ESCAPING THE BIG DATA PARADIGM WITH COMPACT TRANSFORMERS

ALI HASSANI, STEVEN WALTON, NIKHIL SHAH,
ABULIKEMU ABUDUWEILI, JIACHEN LI, HUMPHREY SHI

Akila Peiris

We will cover...

- Introduction
- Related Works
- Method
- Experiments
- Conclusion

INTRODUCTION

(CNN)

- Convolutional neural networks (CNNs)[1]
- Standard for computer vision[2]
  - Invariance to spatial translations
  - Low relational inductive bias
- Improved with residual connections[3]
- Efficiency[4]
  - Sparse interaction
  - Weight sharing
  - Equivariant representations

[1] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541–551, 1989..

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25:1097–1105, 2012.

[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.

[4] Ian Goodfellow, YoshuaBengio, AaronCourville, and Yoshua Bengio. Deep learning. MIT press Cambridge, 2016.

INTRODUCTION

(TRANSFORMERS)

- Attention is All You Need[5]
- Originated in natural language processing
- First major usage on vision: Vision Transformer (ViT)[6]
  - Large-scale training can trump inductive biases
  - *“Transformers lack some of the inductive biases inherent to CNNs, such as translation equivariance and locality, and therefore do not generalize well when trained on insufficient amounts of data.”*
- “Data hungry” paradigm
  - Larger models
  - Larger datasets
- Training transformers from scratch is impossible in most cases

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.
[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

RELATED WORKS

- Vision Transformer[6]
  - Image Tokenization
  - Positional Embedding
  - Transformer Encoder
  - Classification
- Data-Efficient Transformers
  - Data-Efficient Image Transformers (DeiT)[7]
  - Tokens-to-token ViT (T2T- ViT)[8]
- Convolution-inspired Transformers
  - ConViT[9]

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
[7] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877, 2020.
[8] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. arXiv preprint arXiv:2101.11986, 2021.
[9] Stéphane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. arXiv preprint

