A ConvNet for the 2020s

Facebook AI Research & UC Berkeley
CVPR 2022

Presented by: Bibek Poudel

Outline

- Context & Motivation
- Contributions and Results

Context & Motivation

Context & Motivation

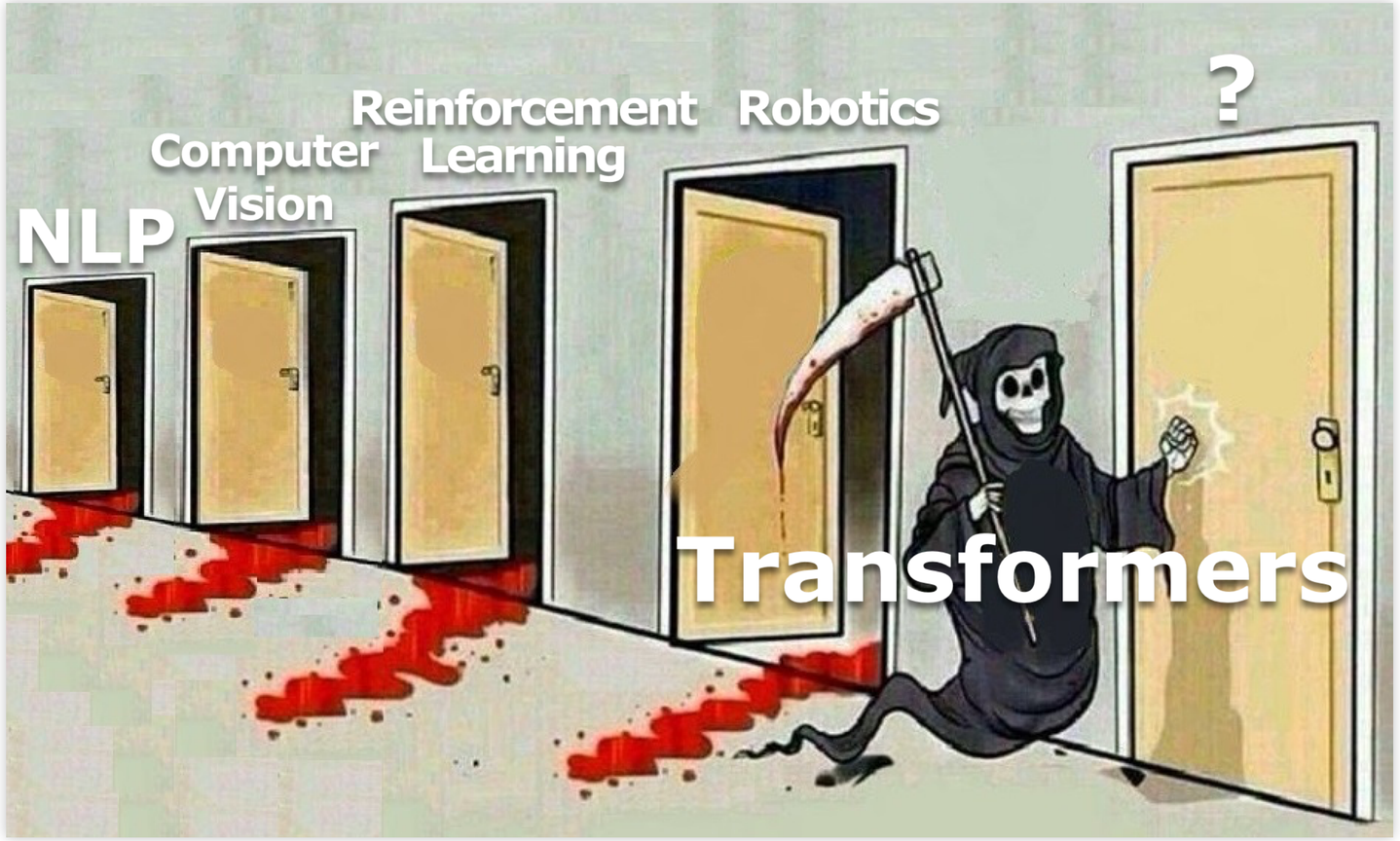


Fig. 1: Transformers are coming for your area of study

Context & Motivation

- Claude 3 release

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	**86.8%** 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	**50.4%** 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	**95.0%** 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	**60.1%** 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	**90.7%** 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	**84.9%** 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	**83.1** 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	**86.8%** 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	**96.4%** 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge *HellaSwag*	**95.4%** 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

