InceptionNeXt: When Inception Meets ConvNeXt
– *Supplementary Material* –

Weihao Yu[1] Pan Zhou[2,3] Shuicheng Yan[4] Xinchao Wang[1*]
[1]National University of Singapore [2]Singapore Management University [3]Sea AI Lab [4]Skywork AI
weihaoyu@u.nus.edu panzhou@smu.edu.sg shuicheng.yan@kunlun-inc.com xinchao@nus.edu.sg
Code: https://github.com/sail-sg/inceptionnext

A. Hyper-parameters

A.1. ImageNet-1K image classification

On ImageNet-1K [2, 8] classification benchmark, following ConvNeXt [6] and ConvNeXt-A trained by timm [12], we adopt the hyper-parameters shown in Table 1 to train InceptionNeXt at the input resolution of 224^2 and fine-tune it at 384^2. Our code is implemented by PyTorch [7] based on timm library [12].

A.2. Semantic segmentation

For ADE20K [16] semantic segmentation, we utilize ConvNeXt as the backbone with UpNet [14] following the configs of Swin [5], and FPN [4] following the configs of PVT [11] and PoolFormer [15]. The backbone is initialized by checkpoints pre-trained on ImageNet-1K at the resolution of 224^2. The peak stochastic depth rates of the InceptionNeXt backbone are shown in Table 2. Our implementation is based on PyTorch [7] and mmsegmentation library [1].

B. Qualitative results

Grad-CAM [9] is employed to visualize the activation maps of different models trained on ImageNet-1K, including RSB-ResNet-50 [3, 13], Swin-T [5], ConvNeXt-T [6] and our InceptionNeXt-T. The results are shown in Figure 1. Compared with other models, InceptionNeXt-T locates key parts more accurately with smaller activation areas.

	InceptionNeXt				
	Train				Finetune
	A	T	S	B	B
Input resolution	224^2	224^2			384^2
Epochs	450	300			30
Batch size	1280	4096			1024
Optimizer	AdamW	AdamW			AdamW
Adam ϵ	1e-8	1e-8			1e-8
Adam (β_1, β_2)	(0.9, 0.999)	(0.9, 0.999)			(0.9, 0.999)
Learning rate	1e-3	4e-3			5e-5
Learning rate decay	Cosine	Cosine			Cosine
Gradient clipping	None	None			None
Warmup epochs	5	20			None
Weight decay	0.05	0.05			0.05
Rand Augment	5/uniform	9/0.5			9/0.5
Repeated Augmentation	off	off			off
Cutmix	1.0	1.0			1.0
Mixup	0.2	0.8			0.8
Cutmix-Mixup switch prob	0.5	0.5			0.5
Random erasing prob	0.1	0.25			0.25
Label smoothing	0.1	0.1			0.1
Peak stochastic depth rate	0.1	0.1	0.3	0.4	0.7
Dropout in classifier	0.0	0.0			0.5
LayerScale initialization	1e-6	1e-6			Pre-trained
Random erasing prob	0.1	0.25			0.25
EMA decay rate	None	None			0.9999

Table 1. **Hyper-parameters of InceptionNeXt on ImageNet-1K image classification.**

Method	InceptionNeXt stochastic depth rate		
	T	S	B
UperNet [14]	0.2	0.3	0.4
FPN [4]	0.1	0.2	0.2

Table 2. **Stochasic depth rate of InceptionNeXt backbone with UperNet and FPN for ADE20K semantic segmentation.**

*Corresponding Author.

