AdaViT: Adaptive Vision Transformers for Efficient Image Recognition
Appendix

Lingchen Meng1,2,3* Hengduo Li4* Bor-Chun Chen5 Shiyi Lan4
Zuxuan Wu1,2† Yu-Gang Jiang1,2 Ser-Nam Lim5
1Shanghai Key Lab of Intelligent Info. Processing, School of Computer Science, Fudan University
2Shanghai Collaborative Innovation Center on Intelligent Visual Computing
3Biren Technology 4University of Maryland 5Meta AI

The Least Computation

The Most Computation

hammerhead abacus balloon barbell
tabby accordion bakery carpenter’s kit
candle crane house finch space shuttle
chainsaw grand piano limousine stage
stupa war plane street sign honeycomb
stretcher washer pretzel barn

Figure 1. Qualitative results. Images allocated with the least (Left) and the most (Right) computational resources by AdaViT are shown.

A. Qualitative Results

We further provide more qualitative results in addition to those in the main text. Images that are allocated the least/most computational resources by our method are shown in Figure 1, demonstrating that our method learns to use less computation on easy object-centric images and more computation on hard complex images with cluttered background. Figure 3 shows more visualization of the learned usage policies for patch selection, demonstrating the pattern that our method allocates less and less computation gradually throughout the backbone network, which indicates that more redundancy in computation resides in the later stages of the vision transformer backbone.

B. Compatibility to Other Backbones

Our method is by design model-agnostic and thus can be applied to different vision transformer backbones. To verify this, we use DeiT-small [1] as the backbone of AdaViT and show the results in Figure 2. AdaViT achieves better efficiency/accuracy tradeoff when compared with standard
variants of DeiT, and consistently outperforms its Random+ baseline by large margins, as demonstrated in Figure 2(a) and 2(b) respectively.

We further show the visualization of patch selection usage policies with DeiT-small as the backbone as well in Figure 3. A similar trend of keeping more computation at earlier layers and gradually allocating less computation throughout the network is also observed.

References


[2] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng
Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In ICCV, 2021. 2