Fig. 2: Claude 3 leads “general intelligence” benchmarks

https://www.anthropic.com/news/claude-3-family

Context & Motivation

Computation at Scale > Algorithmic Breakthrough

Context & Motivation

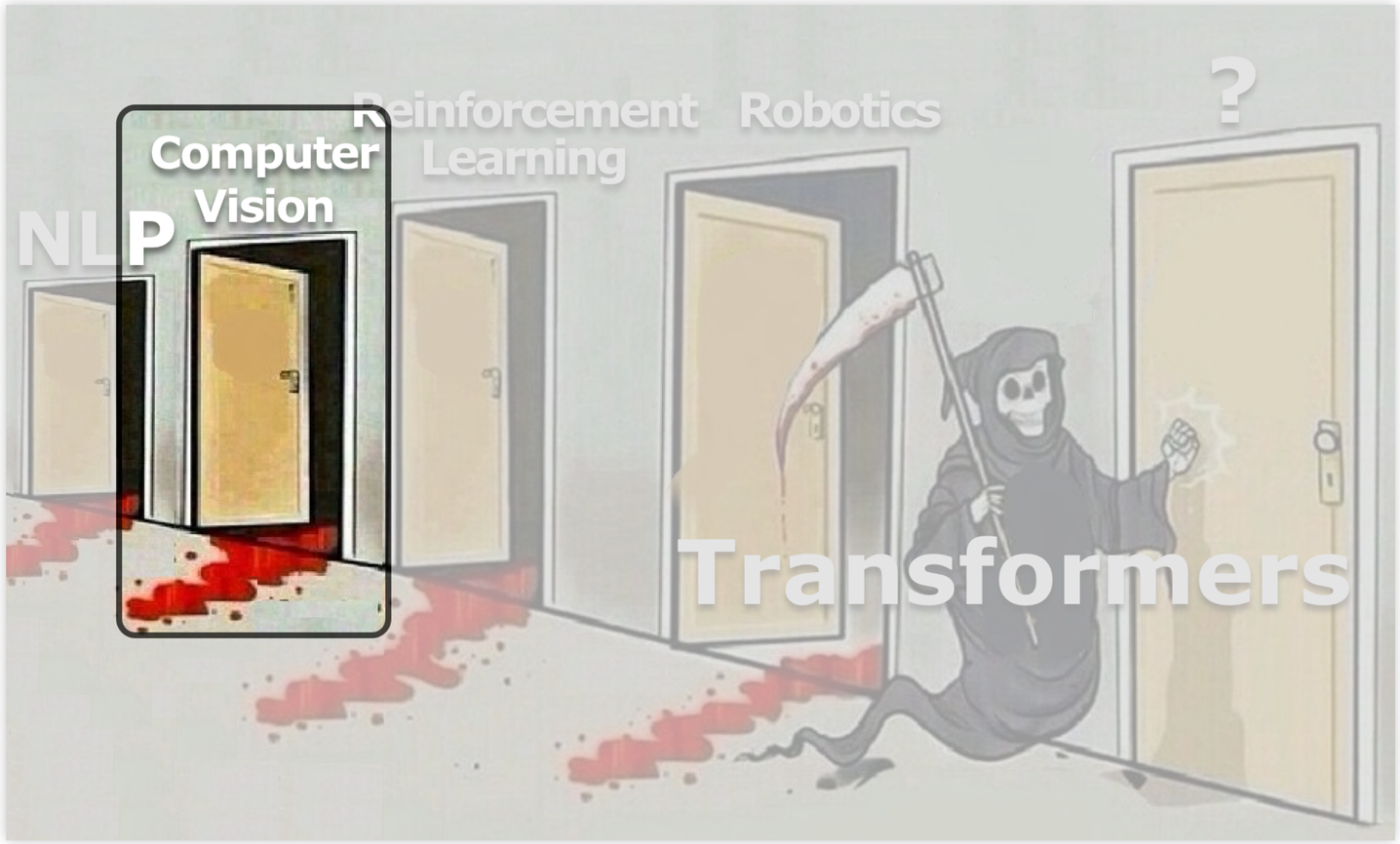


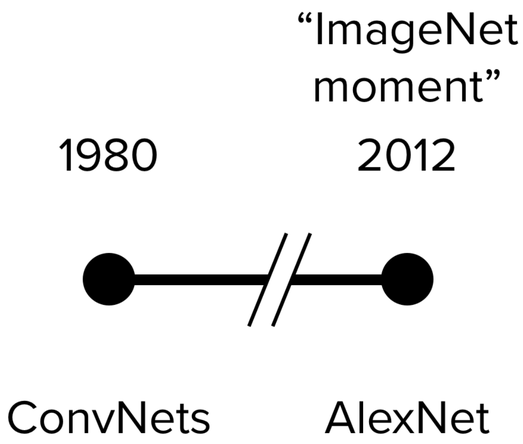
Fig. 3: Transformers in Computer Vision

Context & Motivation

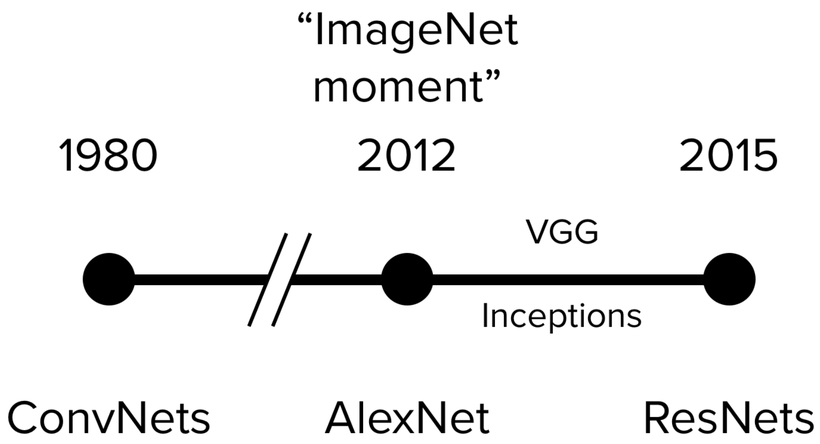
1980

ConvNets

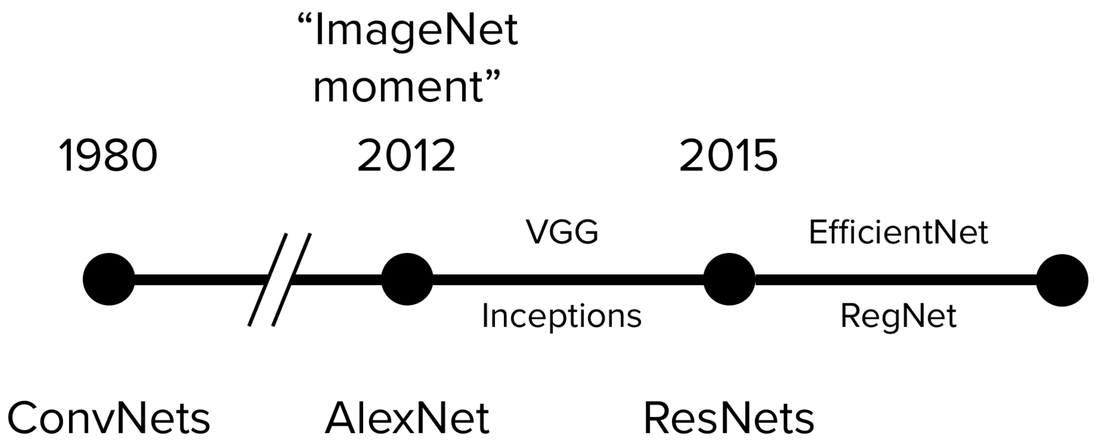
Context & Motivation



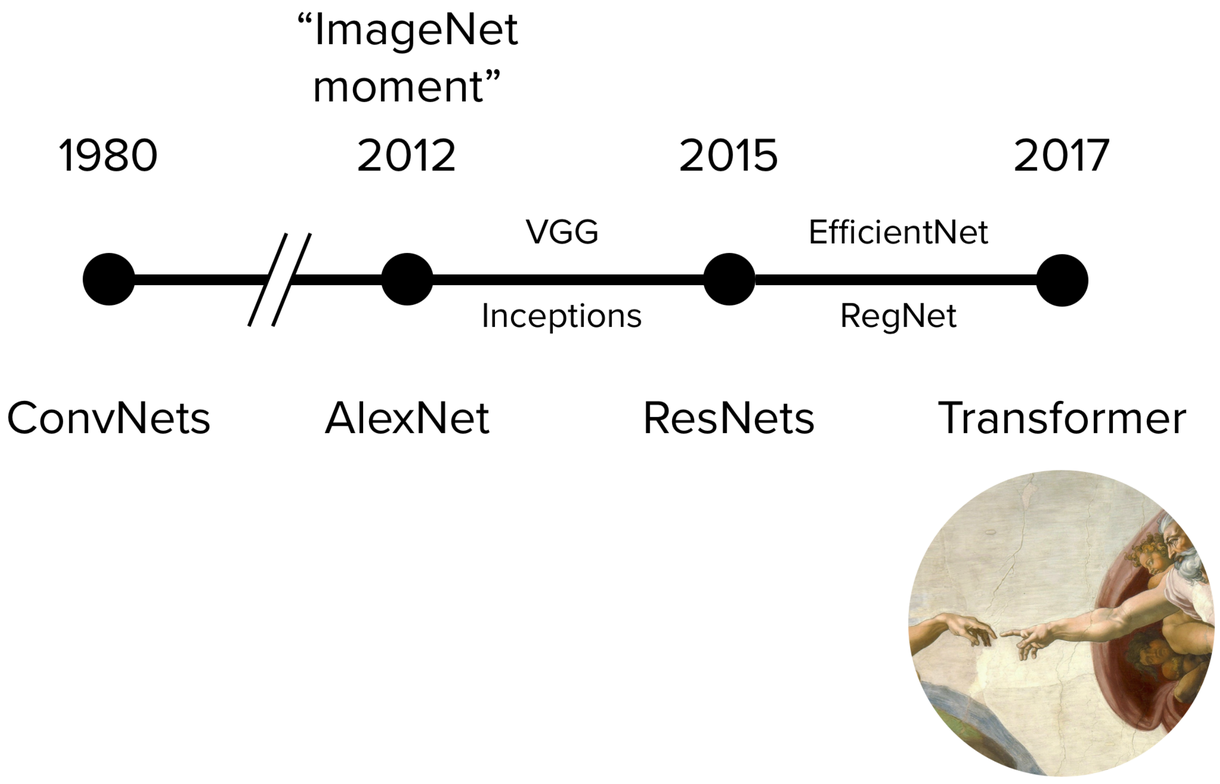
Context & Motivation



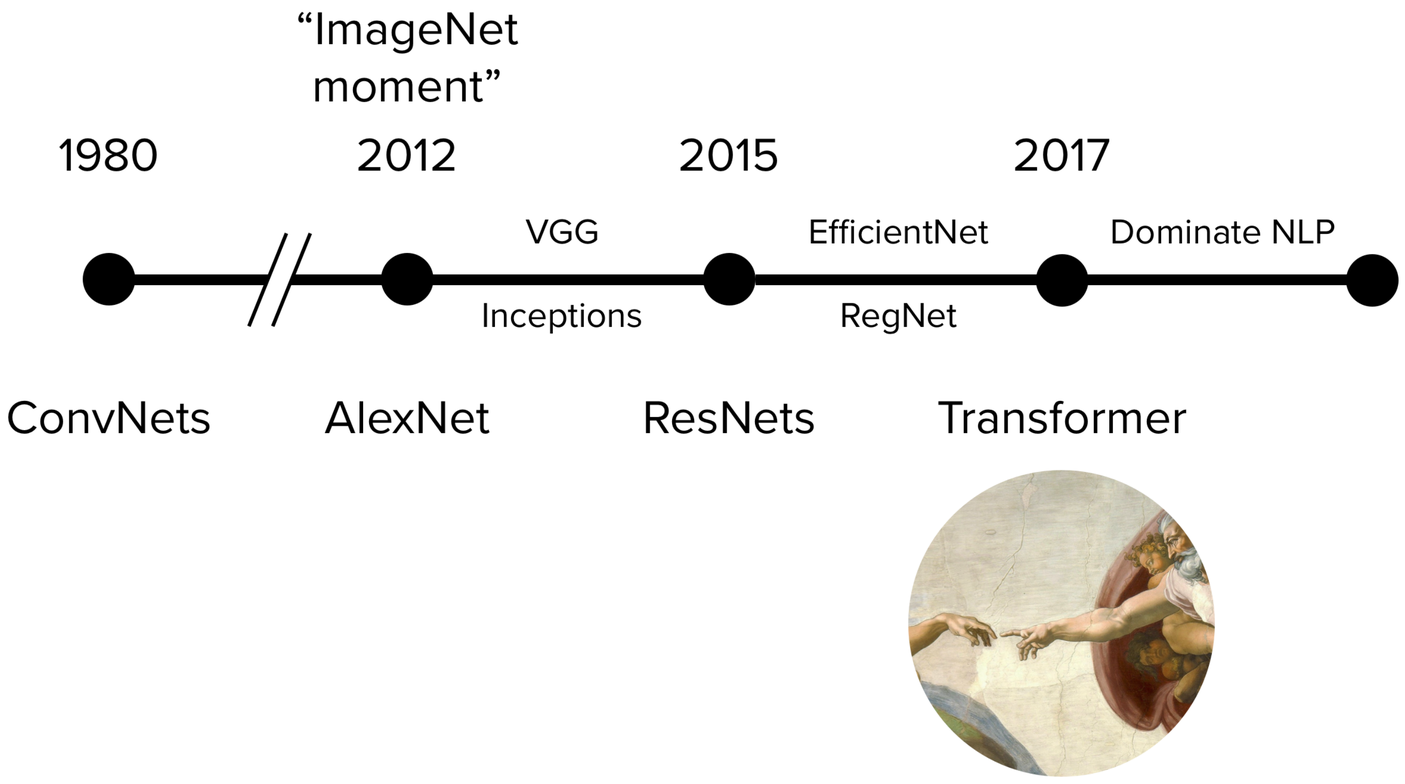
Context & Motivation



