EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

Aarav Khanna, Ashley Liu
Computer Science
Cornell University

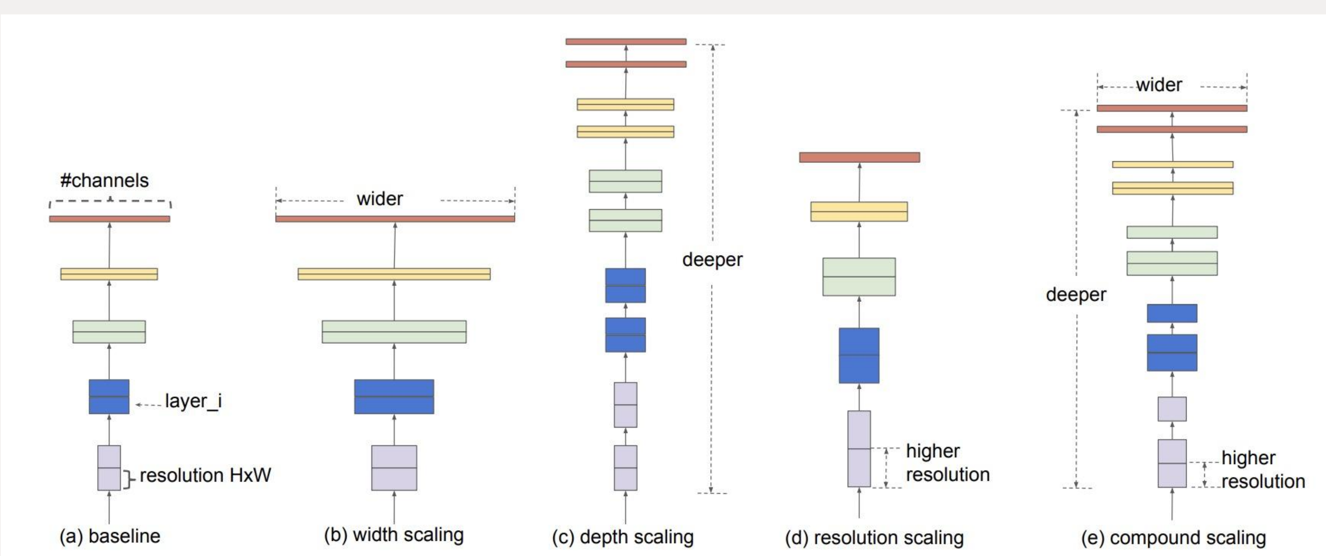
Introduction/Background

Background/Motivation:

- Traditional CNNs are often scaled arbitrarily by changing width, depth, or resolution independently
- Lots of new ConvNet architectures being created with excessive computational costs for little performance gains

Is there a principled way to scale CNNs to achieve better accuracy and efficiency?

Methodology (Compound Scaling)



Source: Tan & Le, 'EfficientNet: Rethinking Model Scaling for CNNs', ICML 2019

- **Key Principle**: Balance network width, depth, and resolution during scaling for optimal accuracy/efficiency
- **Implementation Formula**:
  - depth: $d = \alpha^\phi$
  - width: $w = \beta^\phi$
  - resolution: $r = \gamma^\phi$
  - constraint: $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$
- **Computational Efficiency**: With constraint $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$, total FLOPS increase by 2^ϕ for any new ϕ value
- **Optimal Scaling Coefficients (grid search)**: $\alpha = 1.2$, $\beta = 1.1$, $\gamma = 1.15$
- Fix α, β, γ as constants and scale up baseline network with different ϕ values to get B1 to B7

Methodology (Training & Evaluation)

We built the EfficientNet architecture and trained models B0 through B7 on the CIFAR-100 dataset.

Base Architecture Implementation: Created the foundational EfficientNet-B0 network with MBConv blocks (Mobile Inverted Bottleneck Convolution) and squeeze-and-excitation optimization.

