Harnessing Cheap DNNs in Cascade Inference Pipelines

Tiantu Xu
(Teammate: Shuang Zhai)

Introduction

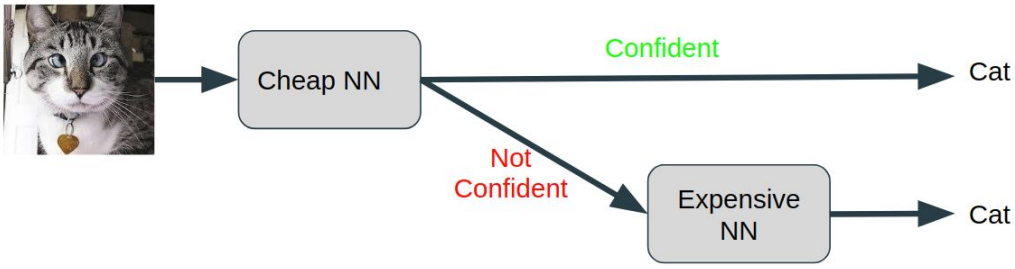
The emergence of millions of surveillance cameras in major cities like Beijing and London calls for fast retrospective image analytics. High accuracy deep neural networks for image classification are notoriously computationally expensive. Due to the diversity of different DNN models, some models can run at a much faster speed with moderate accuracy drop. By composing 2 models into cascades, there is a rich trade-off space between **accuracy** and **inference speed.**

Related Work

Applying specialized/cheap DNNs in model cascades is not a brand new idea. Prior work [1,2,3] has composed the cascade models with fast decent-accuracy DNNs and slow high-accuracy DNN models. However, there lacks a work that systematically studies how to pick up the appropriate cheap NN.

Major Contributions

The figure below shows a sample cascade inference pipeline.



Contribution 1: Choosing appropriate models

Toward choosing appropriate cheap NNs, we propose 2 options:

(1) **Inter-model approach**: Picking up different models

(2) **Intra-model approach**: Picking up one single model, and modifies the number of convolutional layers

Contribution 2: Online Aggregation

We can provide the aggregated results on-the-fly (Progress & Aggregated Accuracy) based on the progress of the inference. The user can choose to early abort the inference and trade time for lower accuracy.

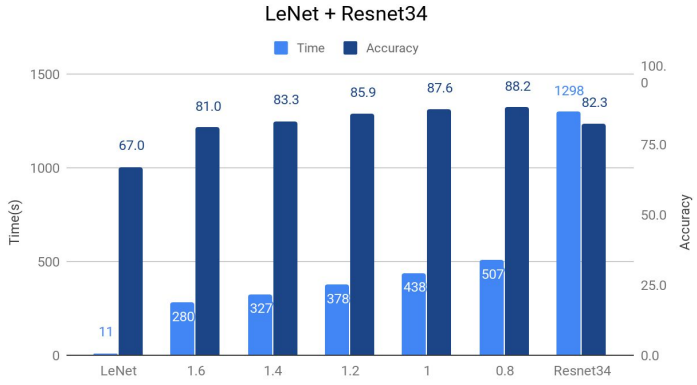
Implementation

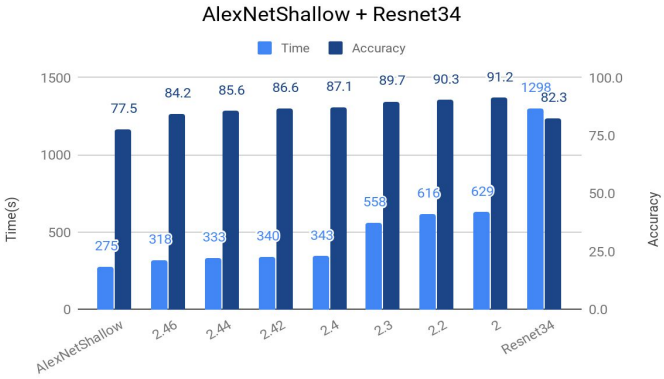
We systematically search for the options of our model. For inter-model selection, we choose LeNet & ResNet18 as our candidate; For intra-model selection, we choose AlexNet by removing 1 or 2 of its own convolutional layer. For our expensive/full NN, we choose ResNet34. To make the cheap model specialized, we use CIFAR-10 as our training & validation set.