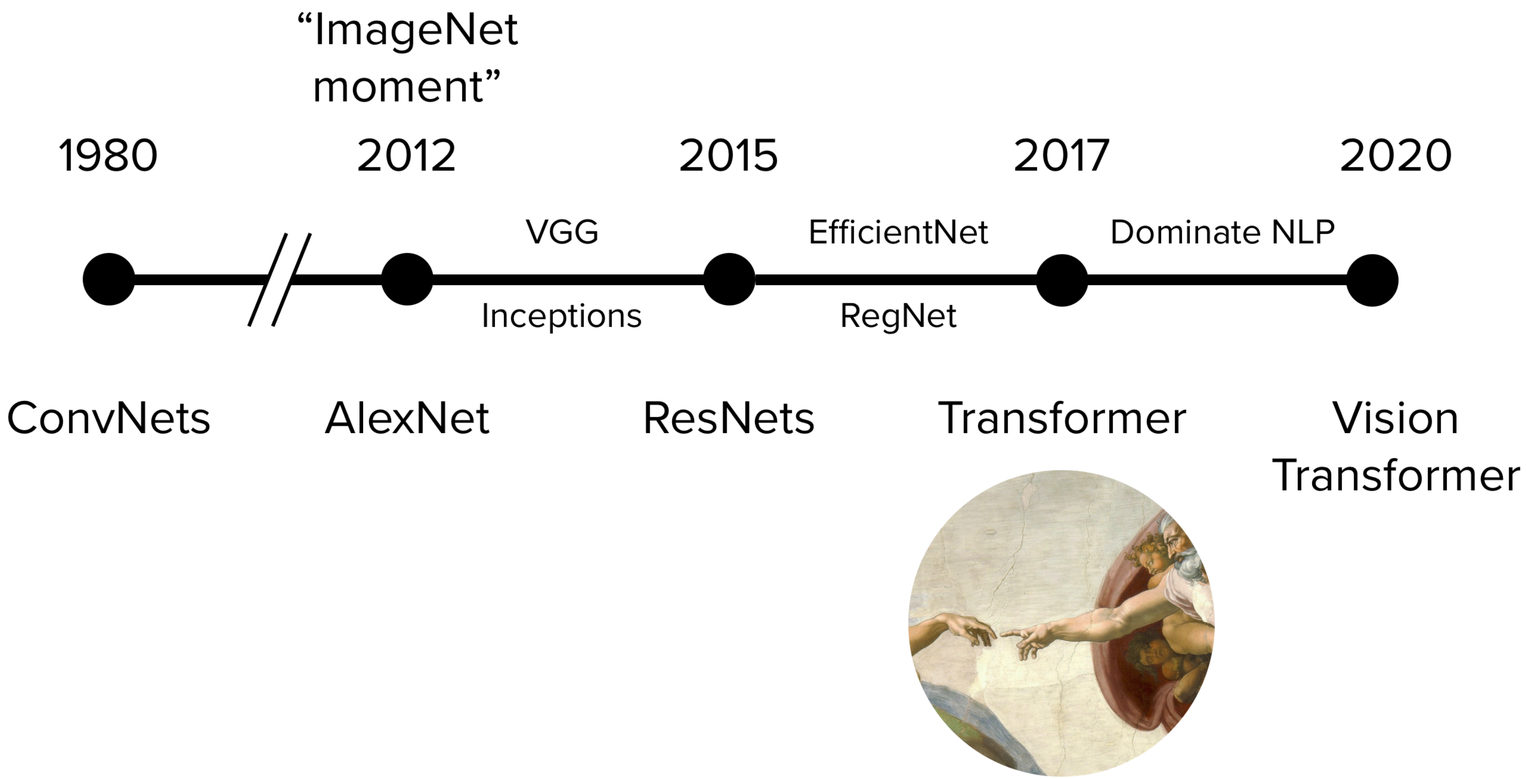
Context & Motivation



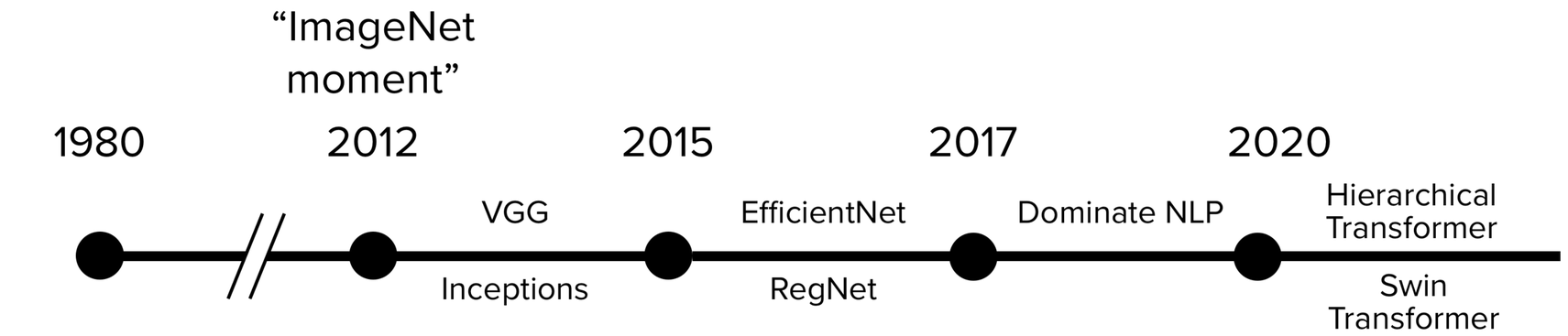
Context & Motivation



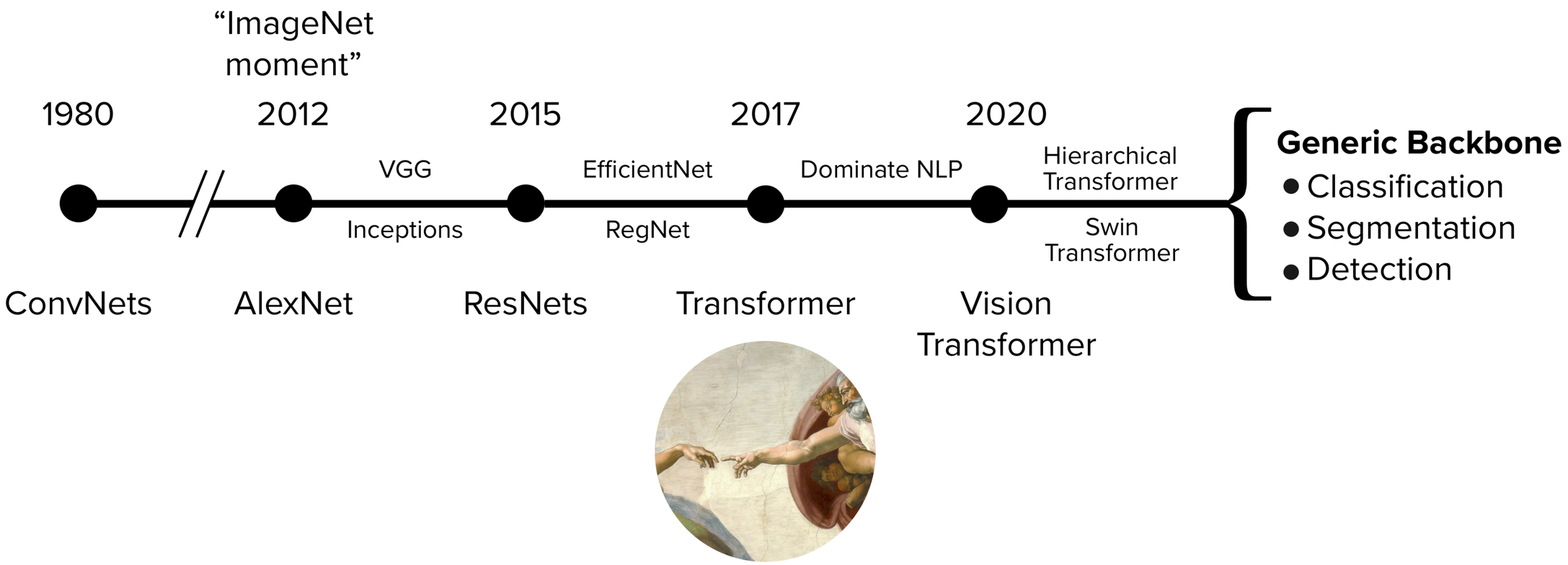
Context & Motivation



Context & Motivation



Context & Motivation



Context & Motivation

- Successful because of the “Transformer” part:
  - Scaling?
  - Multi-head attention?

Context & Motivation

- Successful because of the “Transformer” part:
  - Scaling?
  - Multi-head attention?

But..

- Swin-T introduce “ConvNet priors”:

Context & Motivation

- Successful because of the “Transformer” part:
  - Scaling?
  - Multi-head attention?