Compound Scaling: Applied the compound scaling method to generate models B1-B7 by systematically increasing width, depth, and resolution according to the paper's scaling coefficients.

Training Setup:

- Dataset: CIFAR-100
- Optimizer: SGD with momentum
- Loss Function: Cross-Entropy
- Training Duration: 200 epochs for each model variant
- Computing Resources: Google Colab GPUs

Evaluation Metrics: Top-1 accuracy on validation and test sets

Source: Ahmed, Tashin & Sabab, Noor. (2022). Classification and Understanding of Cloud Structures via Satellite Images with EfficientUNet..

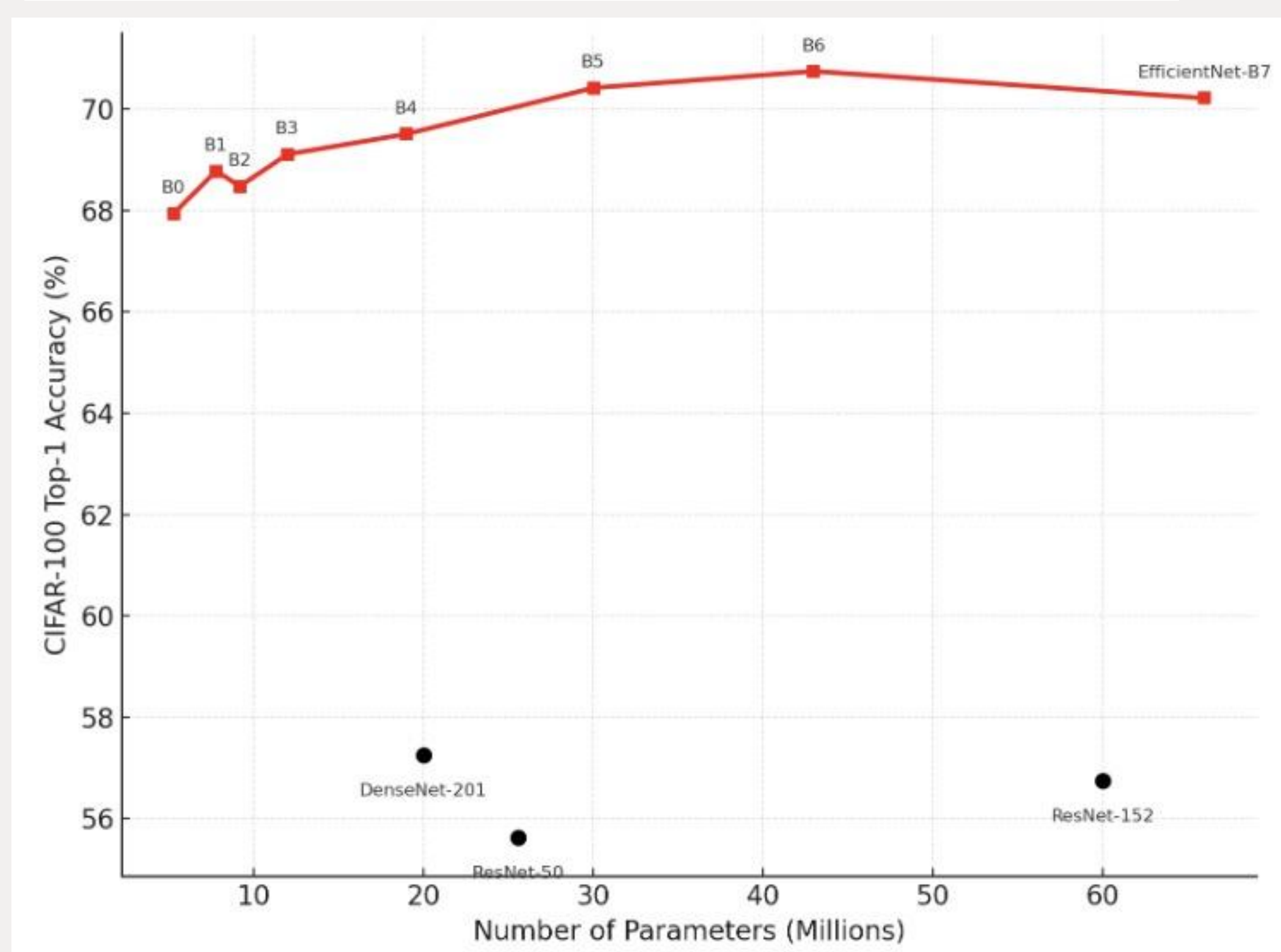
Results

Table 1: **CIFAR-100 Validation vs. Test Top-1 Accuracy**

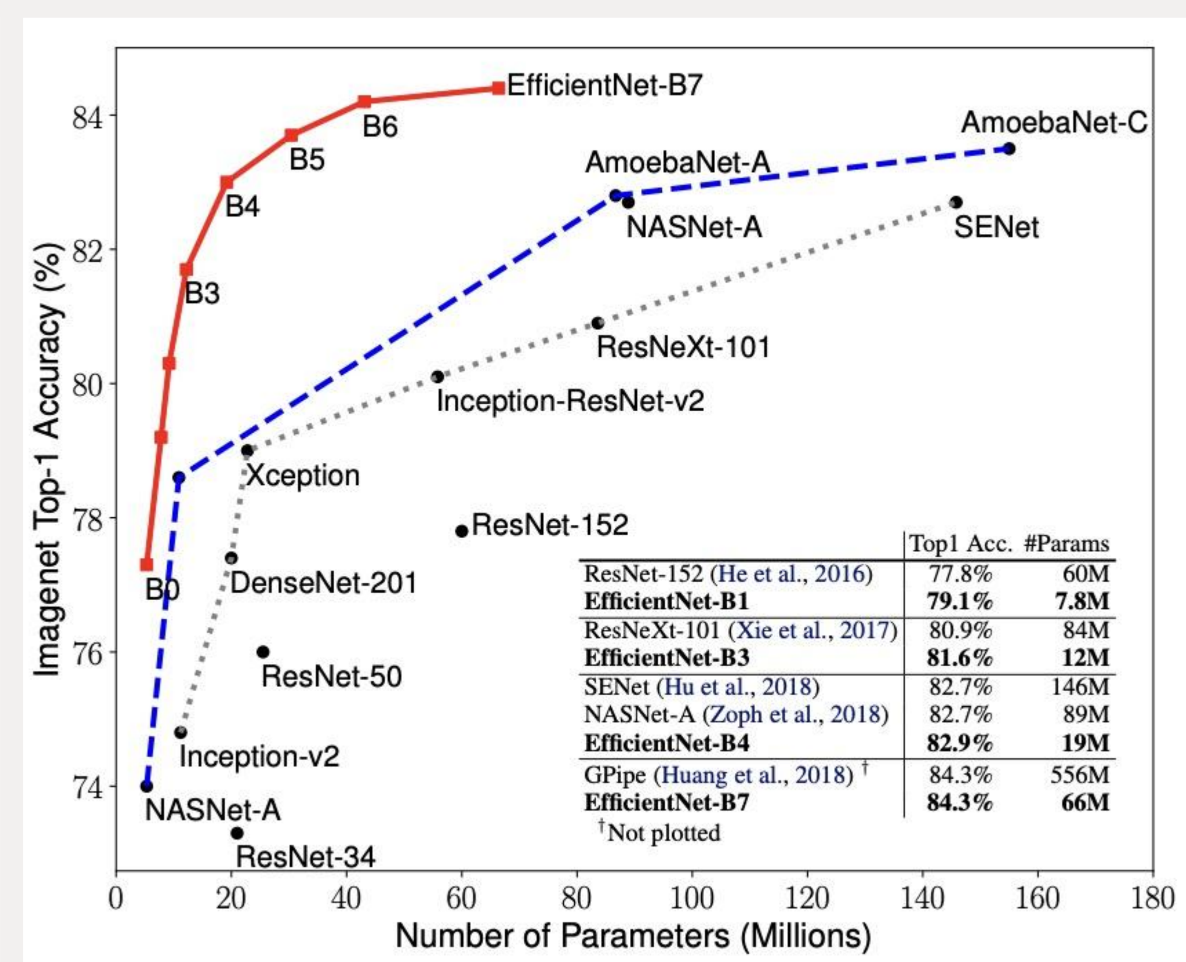
	B0	B1	B2	B3	B4	B5	B6	B7
Val top-1	63.14	64.26	63.94	65.98	67.04	66.44	66.56	67.48
Test top-1	67.94	68.78	68.48	69.11	69.51	70.42	70.75	70.22