Evaluation

The evaluation is divided into 2 parts: the impact of choices of (1)threshold and (2)models on the hybrid model

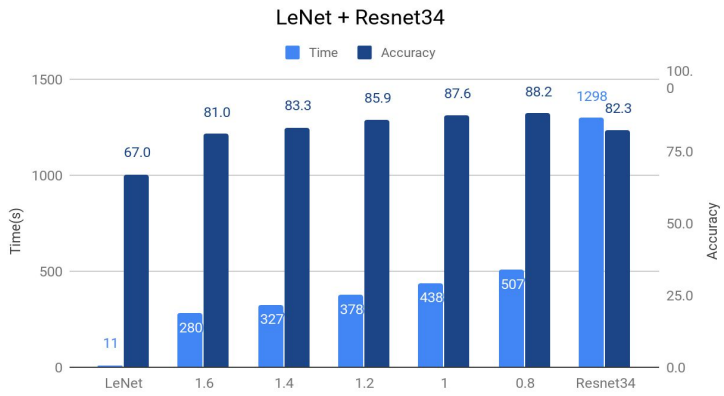
Part 1: Choice of thresholds (Confidence Level)

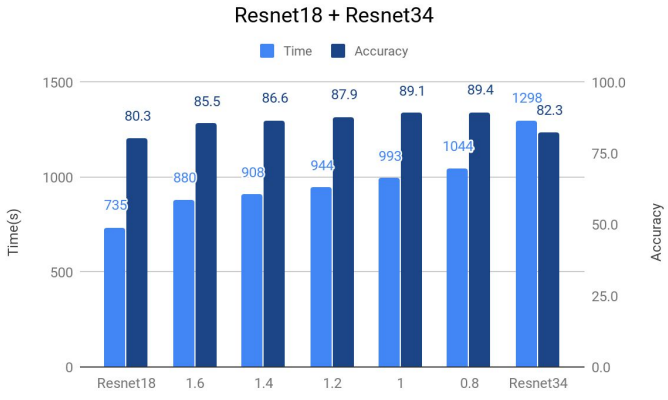




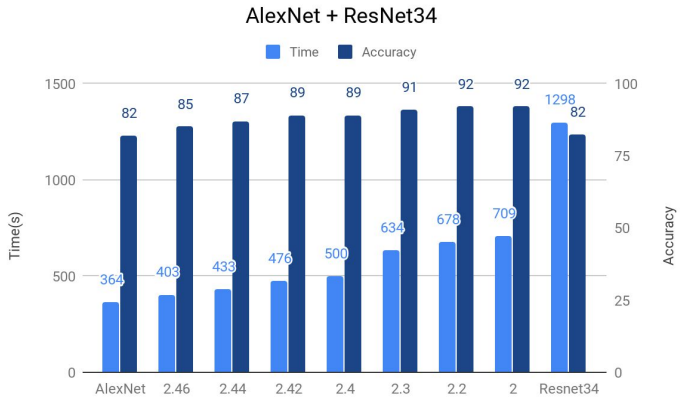
From the results, we can see that by choosing a different threshold in the same model, it provides various trade-offs between the accuracy and inference speed. By choosing a lower threshold, the more frames will be filtered out by the cheap NN, thus making the inference speed fast; by choosing a higher threshold, more images will go to the high accuracy expensive NN, thus spending more time. Note that in AlexNetShallow, we removed 1 conv layer from original AlexNet.

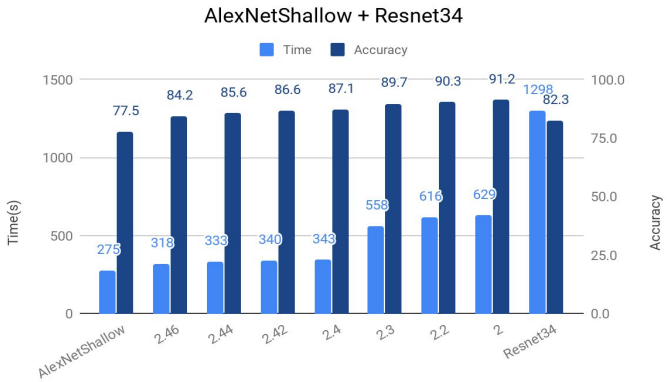
Part 2: Choice of models





From above results, we can see that by choosing different models, it provides various trade-offs between the accuracy and inference speed. LeNet is cheaper in computation, but it also has lower accuracy; ResNet18 is comparatively computationally expensive, but has higher accuracy.





From above results, we can see that by removing conv layers in the model, we can have various trade-offs between the accuracy and inference speed. AlexNetShallow is cheaper in computation compared to original AlexNet, but has a little bit lower accuracy.

Conclusion & Future Ideas

By choosing appropriate cheap models, the DNN inference pipeline can have up to 4.6x speedup. By choosing different threshold selection **further** provides trade-offs between accuracy and inference speed.

This idea can be applied to answer queries from the surveillance videos like “Find a bike in this surveillance camera view from 3:00 pm to 5:00 pm”. When answering video queries, we can further apply cheaper early filters like difference/motion filters, which can run at the speed of 100k frames per second.

References

1. Han, Seungyeop, et al. "MCDNN: An approximation-based execution framework for deep stream processing under resource constraints." *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2016.
2. Shen, Haichen, et al. "Fast video classification via adaptive cascading of deep models." *arXiv preprint* (2017).
3. Kang, Daniel, et al. "NoScope: optimizing neural network queries over video at scale." *Proceedings of the VLDB Endowment* 10.11 (2017): 1586-1597.