But..

- Swin-T introduce “ConvNet priors”:
  - Hierarchical feature maps

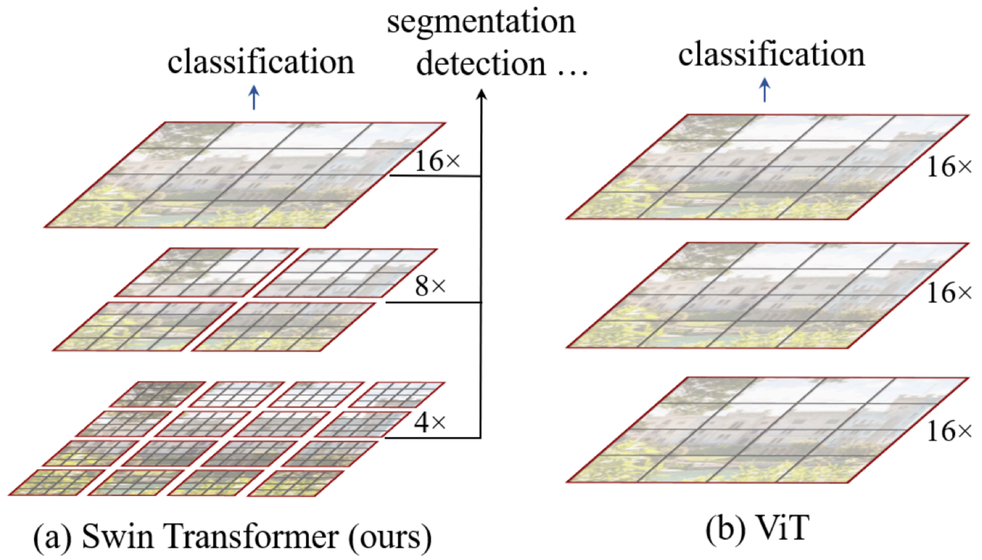


Fig. 4: Hierarchical feature maps in Swin Transformer

“Swin Transformer: Hierarchical Vision Transformer using Shifted Windows” (https://arxiv.org/pdf/2103.14030.pdf)

Context & Motivation

- Successful because of the “Transformer” part:
  - Scaling?
  - Multi-head attention?

But..

- Swin-T introduce “ConvNet priors”:
  - Hierarchical feature maps
  - Locality

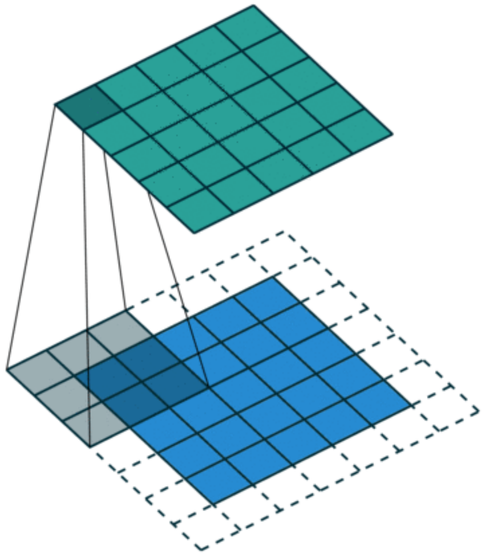


Fig. 5: Sliding window in CNN

Is attention really “all you need”?