Table 8. **ImageNet Validation vs. Test Top-1/5 Accuracy.**

	B0	B1	B2	B3	B4	B5	B6	B7
Val top1	77.11	79.13	80.07	81.59	82.89	83.60	83.95	84.26
Test top1	77.23	79.17	80.16	81.72	82.94	83.69	84.04	84.33



Source: Our implementation on CIFAR-100



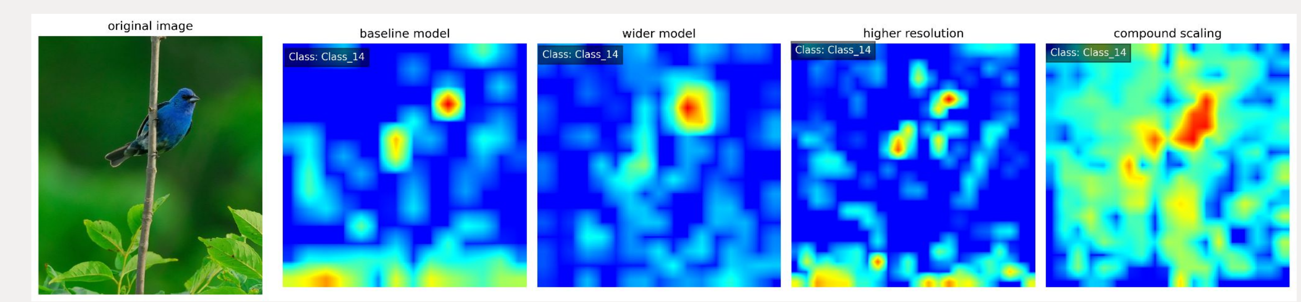
Source: Tan & Le, 'EfficientNet: Rethinking Model Scaling for CNNs', ICML 2019

Consistent Accuracy Improvement with Compound Scaling
Accuracy-Parameter Trade-off
Validation-Test Correlation

Conclusion

Our reimplementation confirms the effectiveness of EfficientNet's compound scaling method on the CIFAR-100 dataset.

- Systematic scaling of width, depth, and resolution leads to consistent performance improvements.
- The compound scaling approach produces models with better accuracy-parameter trade-offs than conventional scaling methods.
- Even with computational constraints and a different dataset (CIFAR-100 vs. ImageNet), the core principles of EfficientNet hold true.



Source: Class Activation Map (CAM) we generated for different scaling methods

Future Work

- Investigate transfer learning capabilities of pretrained EfficientNets on domain-specific tasks.
- Incorporate recent advancements like attention mechanisms or neural architecture search to further improve EfficientNet designs.
- Apply quantization and pruning techniques to further reduce model size while maintaining accuracy.

References

1. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ICML 2019.*
2. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *CVPR 2016.*
3. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.C. (2018). 4. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *CVPR 2018.*
4. Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K.Q. (2017). Densely Connected Convolutional Networks. *CVPR 2017.*
5. Ahmed, Tashin & Sabab, Noor. (2022). Classification and Understanding of Cloud Structures via Satellite Images with EfficientUNet. SN Computer Science.