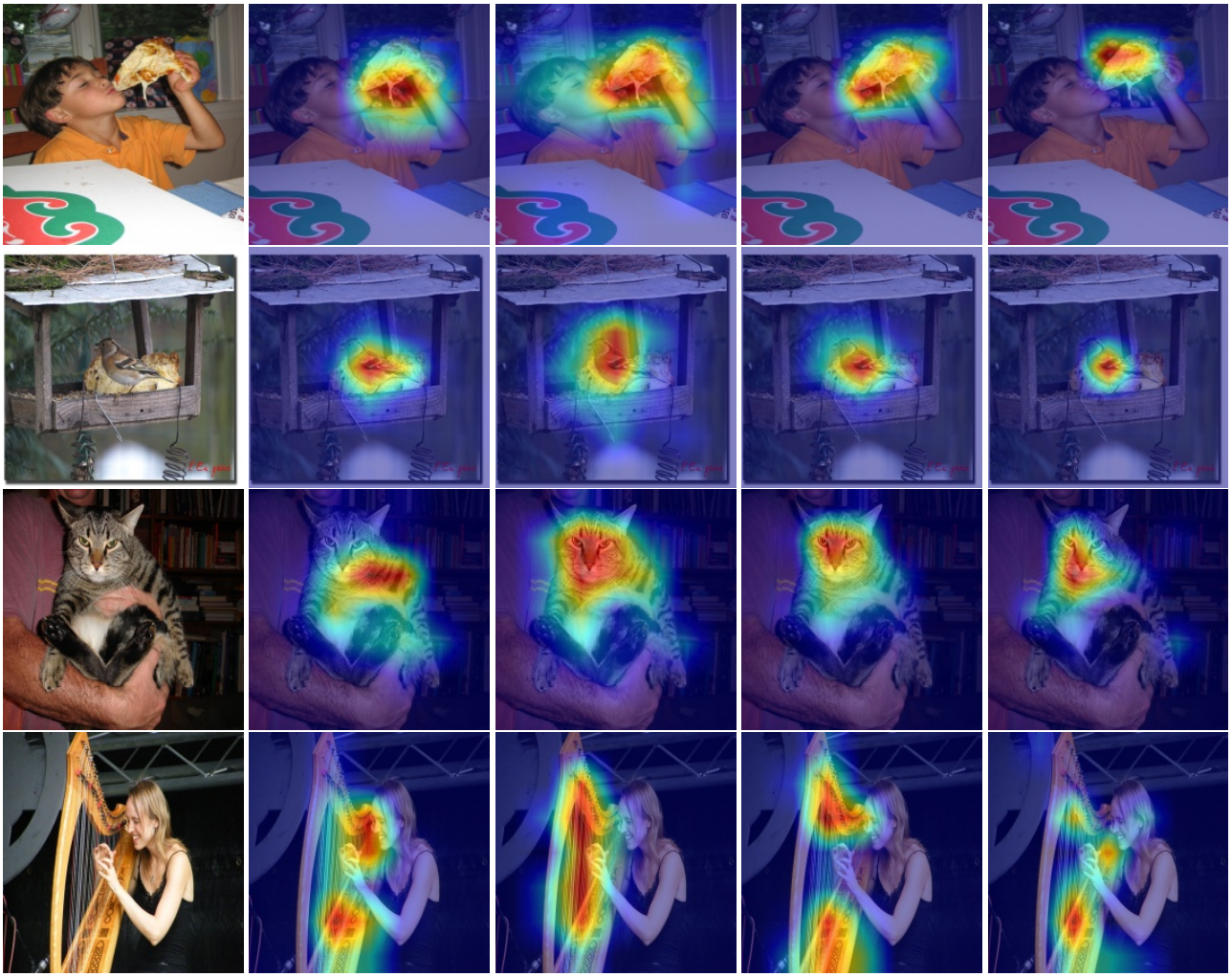

Figure 1. Grad-CAM [9] activation maps of different models trained on ImageNet-1K. The visualized images are from the validation set of ImageNet-1K.

References

[1] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[4] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019. 1

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[6] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1, 2

[7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Ad-*

vances in neural information processing systems, 32, 2019. 1

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1

[9] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 2

[10] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2

[11] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 1

[12] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 1

[13] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021. 1, 2

[14] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 1

[15] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 1

[16] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1