METHOD

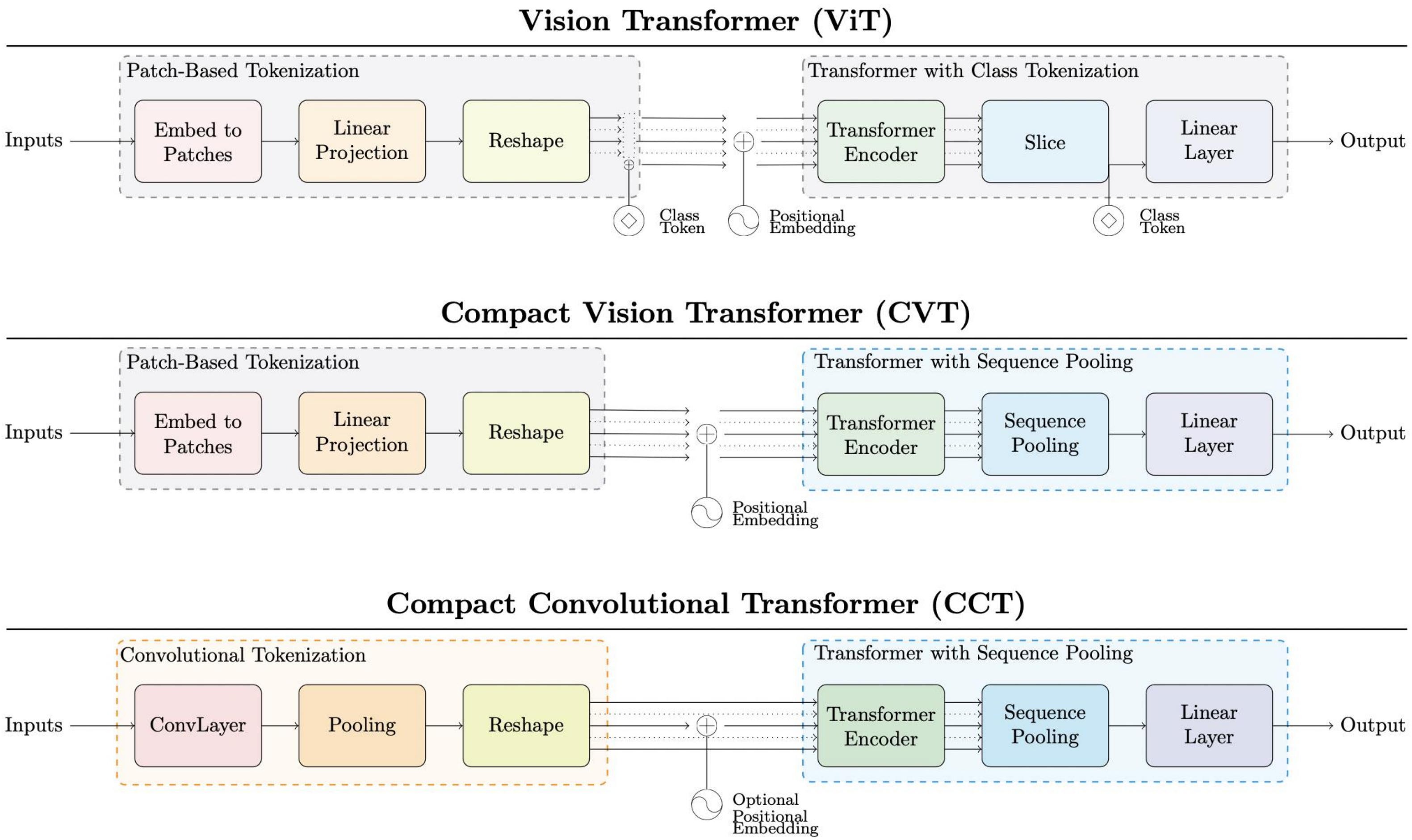

Figure 1: Comparing ViT (top) to CVT (middle) and CCT (bottom). CVT can be thought of as an ablated version of CCT, only utilizing sequence pooling and not a convolutional tokenizer. CVT may be preferable with more limited compute, as the patch-based tokenization is faster.

METHOD

TRANSFORMER-BASED BACKBONE

- Follow the original Transformer[5] and original Vision Transformer (ViT)[6]
- Encoder consists of transformer blocks
  - Multi-Headed Self-Attention (MHSA) layer
  - Multi-Layer Perceptron (MLP) block
- Layer Normalization
- GELU activation
- Dropout
- Positional embeddings
  - Learnable or sinusoidal (sine wave), both of which are effective.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.
[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

METHOD

SMALL AND COMPACT MODELS

- Smaller and more compact vision transformers
  - ViT-Base
    - 12 layer transformer encoder
    - 12 attention heads
    - 64 dimensions per head
    - 2048-dimensional hidden layers in the MLP blocks
    - 85M parameters total
- Proposed variants (ViT-Lite)
  - 2 layers
  - 2 heads
  - 128-dimensional hidden layers
  - Smallest has 0.22M parameters in total
  - Largest only has 3.8M parameters

METHOD

SeqPool

- Traditionally, to map the sequential outputs to a singular class index
  - Transformer-based classifiers follow BERT[10]
  - Global average pooling
- SeqPool - attention-based method which pools over the output sequence of tokens.
  - Output sequence contains information across different parts of the input (improve performance)
  - No additional parameters compared to learnable token
  - One less token being forwarded (decreases computation)

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, 2019.

METHOD

SeqPool

- Maps output sequence using $T : \mathbb{R}^{b\times n\times d} \mapsto \mathbb{R}^{b\times d}$ given $\mathbf{x}_L = \mathrm{f}(\mathbf{x}_0) \in \mathbb{R}^{b\times n\times d}$
  - x_L - output of an L layer transformer encoder f
  - b - batch size
  - n - sequence length
  - d - the total embedding dimension
- x_L is then fed to a linear layer $\mathrm{g}(\mathbf{x}_L) \in \mathbb{R}^{d\times 1}$ with Softmax activation

$$\mathbf{x}'_L = \mathrm{softmax}\left(\mathrm{g}(\mathbf{x}_L)^T\right) \in \mathbb{R}^{b\times 1\times n}$$

- This generates an importance weighting for each input token, applied as follows

$$\mathbf{z} = \mathbf{x}'_L \mathbf{x}_L = \mathrm{softmax}\left(\mathrm{g}(\mathbf{x}_L)^T\right) \times \mathbf{x}_L \in \mathbb{R}^{b\times 1\times d}$$

- The output $z \in \mathbb{R}^{b\times d}$ is produced by flattening

METHOD

Convolutional Tokenizer

- Replace patch and embedding with a simple convolutional block
  - Introduces inductive bias
  - Single convolution
  - ReLU activation
  - Max pooling
- Given an image or feature map $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$

$$\mathbf{x}_0 = \mathrm{MaxPool}(\mathrm{ReLU}(\mathrm{Conv2d}(\mathbf{x})))$$

  - Conv2d operation has d filters (embedding dimension of the transformer backbone)
- The convolution and maxpool can overlap (injecting inductive biases)
- Advantages
  - Maintains locally spatial information.
  - No longer tied to the input resolution strictly divisible by the pre-set patch size
  - Performance gains
  - Gives more flexibility toward removing the positional embedding in the model

EXPERIMENTS

- Datasets
  - CIFAR-10 (MIT License)[11]
  - CIFAR- 100 (MIT License)[11]
  - MNIST
  - Fashion-MNIST
  - Oxford Flowers-102 [12]
  - ImageNet-1k[13]

Small-scale small resolution datasets

Small-scale larger resolution datasets

Medium-scale datasets

Single chanel
(low data density)

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.
[12] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pages 722–729. IEEE, 2008
[13] 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.

