Enhancements in Distributed Matching for Crowd Counting

Priyanshu Raj Jindal

November 6, 2024

1 Introduction

The DM-Count framework addressed crowd counting by viewing it as a distribution matching problem using Optimal Transport (OT), achieving strong results. However, DM-Count's high computational complexity, which grows quadratically with image size, makes it difficult to apply in real-time or on high-resolution images. To address this, I experimented with an improvement, Hierarchical Optimal Transport (HOT), which breaks down the OT problem into smaller regions, each calculated independently to reduce computational load. My goal was to see if HOT could help balance accuracy with better efficiency in processing.

2 Approaches and Results

I explored two main approaches to improve DM-Count's performance. Despite the potential of each approach, practical results fell short of expectations due to both computational and training stability issues.

2.1 Hierarchical Optimal Transport Loss

The first improvement attempted was to replace the standard loss in DM-Count with a **Hierarchical Optimal Transport (HOT) Loss** to better handle crowd density variation. HOT introduces a hierarchical decomposition of the global optimal transport (OT) problem, dividing images into smaller regions to compute transport loss locally, then aggregating the results.

Challenges: The implementation did not achieve the expected reduction in computation time. A key limitation I noticed and believe is the heavy reliance on CPU for computations of splitting images and managing corresponding dot annotation maps, since attempts to vectorize this process for GPU execution led to significant dimensional errors. These errors arose from incompatibilities in hierarchical grid-based operations on the GPU. Thus, HOT failed to provide the anticipated efficiency gains.

2.2 Multifaceted Attention-Based Approach

Given the limitations of HOT, the next approach was inspired by the work of Lin et al., titled *Boosting Crowd Counting via Multifaceted Attention* [1]. This approach introduced multifaceted attention mechanisms to the DM-Count model, incorporating both global and local attention modules to improve the model's responsiveness to crowd density variations. Unlike the baseline DM-Count, which does not distinguish between sparse and crowded regions, this approach aimed to adapt density map predictions based on local crowd distributions.

Implementation: The global and local attention layers were added following a pretrained VGG-19 backbone to enhance the feature extraction capabilities of the model. The global attention module was designed to capture large-scale dependencies, while the local attention module refined the density estimation in specific regions.

Challenges and Unstable Training: This approach resulted in highly unstable training. Potential reasons for this instability include:

- **Parameter Sensitivity:** The multifaceted attention modules required precise parameter tuning to achieve stable results. Without optimal tuning, the attention mechanisms could overfit to specific density regions, leading to high variance in predictions.

- **Compatibility with Backbone Model:** The VGG-19 backbone, although effective for generic feature extraction, may not have provided the optimal features for the attention layers to effectively distinguish between sparse and crowded regions, causing further instability in training.

3 Changes in Code Details

Both approaches are available on GitHub and run with the same commands as the original code. Key updates are outlined below:

3.1 Approach 1: HOT Loss Changes

Only `train_helper.py` was modified for HOT. Notable updates include:

- `compute_adaptive_hot_loss`: This function calculates the HOT loss by dividing the image into a grid and applying OT within each cell. Depending on the image size, the function adjusts the number of grid cells dynamically.
- Mixed Precision Training: `autocast` and `GradScaler` were added to optimize GPU memory usage.

3.2 Approach 2: Attention Module Additions

For the second approach, I made structural changes to incorporate `GlobalAttention` and `LocalAttention` modules:

- **GlobalAttention** captures global relationships in the feature maps to understand broader context.
- **LocalAttention** focuses on specific local patterns, refining details in densely populated areas.
- **Baseline Model Change - Density Map Prediction Head**: A softmax layer was also added at the end to make the output compatible with the OT-based loss function, which requires normalized values.

Minor updates were also made in `train_helper` and `test` files to ensure compatibility with the modified model.

4 Conclusion

In summary, while both approaches offered theoretical benefits, they encountered practical issues during implementation. The HOT approach faced challenges with CPU dependence and dimensional errors on the GPU, while the attention-based model struggled with unstable training. Future work may involve further optimizing HOT for GPU compatibility or experimenting with alternative backbones and attention settings to stabilize training.

References

[1] H. Lin, Z. Ma, R. Ji, Y. Wang, X. Hong, ”Boosting Crowd Counting via Multifaceted Attention,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.