CoAtNet: Marrying Convolution and Attention for All Data Sizes

Zihang Dai, Hanxiao Liu, Quoc V. Le, Mingxing Tan

Google Research, Brain Team

Advances in neural information processing systems, 2021.

경영과학연구실 김윤석

Introduction

- ViT (Vision Transformer) showed good performance with almost only vanilla Transformer layers.
- On the JFT-300M dataset, ViT outperforms ConvNets.
- Large datasets are pre-trained, surpassing the performance of ConvNets.
- ConvNets show limitations in capturing global contexts, whereas ViT shows strength in this regard.
- Unlike traditional Transformer models, ConvNets capture local contexts more effectively.
- Transformers and ConvNets have distinct strengths and limitations when capturing global and local contexts respectively.

Related works

Paper	Key idea
BELLO, Irwan, et al. Attention augmented convolutional networks. 2019	Connecting convolution feature maps to perform relative self-attention operations
SRINIVAS, Aravind, et al. Bottleneck transformers for visual recognition. 2021	Replacing the last three blocks of ResNet with self-attention
VASWANI, Ashish, et al. Scaling local self-attention for parameter efficient visual backbones. 2021	Applying the filter operation characteristics of ConvNets to attention operations
LIU, Ze, et al. Swin transformer: Hierarchical vision transformer using shifted windows. 2021	Introducing window self-attention and shifted window self-attention to perform attention operations hierarchically, similar to CNN

Problem statements

- This paper investigates efficient integration to achieve a trade-off between convolution layers and attention layers.

- The trade-off between generalization ability and model capacity.
- Effective fusion of local pattern recognition and global pattern recognition.
- Full integration of different layers.

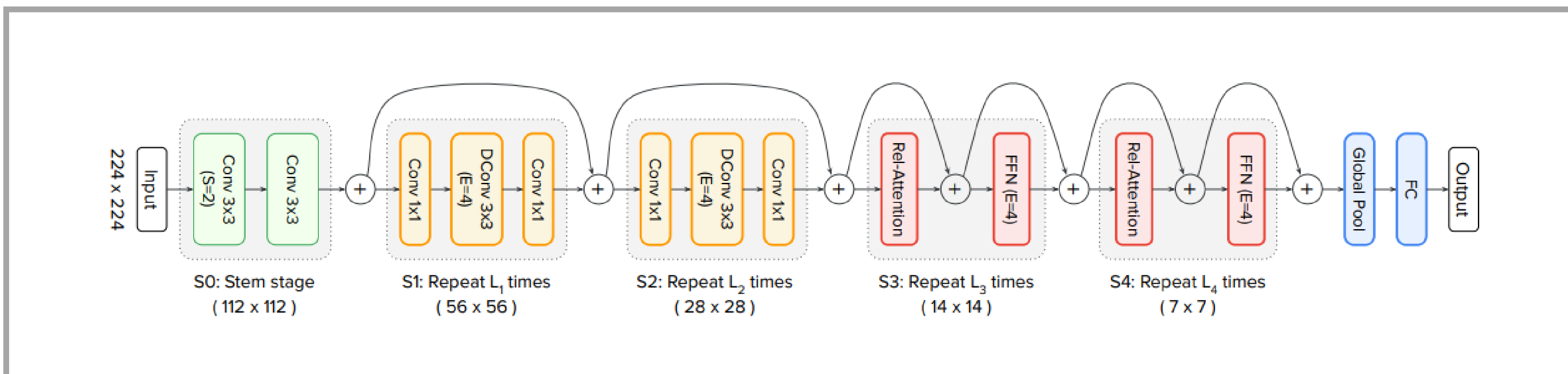
Key idea

- The key idea is the combination of MBConv and Attention-FFN.

  - MBConv and Attention-FFN share structural similarity by using inverted bottlenecks.
  - Both Depthwise Convolution and self-attention can be expressed as weighted average calculations on defined inputs.
  - The trade-off between generalization ability and model capacity is determined through comparative experiments based on model configurations.

Overall architecture

- The Stem stage is responsible for transforming the input image into a lower-dimensional feature map.
- Stage 1 and Stage 2 serve the purpose of reducing dimensionality and increasing channels through MBConv blocks.
- Stage 3 and Stage 4 perform relative attention (rel-attention) on feature maps, utilizing pooling and FFN to generate lower-dimensional feature maps.
- Resolution is reduced by half in all stages.



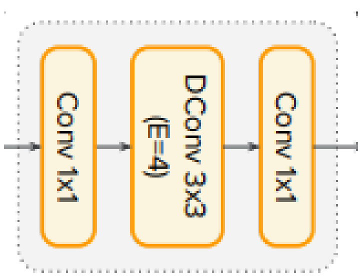
MBConv block

- MBConv is an idea that originated from MobileNetV2 and is widely used in recent technologies like EfficientNet.
- Channel expansion in MBConv enables the learning of various features.
- MBConv streamlines the model through the use of pointwise convolutions and depthwise convolutions.
- To achieve efficient integration in CoAtNet, ConvNet uses MBConv blocks.