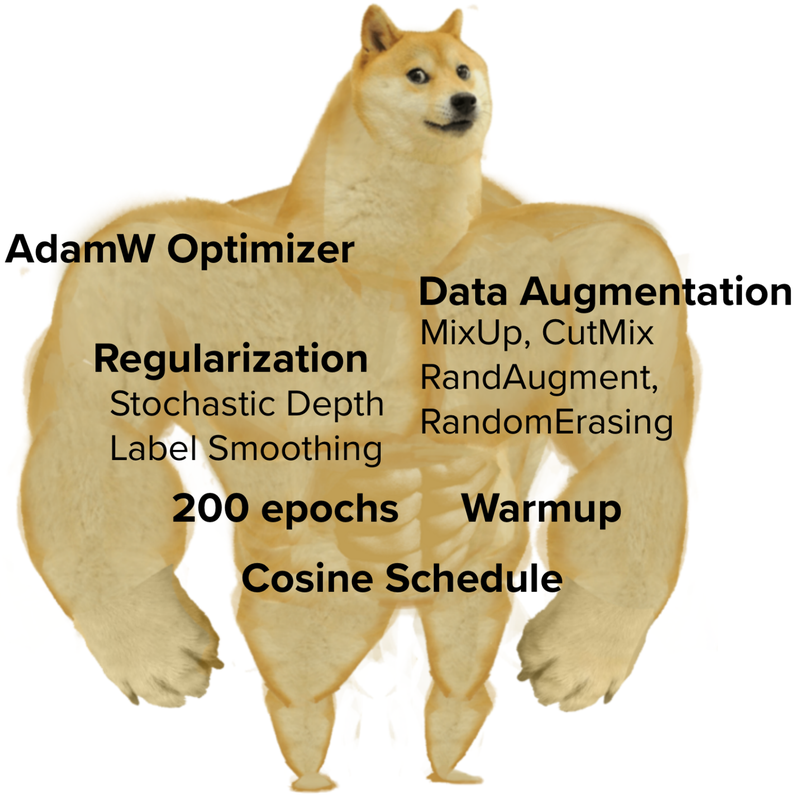
Contributions & Results

Contributions & Results

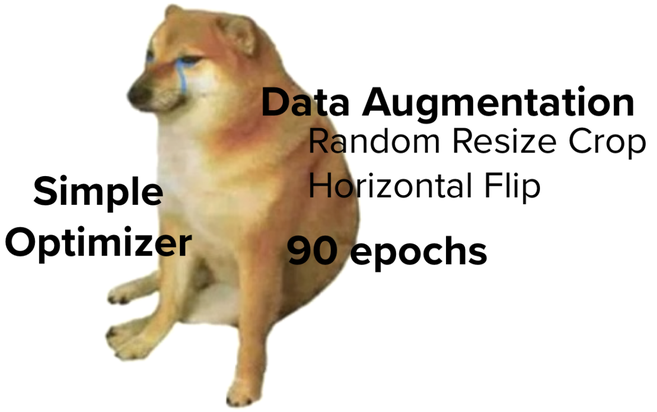
- “Modernize” pure convolution models (ResNet-50 and ResNet-200)

Contributions & Results

- Step 1: Training recipe



Vision Transformer
Training



ResNet 50
Training

Contributions & Results

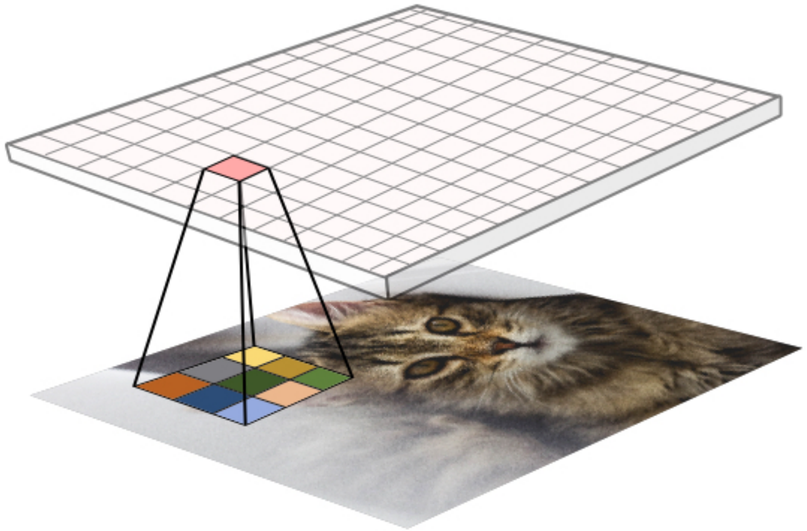
- Step 2: Macro Design Choices
    - Compute Ratio (3, 4, 6, 3) to (3, 3, 9, 3)
    - “Patchify” i.e., non-overlapping convolutions

Contributions & Results

- Step 3: ResNeXt-ify
- Step 4: Inverted Bottleneck

Contributions & Results

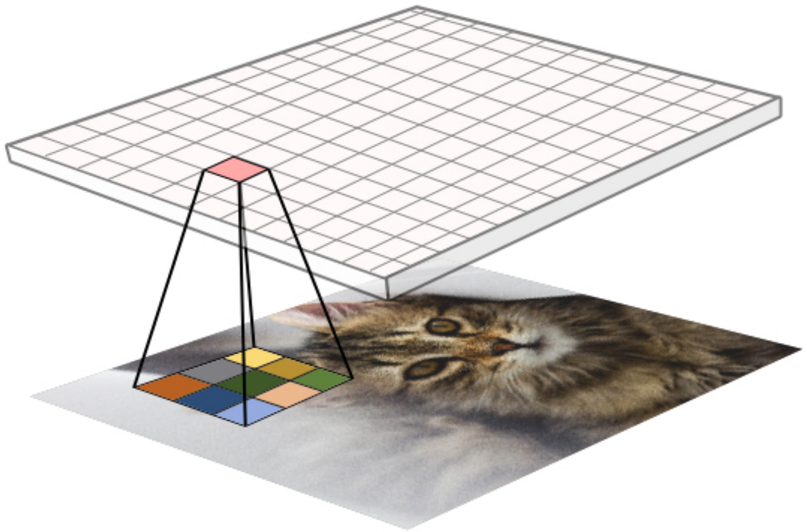
- Step 5: Larger Kernel Size



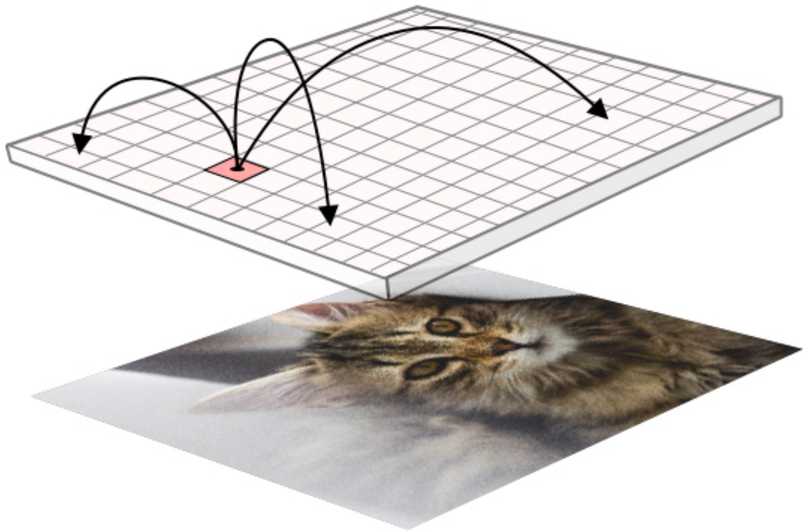
Convolutional
Neural Network

Contributions & Results

- Step 5: Larger Kernel Size
- Global receptive filed



Convolutional
Neural Network



Vision
Transformer

Contributions & Results

- Step 6: Micro Design Choices
    - Change ReLU to GELU
    - Fewer activations

Contributions & Results

- Step 6: Micro Design Choices
    - Change ReLU to GELU
    - Fewer activations
    - Fewer Norm Layers
    - Change Batch Norn to Layer Norm
    - Separate Downsampling layers