EXPERIMENTS

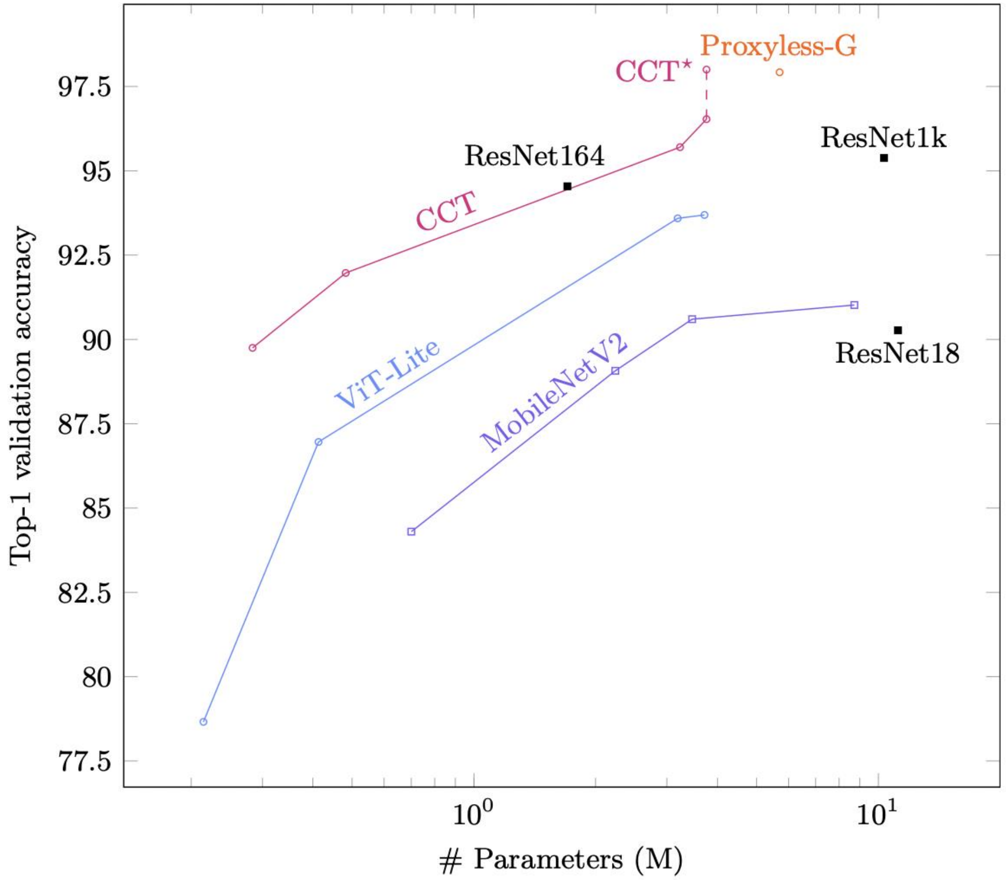

Figure 2: CIFAR-10 accuracy *vs*. model size (sizes < 12M). CCT was trained longer.

EXPERIMENTS

Existing

Model	C-10	C-100	Fashion	MNIST	# Params	MACs
Convolutional Networks (Designed for ImageNet)						
ResNet18	90.27%	66.46%	94.78%	99.80%	11.18 M	0.04 G
ResNet34	90.51%	66.84%	94.78%	99.77%	21.29 M	0.08 G
MobileNetV2/0.5	84.78%	56.32%	93.93%	99.70%	0.70 M	$<$ **0.01** G
MobileNetV2/2.0	91.02%	67.44%	95.26%	99.75%	8.72 M	0.02 G
Convolutional Networks (Designed for CIFAR)						
ResNet56[16]	94.63%	74.81%	95.25%	99.27%	0.85 M	0.13 G
ResNet110[16]	95.08%	76.63%	95.32%	99.28%	1.73 M	0.26 G
ResNet1k-v2⋆[17]	95.38%	–	–	–	10.33 M	1.55 G
Proxyless-G[5]	97.92%	–	–	–	5.7 M	–
Vision Transformers						
ViT-12/16	83.04%	57.97%	93.61%	99.63%	85.63 M	0.43 G