- MBConv block은 Conv 1x1에서 채널이 확장 되고 DConv 연산 후 Conv 1x1에서 채널 수를 되돌림

Conv 1x1(pointwise conv): 1x1xC 필터를 사용함
DConv 3x3(depthwise conv): 3x3x1가 C개 있음

C: Input feature map's channel



MBConv block

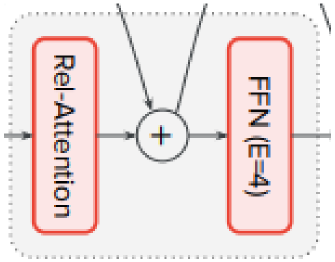
Rel-Attention block

- The Rel-Attention block performs relative attention operations on reduced lower-dimensional feature maps.
- Attention allows the model to learn global information.
- Finally, the resolution is reduced and the channels are increased through a pooling layer.

- Rel-Attention performs rel-attention operations on each pixel vector of the feature map.

Rel-Attention: $Attention(Q,K,V,R) = softmax\left(\frac{QK^T+QR^T}{\sqrt{d_k}}\right)V$

FFN: It operates by increasing and then reducing the feature map by a factor of 4.



Rel-Attention block

CoAtNet model family

- The authors propose a total of 5 CoAtNet model architectures.
- The models are categorized based on their depth and feasible input resolutions.

Stages	Size	CoAtNet-0		CoAtNet-1		CoAtNet-2		CoAtNet-3		CoAtNet-4	
S0-Conv	$1/2$	L=2	D=64	L=2	D=64	L=2	D=128	L=2	D=192	L=2	D=192
S1-MbConv	$1/4$	L=2	D=96	L=2	D=96	L=2	D=128	L=2	D=192	L=2	D=192
S2-MBConv	$1/8$	L=3	D=192	L=6	D=192	L=6	D=256	L=6	D=384	L=12	D=384
S3-$\mathrm{TFM_{Rel}}$	$1/16$	L=5	D=384	L=14	D=384	L=14	D=512	L=14	D=768	L=28	D=768
S4-$\mathrm{TFM_{Rel}}$	$1/32$	L=2	D=768	L=2	D=768	L=2	D=1024	L=2	D=1536	L=2	D=1536

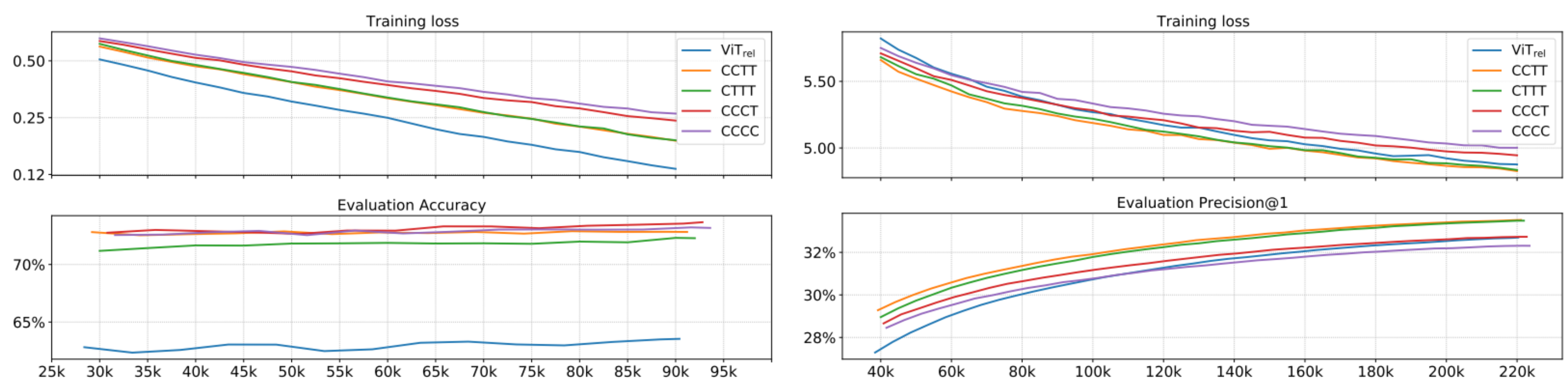
Experiments

- Experiments focus on image classification

- Experiments on model capacity and generalization ability
- Comprehensive Performance Evaluation on ImageNet-1k
- Experiments on Pre-training Performance
- Experiments on CoAtNet Configurations

Experiments on model capacity and generalization ability

- (a) is an experiment comparing performance on ImageNet-1k without pre-training.
- In (a), ViT shows lower generalization performance.
- (b) is an experiment comparing performance on JFT-300M.
- C-C-T-T and C-T-T-T show good performance.



(a) ImageNet-1K

(b) JFT

Performance Evaluation on ImageNet-1k

- CoAtNet outperforms models with similar parameter counts.

