Early Convolutions Help Transformers See Better

TeteXiao, Mannat Singh, EricMintun, Trevor Darrell, Piotr Dollár, Ross Girshick

Argument #1

Technical Novelty

- Builds on the ViT and proposes simple change to stem
- Opens up research question regarding impacts of small convolutions in early ViT training
- Theoretical question of “what does a computer need to understand”
    - Can paint a picture of what the computer actually sees, or what helps it see

Argument #2

Strong empirical results

- Very intricate data and testing across multiple types of stability
  - Proposes architecture changes that make SGD a viable optimizer for ViT
  - Is able to reach competitive or better stability with AdamW optimizer
  - Convergence of ViTs can is brought closer to SOTA CNN
  - Is able to outperform SOTA CNN and ViT when smaller scale pre training data is available (ImageNet-21k)

Argument #3

Ease of implementation, easy to read

- Simple change to the patch encoding
  - Old: a stride-p p by p convolution for patch encoding
  - New: change the stride value of the convolution
- Paper is very easy to follow through
  - Clear definitions for unclear terms (ViT C, ViT P)
  - Metrics for stability clearly detailed (ex. Training length, optimizer, and hyperparameter stability)
  - Solid, cleanly summarized conclusion

ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases

Sabiq Muhtadi, Charlie Arleth and Cheng Che Tsai (Michael)

3 Arguments

1. Technical Contribution

ConViT identifies a limitation of the standard ViT and proposes a very intricate and effective solution to bring together the best of both CNN's and ViT's .

- GPSA provides the ConViT the freedom to decide whether it will behave as a convolutional layer, or as a self-attention layer.
- GPSA blocks provide more freedom to modify the ViT architecture as needed.
- Performance vs. optimization (technical contribution vs. engineering paper).

2. Results - Charlie

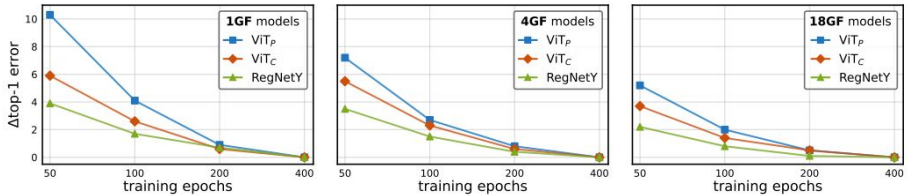
ConViT has a significant improvement over DeiT, especially at higher numbers of parameters.

S	DeiT	6	384	22M	4.3G	587	79.8	-
	ConViT	9	432	27M	5.4G	305	**81.3**	**95.7**
S+	DeiT	9	576	48M	10G	480	79.0	94.4
	ConViT	9	576	48M	10G	382	**82.2**	**95.9**
B	DeiT	12	768	86M	17G	187	81.8	-
	ConViT	16	768	86M	17G	141	**82.4**	**95.9**
B+	DeiT	16	1024	152M	30G	114	77.5	93.5
	ConViT	16	1024	152M	30G	96	**82.5**	**95.9**

One of ViTC's main benefits is that it does better training on less epoch's but ConViT also does better training on less data

model	flops (B)	params (M)	acts (M)	time (min)	batch size	epochs 100	200	400	IN 21k
ResNet-50	4.1	25.6	11.3	3.4	2048	22.5	21.2	20.7	21.6
ResNet-101	7.8	44.5	16.4	5.5	2048	20.3	19.1	18.5	19.2
ResNet-152	11.5	60.2	22.8	7.7	2048	19.5	18.4	17.7	18.2
ResNet-200	15.0	64.7	32.3	10.7	1024	19.5	18.3	17.6	17.7
RegNetY-1GF	1.0	9.6	6.2	3.1	2048	23.2	22.2	21.5	-
RegNetY-4GF	4.1	22.4	14.5	7.6	2048	19.4	18.3	17.9	18.4
RegNetY-16GF	15.5	72.3	30.7	17.9	1024	17.1	16.4	16.3	15.6
RegNetY-32GF	31.1	128.6	46.2	35.1	512	16.2	15.9	15.9	15.0
RegNetZ-1GF	1.0	11.0	8.8	4.2	2048	20.8	20.2	19.6	-
RegNetZ-4GF	4.0	28.1	24.3	12.9	1024	17.4	16.9	16.6	-
RegNetZ-16GF	16.0	95.3	51.3	32.0	512	16.0	15.9	15.9	-
RegNetZ-32GF	32.0	175.1	79.6	55.3	256	16.3	16.2	16.1	-

model	flops (B)	params (M)	acts (M)	time (min)	batch size	epochs 100	200	400	IN 21k
EffNet-B2	1.0	9.1	13.8	5.9	2048	21.4	20.5	19.9	-
EffNet-B4	4.4	19.3	49.5	19.4	512	18.5	17.8	17.5	-
EffNet-B5	10.3	30.4	98.9	41.7	256	17.3	17.0	17.0	-
ViT$_P$-1GF	1.1	4.8	5.5	2.6	2048	33.2	29.7	27.7	-
ViT$_P$-4GF	3.9	18.5	11.1	3.8	2048	23.3	20.8	19.6	20.6
ViT$_P$-18GF	17.5	86.6	24.0	11.5	1024	19.9	18.4	17.9	16.4
ViT$_P$-36GF	35.9	178.4	37.3	18.8	512	19.9	18.8	18.2	15.1
ViT$_C$-1GF	1.1	4.6	5.7	2.7	2048	28.6	26.1	24.7	-
ViT$_C$-4GF	4.0	17.8	11.3	3.9	2048	20.9	19.2	18.6	18.8
ViT$_C$-18GF	17.7	81.6	24.1	11.4	1024	18.4	17.5	17.0	15.1
ViT$_C$-36GF	35.0	167.8	36.7	18.6	512	18.3	17.6	16.8	14.2



3. ConViT is more flexible and scalable

a. Their hard-coded convolution layers are not easily parallelized, and scalable.

b. Side evidence: why using ⅗ ViT-L instead of full ViT-L?

model	ref model	hidden size	MLP mult	num heads	num blocks	flops (B)	params (M)	acts (M)	time (min)
ViT$_P$-1GF	~ViT-T	192	3	3	12	1.1	4.8	5.5	2.6
ViT$_P$-4GF	~ViT-S	384	3	6	12	3.9	18.5	11.1	3.8
ViT$_P$-18GF	=ViT-B	768	4	12	12	17.5	86.7	24.0	11.5
ViT$_P$-36GF	⅗ ViT-L	1024	4	16	14	35.9	178.4	37.3	18.8

model	hidden size	MLP mult	num heads	num blocks	flops (B)	params (M)	acts (M)	time (min)
ViT$_C$-1GF	192	3	3	11	1.1	4.6	5.7	2.7
ViT$_C$-4GF	384	3	6	11	4.0	17.8	11.3	3.9
ViT$_C$-18GF	768	4	12	11	17.7	81.6	24.1	11.4
ViT$_C$-36GF	1024	4	16	13	35.0	167.8	36.7	18.6