these models. The deep learning advances of the last decade have helped us achieve superhuman capabilities on image (and speech) recognition problems. However, fundamental differences arise when we move on to non-Euclidean manifolds such as point clouds and meshes. Rapidly emerging applications in autonomous vehicles and AR/VR use-cases have led to a lot of progress in deep learning algorithms for 3D perception, but these algorithms still fall short of the low-latency/low-power requirements of these applications. Additionally, recent research on the expressivity/complexity of Graph Neural Networks leads to several challenging questions about developing *scalable* 3D vision models. My long term research goal is to develop deep networks that are robust to the changes in different 3D environments, and concurrently design geometric deep learning architectures that are capable of *efficiently* handling 3D data.

I am also fascinated by the area of unsupervised/self-supervised representation learning. Using self-supervised algorithms that incorporate various inductive biases available in the 3D geometries of the environment can be used to learn generalizable semantic representations. I am inspired by the research in equivariant deep learning, in particular Gauge Equivariant CNNs proposed by my colleagues at QUVA. In my PhD research, I would like to explore different ways of learning meaningful representations for 3D data or more generally multi-modal data (3D + 2D + various sensors) to help models generalize to challenging environments.

My research at Qualcomm AI Research has been focused on algorithm and system design to develop efficient deep networks for resource-constrained edge devices. Specifically, I have worked on developing pipelines for structured/unstructured pruning [1, 2] and low-bit quantization [3, 4] of deep networks as well as conditional computing for various use-cases including image & video classification/segmentation, gaze estimation and hand-pose estimation. My current research at Qualcomm focuses on developing compute-adaptive deep networks, i.e. models that can scale (at runtime, without training) with the changes in available compute (a popular example of this being Once-for-all networks, ICLR 2020). This is especially useful for autonomous driving applications where, for example, you would want a scene-parsing model to have a lower or higher latency depending on the speed of the vehicle. Alongside the conventional model compression methods, incorporating inductive biases can also be a way to develop efficiently parametrized models. For example, in hand-pose estimation pipeline, we are currently studying a structural consistency loss that relies on the fact that the hand-joint estimations cannot lie outside the boundary of the hand. Using such a simple inductive bias can be used to train smaller models that make sensible predictions. Similarly, temporal consistency is a strong inductive bias that can be used to build efficient video processing models. At Voxel51, I worked with Prof. Jason Corso from University of Michigan on developing real-time pipelines for vehicle detection and tracking to perform querying on large-scale video databases. In these implementations, I heavily relied on ideas such as feature-reuse across consecutive frames to refine the predictions of my models. My near-term research goal is to extend and/or combine the above ideas for designing low-complexity architectures capable of modelling long-range interactions across several components in 3D environments.

Hardware-software co-design has been an important aspect of my work as deep networks that are efficient in theory do not run as efficiently on general-purpose (or even domain-specific, e.g. accelerator) hardware platforms. This is wonderfully explained in a recent article by Sara Hooker [5]. For example, in our work on Learned Threshold Pruning [2], we developed a gradient-based algorithm to induce sparsity in deep networks that led to as high as 26× compression on AlexNet with negligible performance loss. But this doesn't translate to a 26× latency improvement on CPUs/GPUs, because the unstructured nature of this sparsity can't be exploited by these platforms. This led to several considerations, such as using block-wise sparsity or using specific sparsity patterns that are amenable to hardware. I also actively participated in the Neural Architecture Search team brainstorming sessions where we discussed ways to incorporate hardware constraints as metrics to be optimized by our automated search algorithms. For the last year, I have been working on a hardware accelerator project (on the deep learning Systems side) where we ran into several considerations about the hardware constraints while designing our deep networks.

Apart from my work at Qualcomm and Voxel51, I have also pursued several projects and internships to upskill myself. My first deep learning related internship was at IBM Research - Bangalore, where I used a Common Representation Learning (CRL) based approach to build a fast catalog search engine for large fashion databases. At IIT Bombay, I pursued my Bachelors thesis under the guidance of Prof. Vikram Gadre where I completed a comprehensive research study on using Scattering Wavelet Networks (ScatNets) for robust feature extraction and eventually used it for classification of latent fingerprints (i.e., raw imprints obtained from forensic documents). This work was done in a collaboration with the Department of Cyber Security, Maharashtra. I was awarded the **Undergraduate Research Award** (URA 02) by IIT Bombay for my thesis. One of my parallel work on ScatNets was during my internship with the Image and Signal Processing group at IFPEN, Paris on seismic sensor images - this work was presented at ICASSP 2018 [6].

While the above projects involved large-scale image & video datasets, I also pursued two projects that dealt with learning from limited or noisy data. The first was when I interned with the Watson Languages group at IBM Research - Almaden, where I developed pipelines to train sentiment classification models using noisy labels. I implemented several ideas from semi-supervised learning and otherwise (e.g. the Noise Adaption Layer, ICLR 2016) and eventually proposed a Teacher-Student learning method based on a curriculum learning to tackle this problem [7]. The second project involved segmentation of anatomical structures in chest radiographs, where we proposed an LP-based active learning framework to utilize weaker forms of annotations (bounding boxes and landmark points) to train a segmentation model in a *mixed-supervision setting* [8].

Moreover, a predilection for teaching and my academic performance led me to serve as a Graduate Student Instructor (GSI) / Teaching Assistant for *Computational Data Science*, *Logic for Computer Science*, *Quantum Mechanics & Applications* and *Wavelets* courses at UMich as well as IIT Bombay. I was one of the top nominees for the **Towner Prize for Outstanding GSIs** for my work at UMich. These enriching teaching experiences exposed me to new tools that I couldn't have learnt otherwise. For example, in the Computational Data Science under Prof. Raj Nadakuditi's guidance, we created engaging lectures and assignments using Jupyter notebooks in Julia that gave the students hands-on experience with a variety of algorithms. I also helped setup challenging projects for this course in Julia that really taught me about the different considerations regarding logistics and evaluation to help the students best learn their concepts. I am currently mentoring John Yang, last year PhD with Prof. Nojun Kwak, on his summer internship with my team, where we are working on developing a Conditional Compute based architecture for 3D hand-pose estimation.

⟨ I will add University specific professors information here. ⟩

References

[1] Yash Bhalgat, Yizhe Zhang, Jamie Lin, and Fatih Porikli. Structured convolutions for efficient neural network design. In *Advances in Neural Information Processing Systems*, 2020.

[2] Kambiz Azarian, Yash Bhalgat, Jinwon Lee, and Tijmen Blankevoort. Learned threshold pruning. *arXiv preprint arXiv:2003.00075, under review*, 2020.

[3] Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[4] Yash Bhalgat, Jangho Kim, Jinwon Lee, Chirag Patel, and Nojun Kwak. Qkd: Quantization-aware knowledge distillation. *arXiv preprint arXiv:1911.12491*, 2019.

[5] Sara Hooker. The hardware lottery. *arXiv preprint arXiv:2009.06489*, 2020.

[6] Y. Bhalgat, J. Charléty, and L. Duval. Catseyes: categorizing seismic structures with tessellated scattering wavelet networks. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[7] Yash Bhalgat, Zhe Liu, Pritam Gundecha, Jalal Mahmud, and Amita Misra. Teacher-student learning paradigm for tri-training: An efficient method for unlabeled data exploitation. In *KONVENS*, 2019.

[8] Yash Bhalgat, Meet Shah, and Suyash Awate. Annotation-cost minimization for medical image segmentation using suggestive mixed supervision fully convolutional networks. In *Medical Imaging meets NeurIPS workshop*, 2018.