Proposed

Model	C-10	C-100	Fashion	MNIST	# Params	MACs
Vision Transformers						
ViT-Lite-7/16	78.45%	52.87%	93.24%	99.68%	3.89 M	0.02 G
ViT-Lite-7/8	89.10%	67.27%	94.49%	99.69%	3.74 M	0.06 G
ViT-Lite-7/4	93.57%	73.94%	95.16%	99.77%	3.72 M	0.26 G
Compact Vision Transformers						
CVT-7/8	89.79%	70.11%	94.50%	99.70%	3.74 M	0.06 G
CVT-7/4	94.01%	76.49%	95.32%	99.76%	3.72 M	0.25 G
Compact Convolutional Transformers						
CCT-2/3×2	89.75%	66.93%	94.08%	99.70%	**0.28** M	0.04 G
CCT-7/3×2	95.04%	77.72%	95.16%	99.76%	3.85 M	0.29 G
CCT-7/3×1	96.53%	80.92%	**95.56%**	**99.82%**	3.76 M	1.19 G
CCT-7/3×1⋆	**98.00%**	**82.72%**	–	–	3.76 M	1.19 G

Table 1: Top-1 validation accuracy comparisons. ⋆ variants were trained longer (see Table 2)

Note: MACs - multiply-and-accumulate operations (measure of computational complexity, lesser the better)

EXPERIMENTS

# Epochs	Pos. Emb.	CIFAR-10	CIFAR-100
300	Learnable	96.53%	80.92%
1500	Sinusoidal	97.48%	82.72%
5000	Sinusoidal	**98.00%**	**82.87%**

Table 2: Top-1 accuracy on CIFAR-10/100 when a CCT model with 7 transformer encoder layers, and a 1-layer convolutional tokenizer with 3×3 kernel size is trained longer.

EXPERIMENTS

Model	Top-1	# Params	MACs	Training Epochs
ResNet50 [16]	77.15%	25.55 M	4.15 G	120
ResNet50 (2021) [44]	79.80%	25.55 M	4.15 G	300
ViT-S [19]	79.85%	**22.05** M	**4.61** G	300
CCT-14/7×2	**80.67%**	22.36 M	5.53 G	300
DeiT-S [19]	81.16%	22.44M	**4.63** G	300
CCT-14/7×2 Distilled	**81.34%**	**22.36** M	5.53 G	300

Table 3: ImageNet Top-1 validation accuracy comparison (no extra data or pretraining).

Note: MACs - multiply-and-accumulate operations (measure of computational complexity, lesser the better)

EXPERIMENTS

Model	Resolution	Pretraining	Top-1	# Params	MACs
CCT-14/7×2	224	-	97.19%	22.17 M	18.63 G
DeiT-B	384	ImageNet-1k	98.80%	86.25 M	55.68 G
ViT-L/16	384	JFT-300M	99.74%	304.71 M	191.30 G
ViT-H/14	384	JFT-300M	99.68%	661.00 M	504.00 G
CCT-14/7×2	384	ImageNet-1k	**99.76%**	**22.17** M	**18.63** G

Table 4: Flowers-102 Top-1 validation accuracy comparison.

Note: MACs - multiply-and-accumulate operations (measure of computational complexity, lesser the better)

CONCLUSION

- Main contributions
  - Extending transformer-based research to small data regimes
    - ViT-Lite which can be trained from scratch and achieve high accuracy on small scale data sets
  - Introducing Compact Vision Transformer (CVT)
    - Performance improved using SeqPool strategy
  - Introducing Compact Convolutional Transformer (CCT)
    - Increase performance and provide flexibility for input image sizes
- CCT can outperform other transformer based models on small datasets
  - Significant reduction in computational costs and memory constraints
- Important to many scientific domains where data is far more limited

REFERENCES

[1] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541–551, 1989..

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25:1097–1105, 2012.

[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.

[4] IanGoodfellow, YoshuaBengio, AaronCourville, and Yoshua Bengio. Deep learning. MIT press Cambridge, 2016.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve´ Je´gou. Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877, 2020.

[8] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. arXiv preprint arXiv:2101.11986, 2021.

[9] Ste´phane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. arXiv preprint arXiv:2103.10697, 2021.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, 2019.

REFERENCES

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.
[12] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pages 722–729. IEEE, 2008
[13] 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.

THANK YOU

Q&A