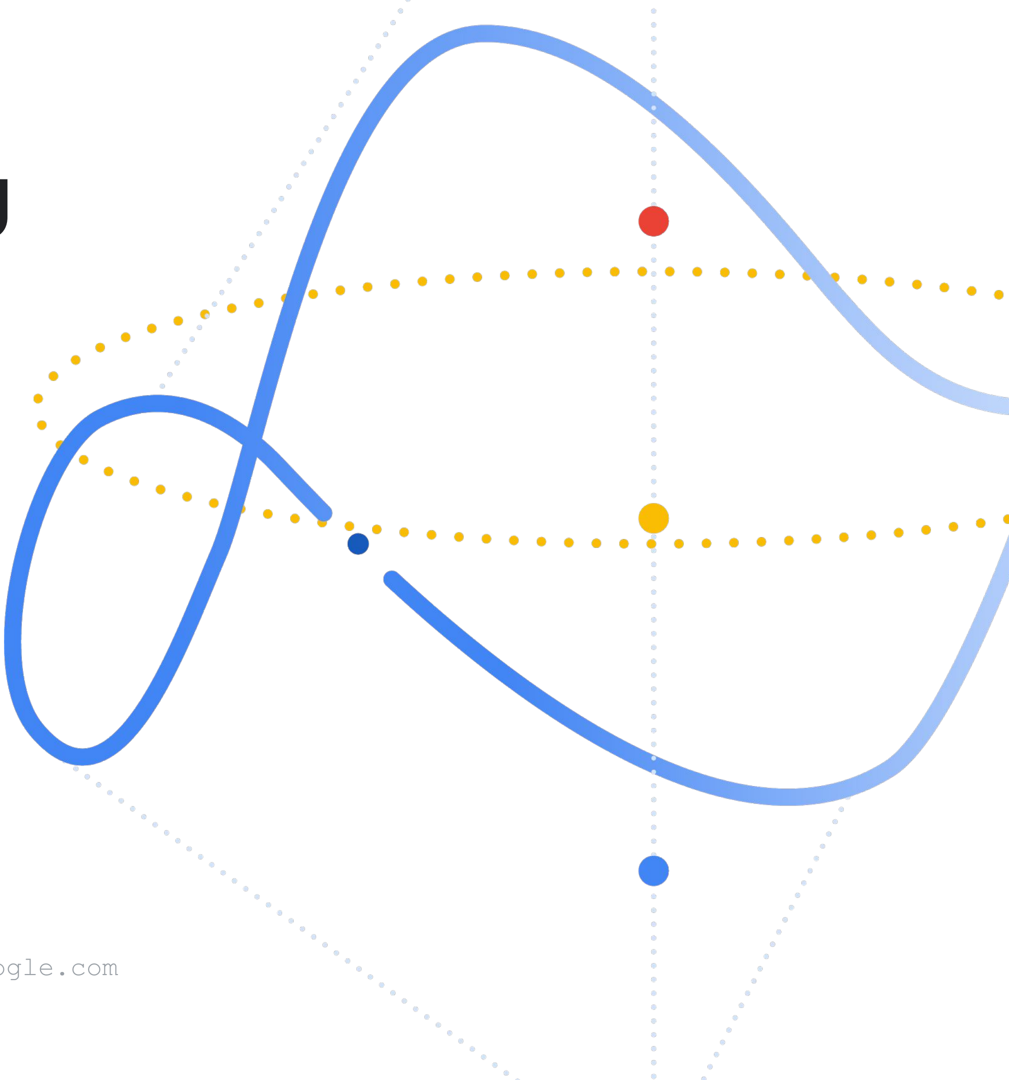
Google AI

When Ensembling Smaller Models is More Efficient than Single Large Models

WebVision 2020



Dan Kondratyuk, Mingxing Tan, Matthew Brown, Boqing Gong
{dankondratyuk,tanmingxing,mtbr,bgong}@google.com

Model Ensembles

- Train multiple models and average their predictions during inference
  - E.g., train a neural network architecture with different random initializations
- Easy method to reduce prediction error
- Introduces heavy efficiency penalties
  - Most commonly reserved for the largest models
- Can **small** ensembles be ***efficient***?