Contributions & Results

- Accuracy in Task: Classification, Segmentation, Detection

Contributions & Results

- Accuracy: Classification

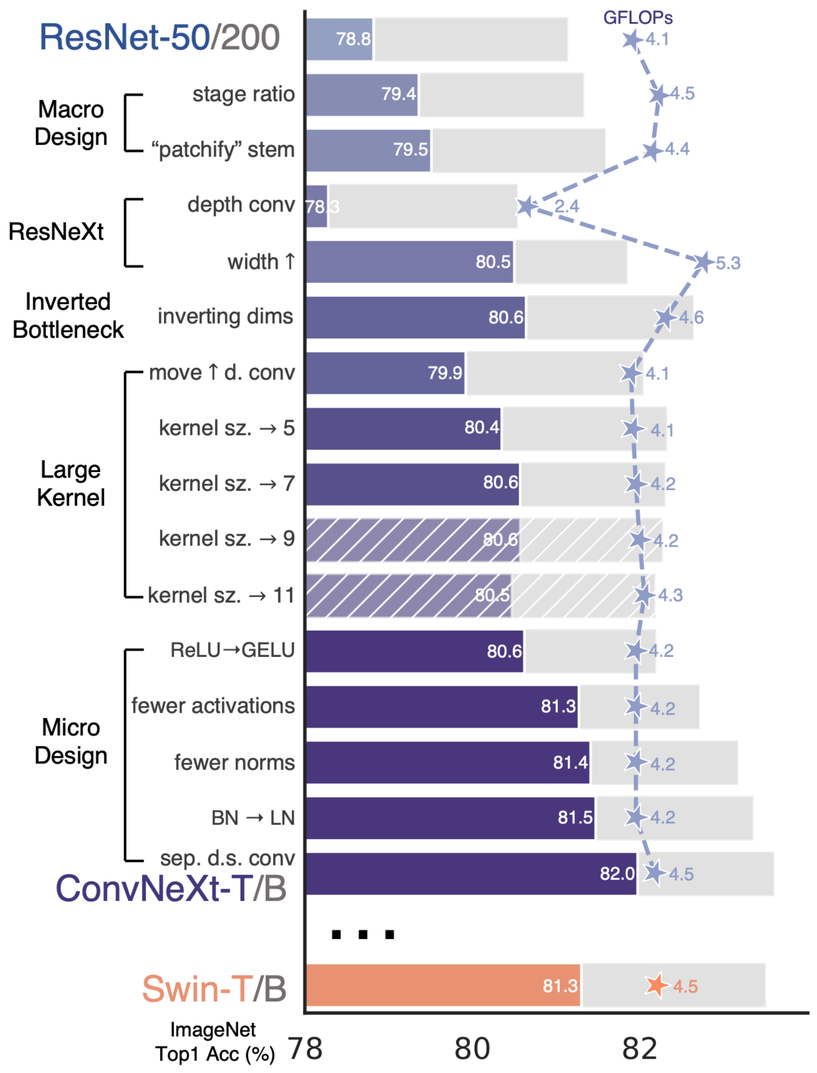


Fig. 6: ConvNext vs. Swin-T

Contributions & Results

- Accuracy: Classification

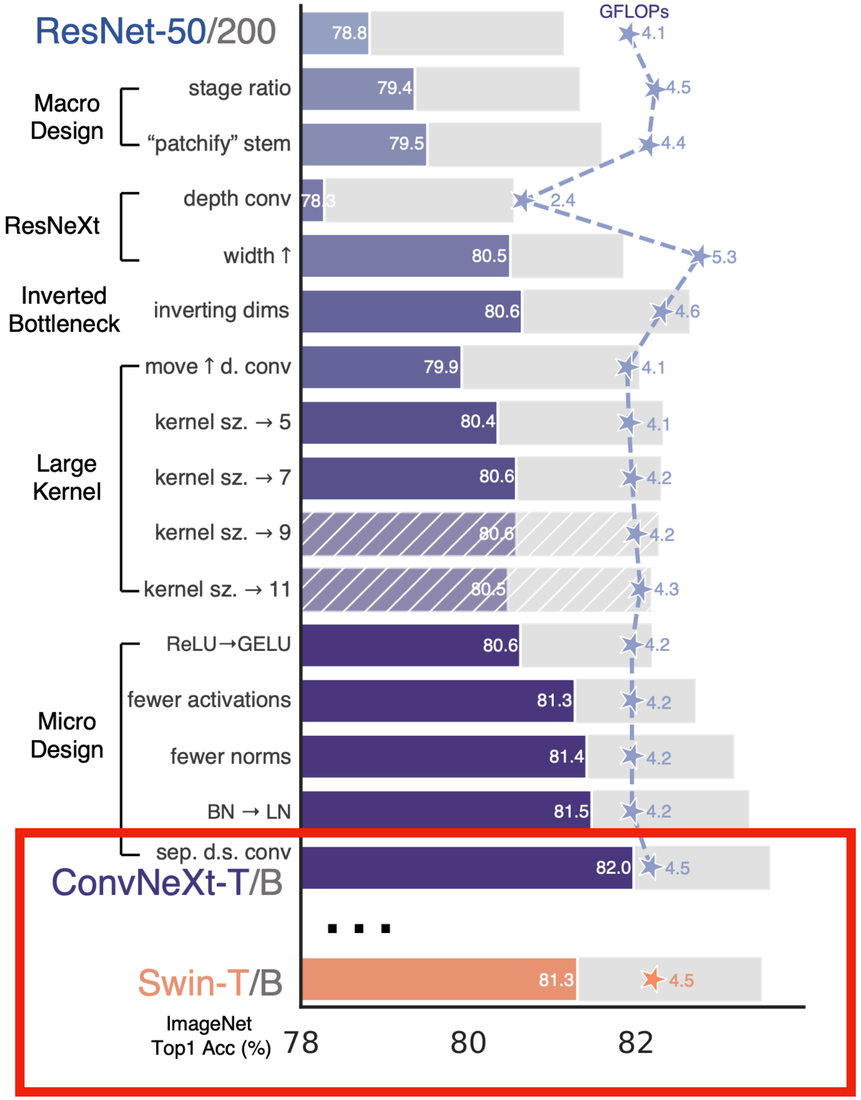


Fig. 6: ConvNext vs. Swin-T

Contributions & Results

- Accuracy: Classification

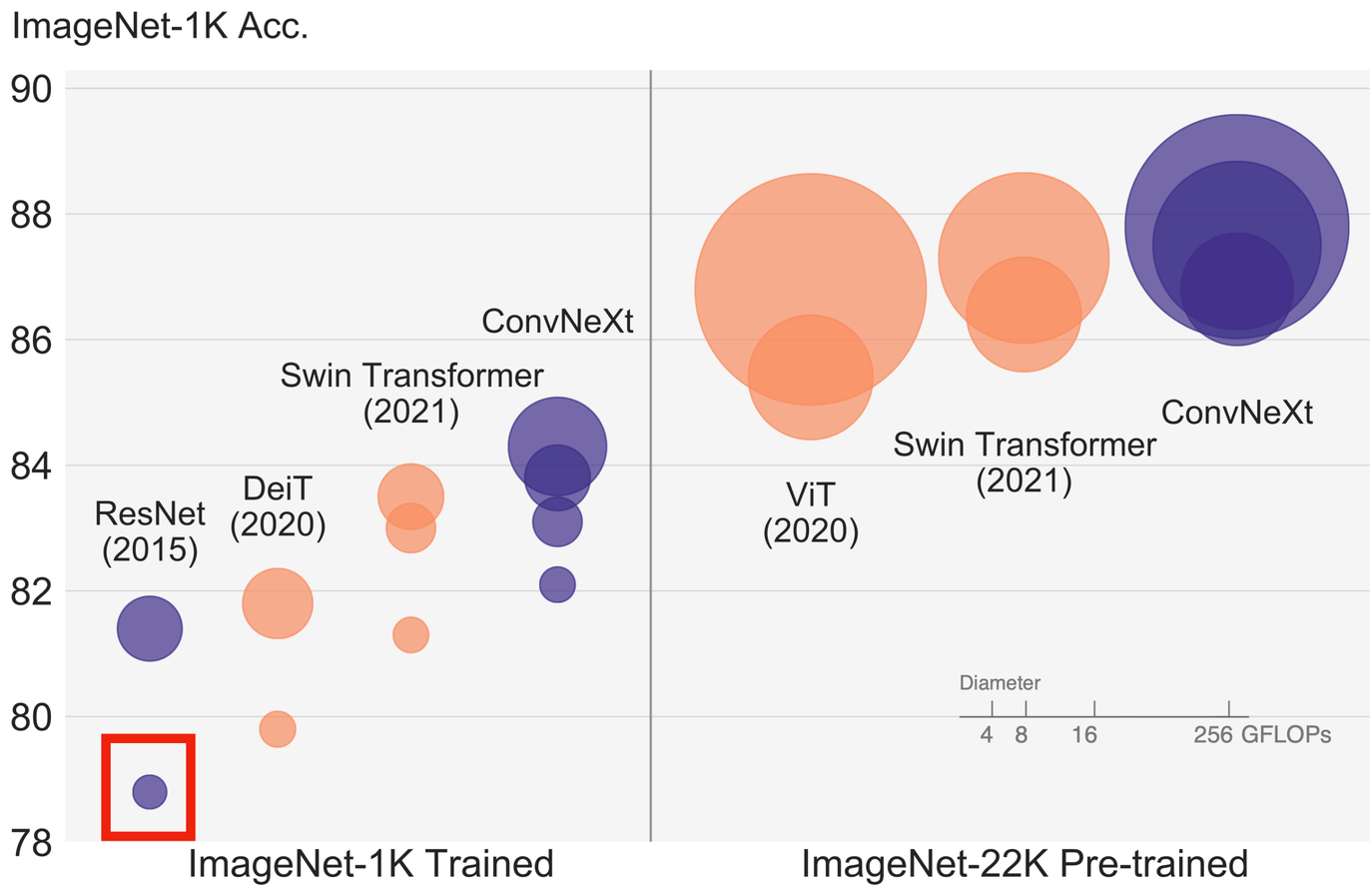


Fig. 7: Performance of various compute classes

Contributions & Results

- Accuracy: Classification

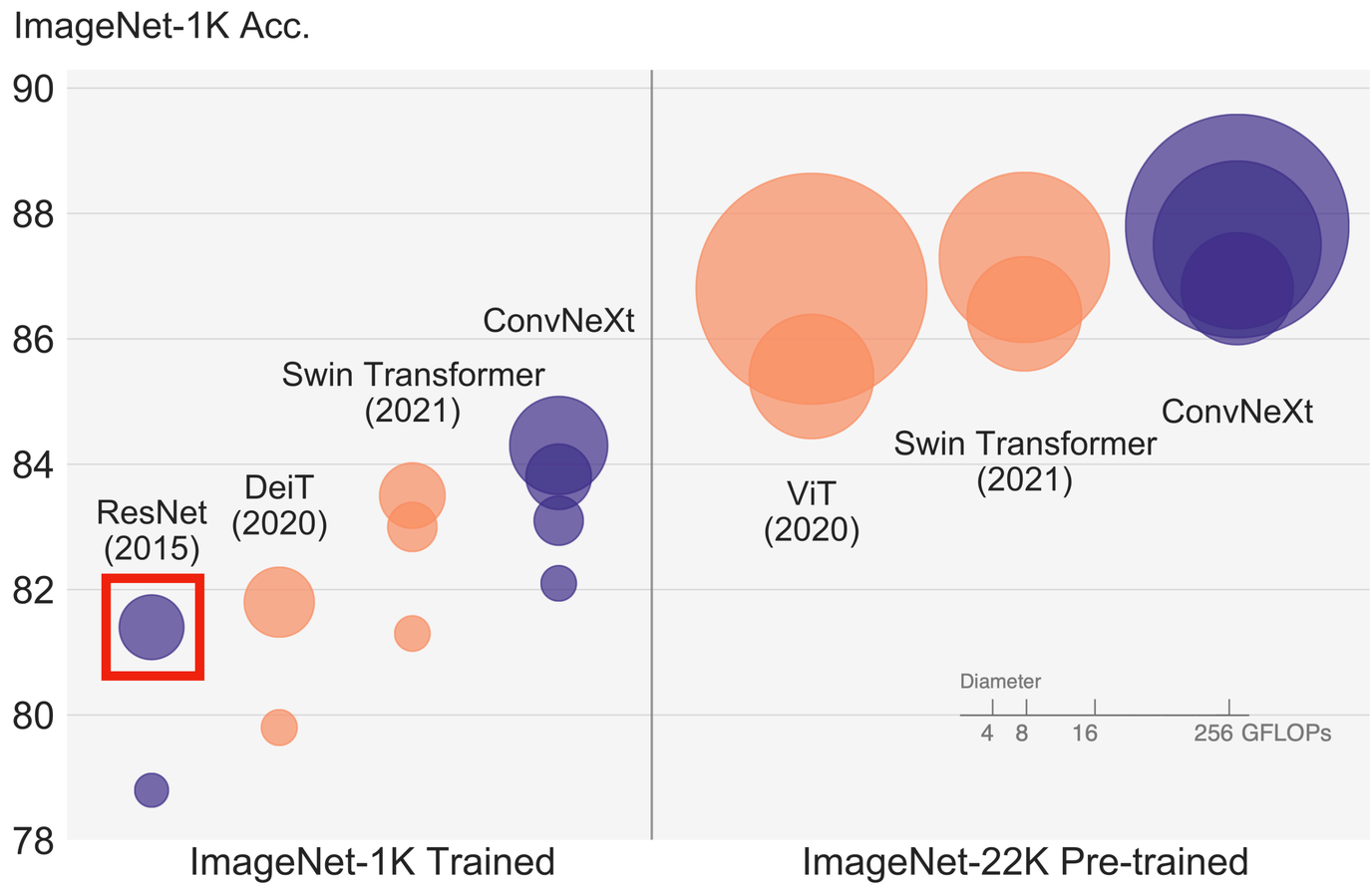


Fig. 7: Performance of various compute classes

Contributions & Results

- Accuracy: Classification

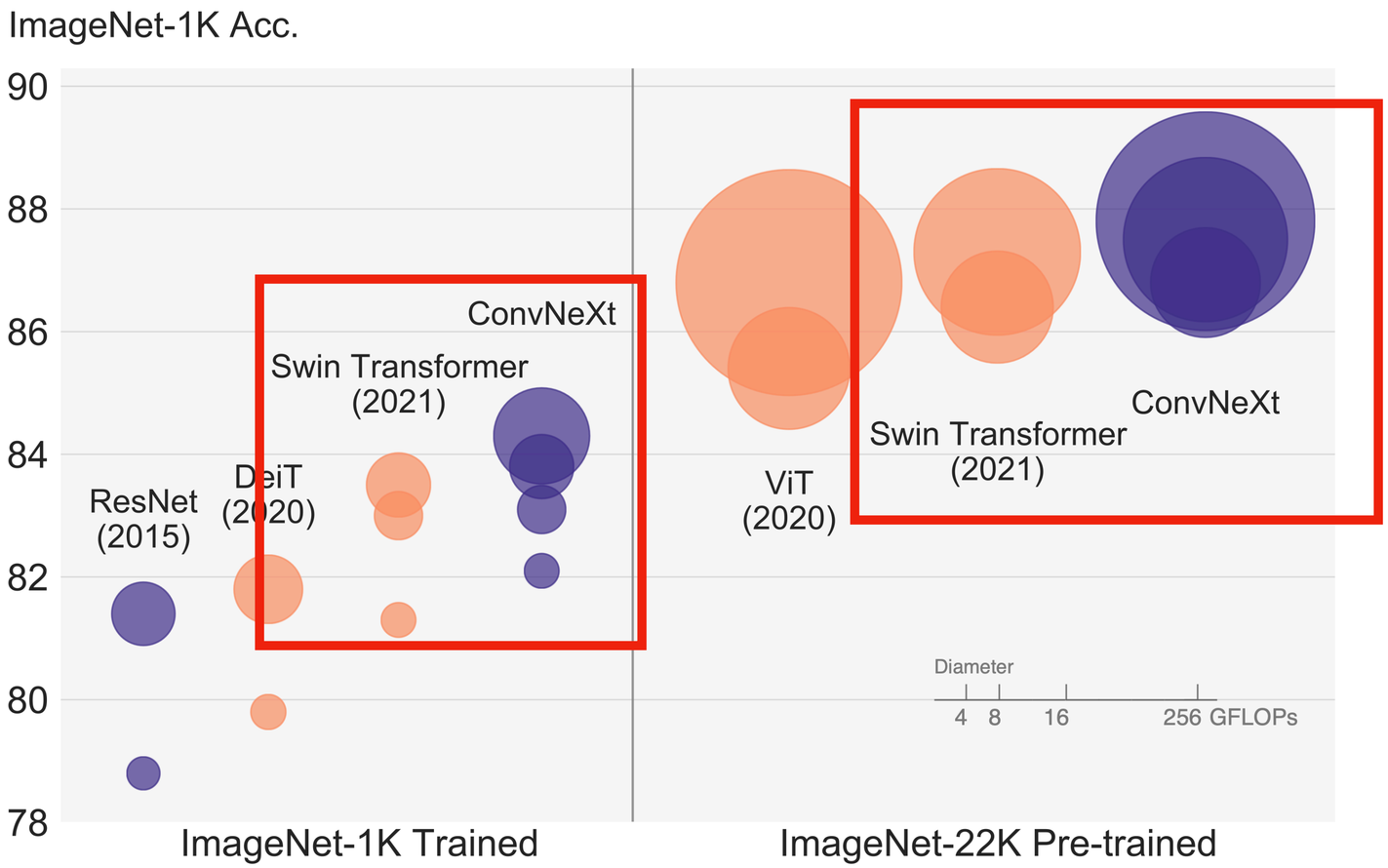


Fig. 7: Performance of various compute classes

Contributions & Results

- Accuracy: Object Detection & Segmentation

backbone	FLOPs	FPS	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
Mask-RCNN 3× schedule								
○ Swin-T	267G	23.1	46.0	68.1	50.3	41.6	65.1	44.9
● ConvNeXt-T	262G	25.6	**46.2**	67.9	50.8	**41.7**	65.0	44.9
Cascade Mask-RCNN 3× schedule								
● ResNet-50	739G	16.2	46.3	64.3	50.5	40.1	61.7	43.4
● X101-32	819G	13.8	48.1	66.5	52.4	41.6	63.9	45.2
● X101-64	972G	12.6	48.3	66.4	52.3	41.7	64.0	45.1
○ Swin-T	745G	12.2	50.4	69.2	54.7	43.7	66.6	47.3
