Report Summary: DynamicViT - Efficient Vision Transformers with Dynamic Token Sparsification

Overview
DynamicViT introduces a novel approach to accelerate vision transformers by dynamically pruning redundant tokens during inference while maintaining model accuracy. The key insight is that vision transformers often rely on only a subset of informative tokens for final predictions, allowing for significant computational savings without performance degradation.

Key Contributions
1. **Dynamic Token Sparsification**:
   - Implements hierarchical token pruning based on input-specific importance scores
   - Progressively removes up to 66% of input tokens across network layers
   - Achieves 31-37% FLOP reduction and >40% throughput improvement with <0.5% accuracy drop

2. **Technical Innovations**:
   - Lightweight prediction module combining local and global token features
   - Attention masking strategy for differentiable token pruning during training
   - Gumbel-Softmax for end-to-end optimization of pruning decisions
   - Hierarchical pruning with geometric sequence of keeping ratios

3. **Performance**:
   - Demonstrated on DeiT-S, LV-ViT-S and LV-ViT-M backbones
   - Outperforms structural downsampling and static pruning methods
   - Achieves better FLOPs-accuracy tradeoffs than model width scaling
   - Maintains effectiveness on larger models and higher resolution inputs

Implementation Details
- Three-stage progressive pruning (< Â Á², < 2’ v—F, Á typically 0.7-0.9
- Prediction module uses MLPs to combine local and global features
- Training incorporates distillation and ratio constraint losses
- Inference directly prunes tokens based on predicted importance

Significance
DynamicViT opens a new direction for transformer acceleration by exploiting spatial sparsity in vision tasks. The method demonstrates that vision transformers can achieve CNN-like efficiency through input-adaptive computation while maintaining their performance advantages. The approach is particularly valuable for real-time applications where computational efficiency is critical.

Future Directions
The paper suggests extending this approach to other vision tasks like video classification and dense prediction, as well as exploring joint optimization with other efficiency techniques like quantization and knowledge distillation.