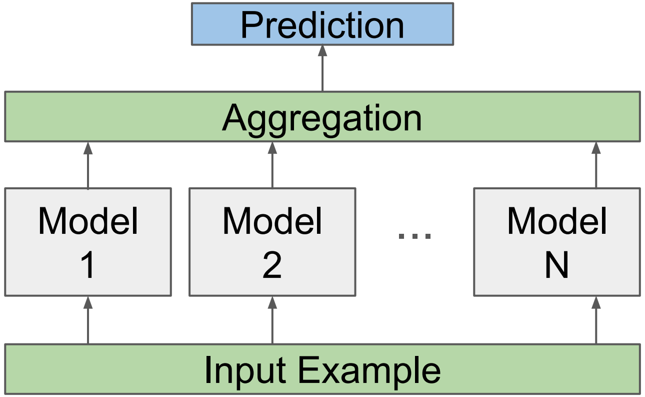


Image Classification - Wide ResNet - CIFAR 10

- Ensembles can be both **more accurate** *and* **more efficient**
  - Each line represents one model architecture
  - Each point indicates the number of models ensembled
  - As model sizes get larger, the **performance gap widens**
  - Larger ensembles produce diminishing returns and become less efficient

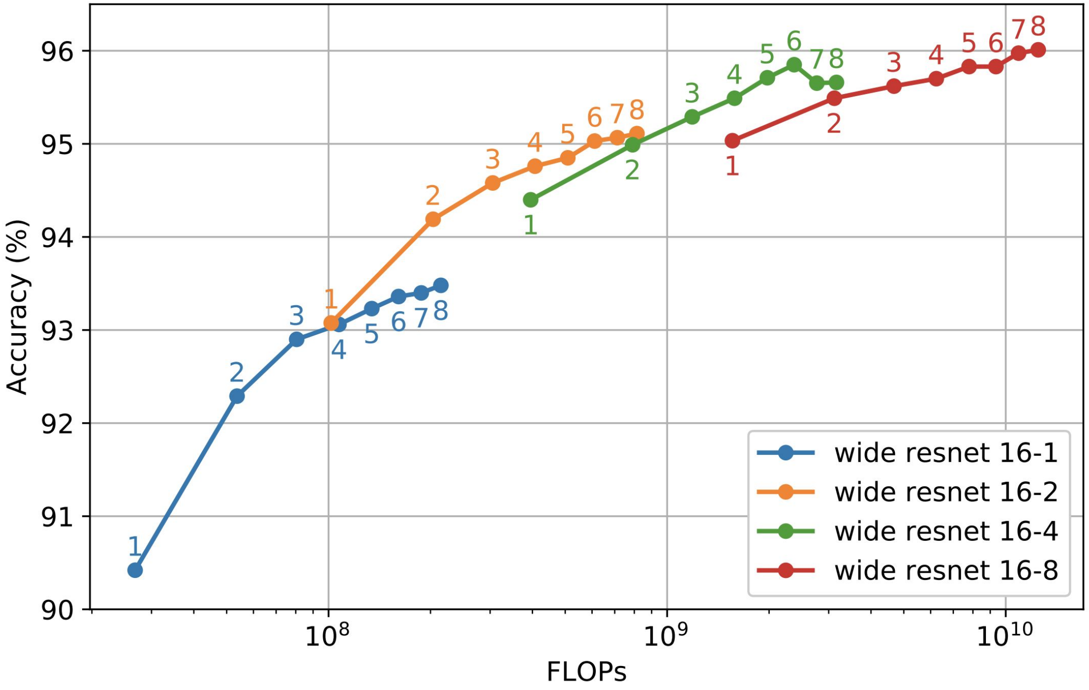
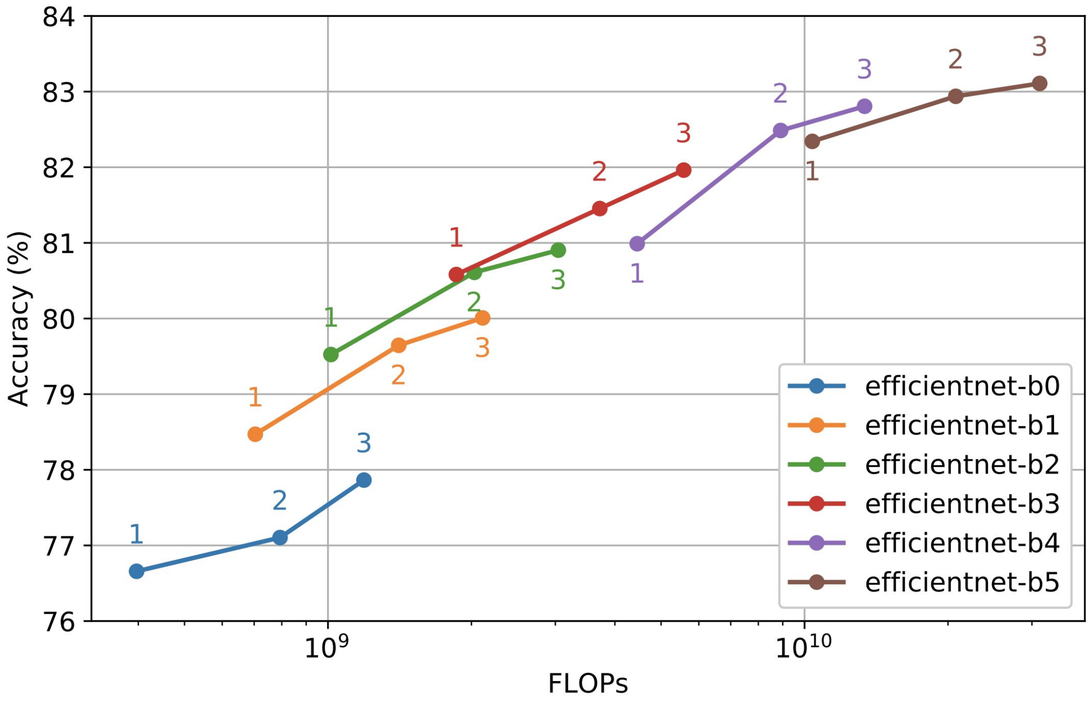