● ConvNeXt-T	741G	13.5	**50.4**	69.1	54.8	**43.7**	66.5	47.3
○ Swin-S	838G	11.4	51.9	70.7	56.3	45.0	68.2	48.8
● ConvNeXt-S	827G	12.0	**51.9**	70.8	56.5	**45.0**	68.4	49.1
○ Swin-B	982G	10.7	51.9	70.5	56.4	45.0	68.1	48.9
● ConvNeXt-B	964G	11.4	**52.7**	71.3	57.2	**45.6**	68.9	49.5
○ Swin-B‡	982G	10.7	53.0	71.8	57.5	45.8	69.4	49.7
● ConvNeXt-B‡	964G	11.5	**54.0**	73.1	58.8	**46.9**	70.6	51.3
○ Swin-L‡	1382G	9.2	53.9	72.4	58.8	46.7	70.1	50.8
● ConvNeXt-L‡	1354G	10.0	**54.8**	73.8	59.8	**47.6**	71.3	51.7
● ConvNeXt-XL‡	1898G	8.6	**55.2**	74.2	59.9	**47.7**	71.6	52.2

Fig. 8: Performance in COCO dataset

Contributions & Results

- Accuracy: Object Detection & Segmentation
- Plus additional tasks

Model	Data/Size	FLOPs / Params	Clean	C (↓)	$\bar{C}$ (↓)	A	R	SK
ResNet-50	$1K/224^2$	4.1 / 25.6	76.1	76.7	57.7	0.0	36.1	24.1
Swin-T [45]	$1K/224^2$	4.5 / 28.3	81.2	62.0	-	21.6	41.3	29.1
RVT-S* [47]	$1K/224^2$	4.7 / 23.3	81.9	49.4	37.5	25.7	47.7	34.7
ConvNeXt-T	$1K/224^2$	4.5 / 28.6	82.1	53.2	40.0	24.2	47.2	33.8
Swin-B [45]	$1K/224^2$	15.4 / 87.8	83.4	54.4	-	35.8	46.6	32.4
RVT-B* [47]	$1K/224^2$	17.7 / 91.8	82.6	46.8	**30.8**	28.5	48.7	36.0
ConvNeXt-B	$1K/224^2$	15.4 / 88.6	**83.8**	**46.8**	34.4	**36.7**	**51.3**	**38.2**
ConvNeXt-B	$22K/384^2$	45.1 / 88.6	86.8	43.1	30.7	62.3	64.9	51.6