Models		Eval Size	#Params	#FLOPs	ImageNet Top-1 Accuracy	
					1K only	21K+1K
Conv Only	EfficientNet-B7	600^2	66M	37B	84.7	-
	EfficientNetV2-L	480^2	121M	53B	85.7	86.8
	NFNet-F3	416^2	255M	114.8B	85.7	-
	NFNet-F5	544^2	377M	289.8B	**86.0**	-
ViT-Stem TFM	DeiT-B	384^2	86M	55.4B	83.1	-
	ViT-L/16	384^2	304M	190.7B	-	85.3
	CaiT-S-36	384^2	68M	48.0B	85.0	-
	DeepViT-L	224^2	55M	12.5B	83.1	-
Multi-stage TFM	Swin-B	384^2	88M	47.0B	84.2	86.0
	Swin-L	384^2	197M	103.9B	-	86.4
Conv+TFM	BotNet-T7	384^2	75.1M	45.8B	84.7	-
	LambdaResNet-420	320^2	-	-	84.8	-
	T2T-ViT-24	224^2	64.1M	15.0B	82.6	-
	CvT-21	384^2	32M	24.9B	83.3	-
	CvT-W24	384^2	277M	193.2B	-	**87.7**
Conv+TFM (ours)	CoAtNet-0	224^2	25M	4.2B	81.6	-
	CoAtNet-1	224^2	42M	8.4B	83.3	-
	CoAtNet-2	224^2	75M	15.7B	84.1	87.1
	CoAtNet-3	224^2	168M	34.7B	84.5	87.6
	CoAtNet-0	384^2	25M	13.4B	83.9	-
	CoAtNet-1	384^2	42M	27.4B	85.1	-
	CoAtNet-2	384^2	75M	49.8B	85.7	87.1
	CoAtNet-3	384^2	168M	107.4B	85.8	87.6
	CoAtNet-4	384^2	275M	189.5B	-	87.9
	+ PT-RA	384^2	275M	189.5B	-	88.3
	+ PT-RA-E150	384^2	275M	189.5B	-	88.4
	CoAtNet-2	512^2	75M	96.7B	85.9	87.3
	CoAtNet-3	512^2	168M	203.1B	**86.0**	87.9
	CoAtNet-4	512^2	275M	360.9B	-	88.1
	+ PT-RA	512^2	275M	360.9B	-	88.4
	+ PT-RA-E150	512^2	275M	360.9B	-	**88.56**

Experiments on pre-training performance

- CoAtNet achieves better performance with fewer parameters than ConvNets and ViT pre-trained on a large dataset (JFT).

Models	Eval Size	#Params	#FLOPs	TPUv3-core-days	Top-1 Accuracy
ResNet + ViT-L/16	384^2	330M	-	-	87.12
ViT-L/16	512^2	307M	364B	0.68K	87.76
ViT-H/14	518^2	632M	1021B	2.5K	88.55
NFNet-F4+	512^2	527M	367B	1.86K	89.2
CoAtNet-3†	384^2	168M	114B	0.58K	88.52
CoAtNet-3†	512^2	168M	214B	0.58K	88.81
CoAtNet-4	512^2	275M	361B	0.95K	89.11
CoAtNet-5	512^2	688M	812B	1.82K	89.77
ViT-G/14	518^2	1.84B	5160B	>30K$^\diamond$	90.45
CoAtNet-6	512^2	1.47B	1521B	6.6K	90.45
CoAtNet-7	512^2	2.44B	2586B	20.1K	**90.88**

Experiments on CoAtNet configuration

- In Table 6, CoAtNet with Rel-Attn shows approximately a 0.4% improvement in performance.
- In Table 7, the V0 layout exhibits the best performance.

Table 6: Ablation on relative attention.

Setting	Metric	With Rel-Attn	Without Rel-Attn
ImageNet-1K	Accuracy (224^2)	84.1	83.8
	Accuracy (384^2)	85.7	85.3
ImageNet-21K	Pre-train Precision@1 (224^2)	53.0	52.8
$\Rightarrow$ ImageNet-1K	Finetune Accuracy (384^2)	87.9	87.4

Table 7: Ablation on architecture layout.

Setting	Models	Layout	Top-1 Accuracy
ImageNet-1K	V0: CoAtNet-2	[2, 2, 6, 14, 2]	84.1
	V1: S2 $\Leftarrow$ S3	[2, 2, 2, 18, 2]	83.4
	V2: S2 $\Rightarrow$ S3	[2, 2, 8, 12, 2]	84.0
ImageNet-21K	V0: CoAtNet-3	[2, 2, 6, 14, 2]	53.0 $\rightarrow$ 87.6
$\Rightarrow$ ImageNet-1K	V1: S2 $\Leftarrow$ S3	[2, 2, 2, 18, 2]	53.0 $\rightarrow$ 87.4

Conclusions

- CoAtNet explores the efficient combination of ConvNets and Transformers.
- CoAtNet is a model that combines the strong generalization ability of ConvNets with the excellent model capacity of Transformers.
- One limitation of the paper is that it only compares results on the image classification task.
- The authors plan to conduct further research on the various applications of CoAtNet across different tasks.