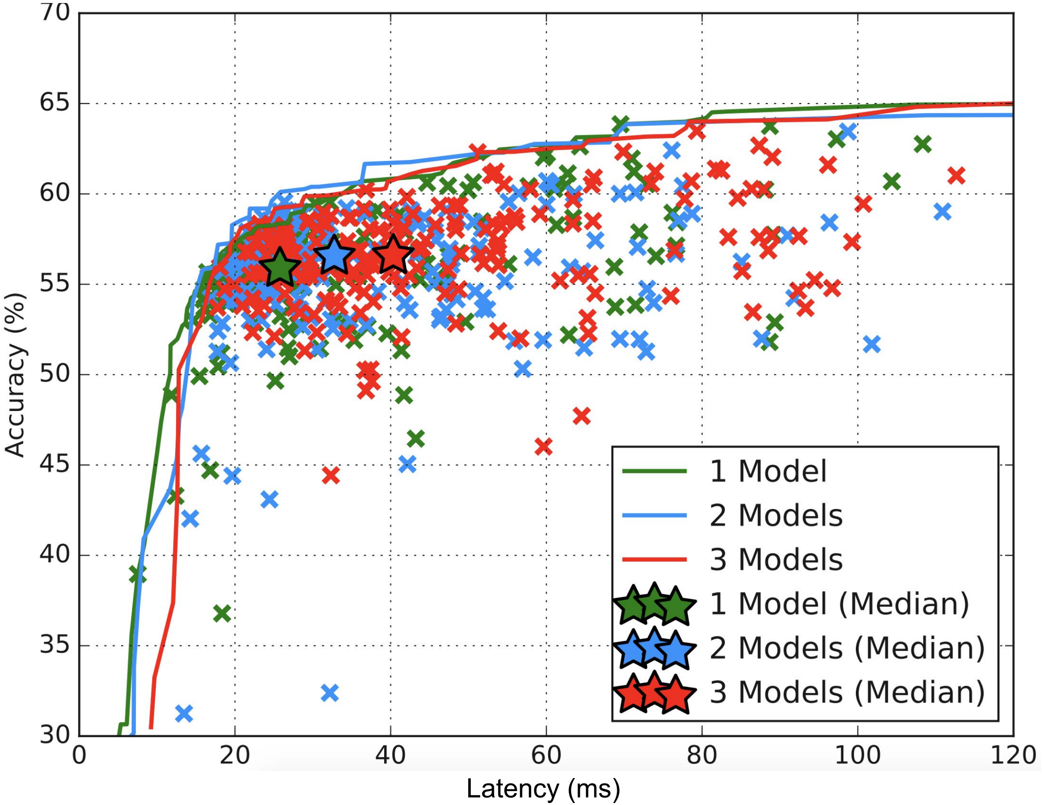
Image Classification - EfficientNet - ImageNet

- This trend appears for highly optimized models on larger datasets as well
  - EfficientNet scales the width, depth, and resolution of each model size



NAS Ensemble - ImageNet

- Can we use NAS to generate diverse ensemble architectures?
  - Can architecture diversity boost the accuracy to FLOPs/latency ratio?
  - Pareto curve shown for model ensembles searched with NAS
  - Surprisingly, a single searched model performs **nearly the same** as a diverse ensemble



Conclusion

- Ensembles of smaller models can be **more accurate** and **more efficient** than single large models, *especially* as model size grows
    - One can use ensembles as a more flexible trade-off between a model's inference speed and accuracy
    - Ensembles can be easily distributed across multiple workers, further increasing efficiency
- A single searched model using NAS can find a well-optimized architecture for ensembling
    - However, ensembling diverse architectures from a search on multiple models performs nearly the same as ensembling one model architecture from the search