CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification (Supplementary Material)

Chun-Fu (Richard) Chen, Quanfu Fan, Rameswar Panda
MIT-IBM Watson AI Lab
chenrich@us.ibm.com, qfan@us.ibm.com, rpanda@ibm.com

Summary This supplementary material contains the following additional comparisons and hyperparameter details. We first provide more comparisons between the proposed CrossViT and DeiT (see Table 1) and then list the training hyperparameters used in main results, ablation studies and transfer learning, in Table 2.

A. More Comparisons and Analysis

To further check the advantages of the proposed CrossViT, we trained the models whose architecture are identical to the L-branch (primary) of our models. E.g., DeiT-9 is the baseline for CrossViT-9. As shown in Table 1, the proposed cross-attention fusion consistently improves the baseline vision transformers regardless of their primary branches and patch embeddings, suggesting that the proposed multi-scale fusion is effective for different vision transformers.

Figure 1 visualizes the features of both branches from the last multi-scale transformer encoder of CrossViT. The proposed cross-attention learns different features in both branches, where the small branch generates more low-level features because there are only three transformer encoders while the features of the large branch are more abstract. Both branches complement each other and hence the ensemble results are better.

Model	Top-1 Acc. (%)	FLOPs (G)	Params (M)
DeiT-9	72.9	1.4	6.4
CrossViT-9	73.9	1.8	8.6
DeiT-9†	75.6	1.5	6.6
CrossViT-9†	**77.1**	2.0	8.8
DeiT-15	80.8	4.9	22.9
CrossViT-15	81.5	5.8	27.4
DeiT-15†	81.7	5.1	23.5
CrossViT-15†	**82.3**	6.1	28.2
DeiT-18	81.4	7.8	37.1
CrossViT-18	82.5	9.0	43.3
DeiT-18†	81.2	8.1	37.9
CrossViT-18†	**82.8**	9.5	44.3

Table 1: **Comparisons with various baselines on ImageNet1K.** See Table 1 of the main paper for model details. † denotes the models using three convolutional layers for patch embedding instead of linear projection.

	Main Results	Transfer
Batch size	4,096	768
Epochs	300	1,000
Optimizer	AdamW	SGD
Weight Decay	0.05	1e-4
Linear-rate Scheduler (Initial LR)	Cosine (0.004)	Cosine (0.01)
Warmup Epochs	30	5
Warmup linear-rate Scheduler (Initial LR)	Linear (1e-6)	
Data Aug.	RandAugment (m=9, n=2)	
Mixup (α)	0.8	
CutMix (α)	1.0	
Random Erasing	0.25	0.0
Instance Repetition*	3	
Drop-path	0.1	0.0
Label Smoothing	0.1	

*: only used for CrossViT-18.

Table 2: **Details of training settings.**

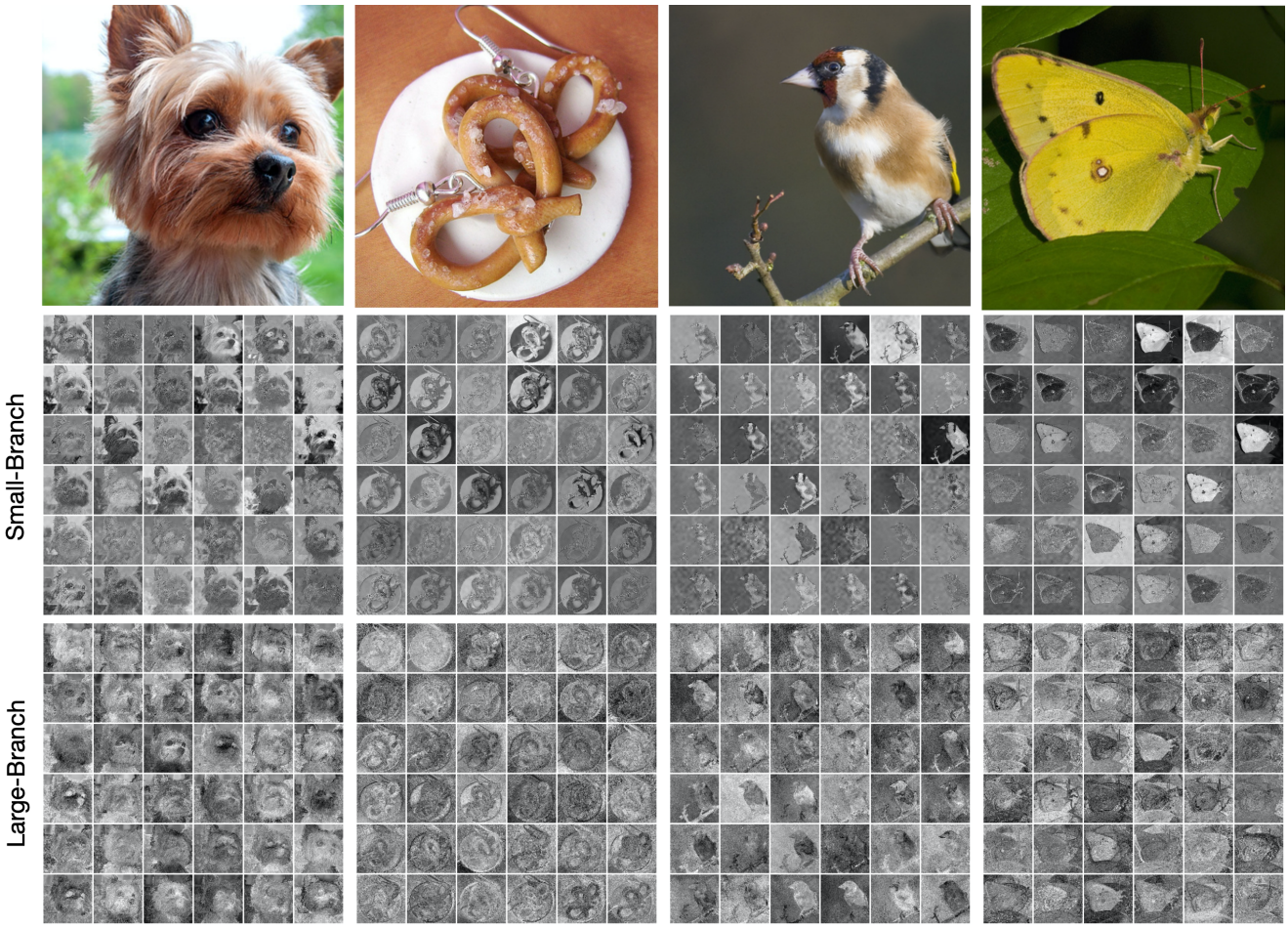

Figure 1: **Feature visualization of CrossViT-S.** Features of patch tokens of both branches from the last multi-scale transformer encoder are shown. (36 random channels are selected.)