ConvNeXt-L	$22K/384^2$	101.0 / 197.8	87.5	40.2	29.9	65.5	66.7	52.8
ConvNeXt-XL	$22K/384^2$	179.0 / 350.2	**87.8**	**38.8**	**27.1**	**69.3**	**68.2**	**55.0**

Fig. 9: Performance in ADE20K dataset

Contributions & Results

- Other desired properties: Robustness, Scalability, Efficiency

Contributions & Results

- Robustness

Model	Data/Size	FLOPs / Params	Clean	C (↓)	$\bar{C}$ (↓)	A	R	SK
ResNet-50	$1K/224^2$	4.1 / 25.6	76.1	76.7	57.7	0.0	36.1	24.1
Swin-T [45]	$1K/224^2$	4.5 / 28.3	81.2	62.0	-	21.6	41.3	29.1
RVT-S* [47]	$1K/224^2$	4.7 / 23.3	81.9	49.4	37.5	25.7	47.7	34.7
ConvNeXt-T	$1K/224^2$	4.5 / 28.6	82.1	53.2	40.0	24.2	47.2	33.8
Swin-B [45]	$1K/224^2$	15.4 / 87.8	83.4	54.4	-	35.8	46.6	32.4
RVT-B* [47]	$1K/224^2$	17.7 / 91.8	82.6	46.8	**30.8**	28.5	48.7	36.0
ConvNeXt-B	$1K/224^2$	15.4 / 88.6	**83.8**	**46.8**	34.4	**36.7**	**51.3**	**38.2**
ConvNeXt-B	$22K/384^2$	45.1 / 88.6	86.8	43.1	30.7	62.3	64.9	51.6
ConvNeXt-L	$22K/384^2$	101.0 / 197.8	87.5	40.2	29.9	65.5	66.7	52.8
ConvNeXt-XL	$22K/384^2$	179.0 / 350.2	**87.8**	**38.8**	**27.1**	**69.3**	**68.2**	**55.0**

Fig. 10: Performance in ImageNet-C dataset

Contributions & Results

- Scalability & Efficiency
    - ConvNext-L
    - ConvNext-XL
    - ~49% higher throughput

Contributions & Results

- Conclusion

attention is NOT “all you need”

Thank You!