Fast Video Classification via Adaptive Cascading of Deep Models

Haichen Shen

Arvind Krishnamurthy

Seungyeop Han

Matthai Philipose

University of Washington

Rubrik

Microsoft

Recognizing entities in every frame of videos

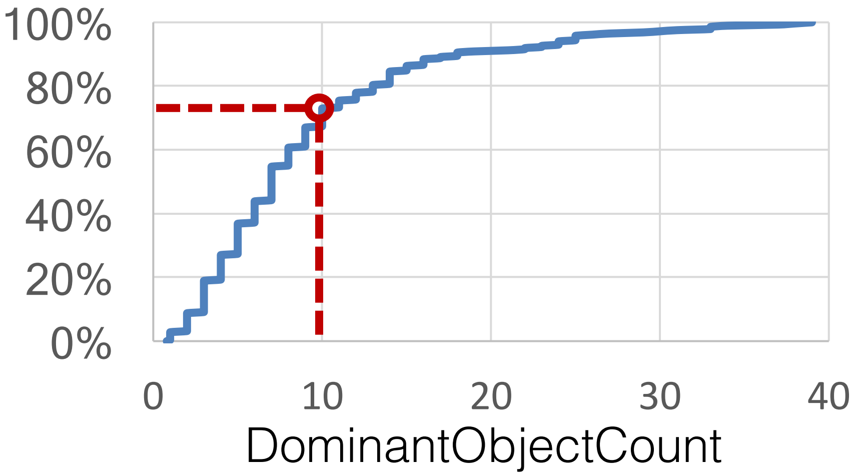
- Convolutional neural networks (“Oracle” model)
  - ✔ High accuracy in recognizing thousands of classes
  - ✗ Expensive to execute
- Simpler convolutional neural networks (”Compact” model)
  - ✗ Low accuracy in recognizing thousands of classes
  - ✔ Cheap to execute

How can we reconcile this?

Object Skew in 1-minute video segments

- DominantObjectCount: # of objects that account for 80% of all object occurrences in 1-minute segments

- Day-to-day video contains a tiny subset of classes in a short interval.



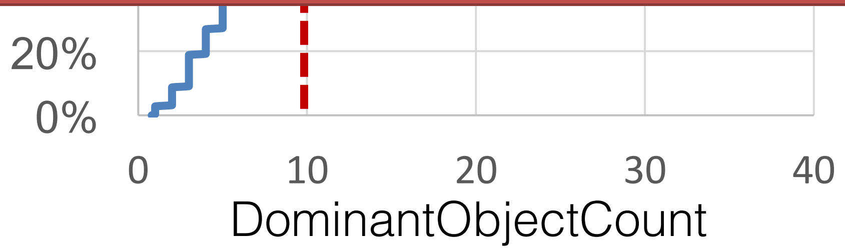
70% of segments have DominantObjectCount <= 10

Object Skew in 1-minute video segments

- DominantObjectCount: # of objects that account for 80% of all object occurrences in 1-minute segments

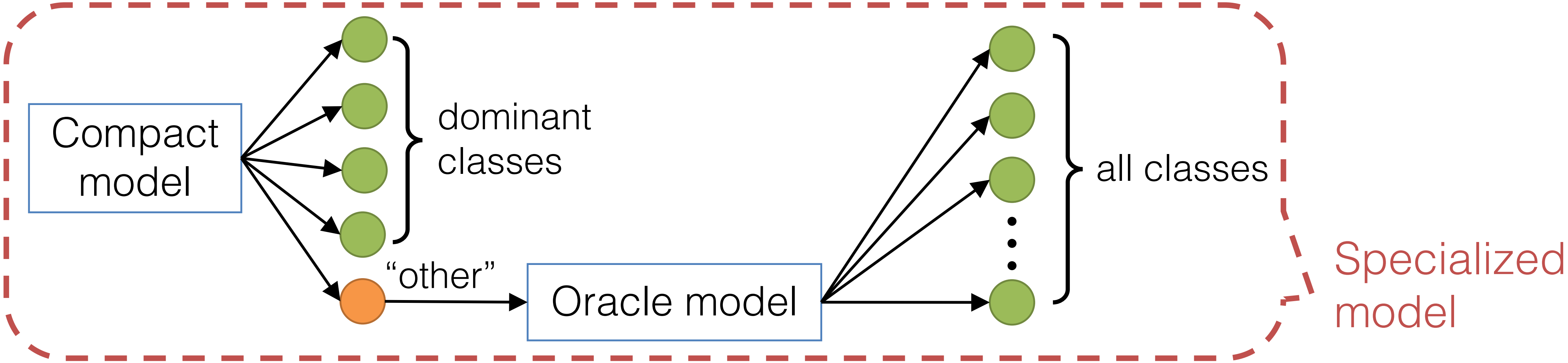
- Day-to-day video contains a tiny subset of classes in a short interval.

Can we exploit temporal skew in a video to accelerate the recognition speed?



DominantObjectCount <= 10

Approach: Cascade oracle model with a less expensive "compact" model



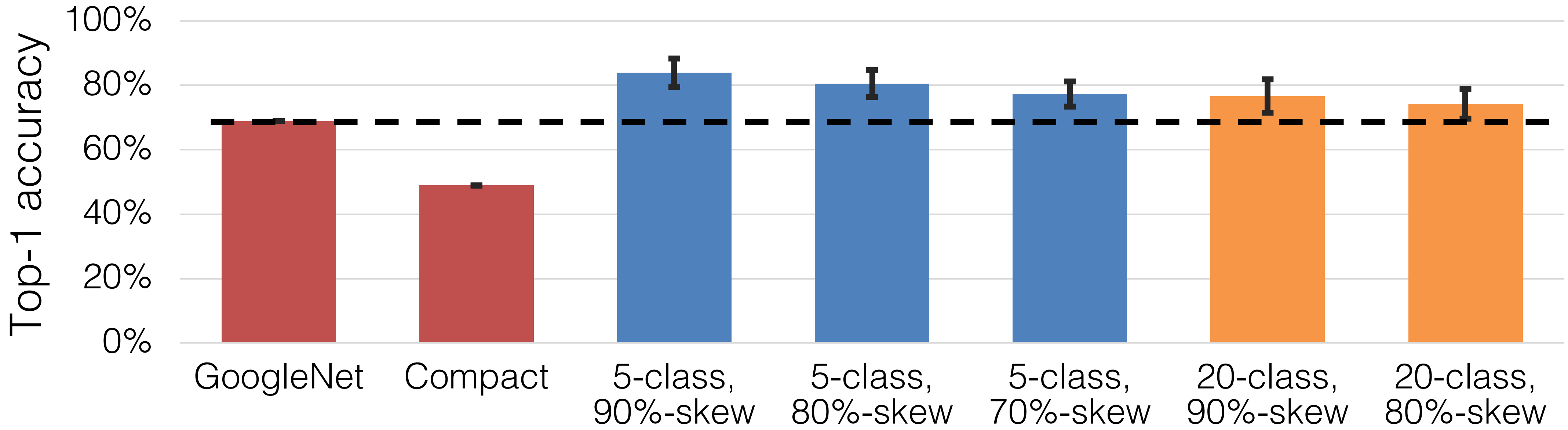
Challenges:

- Can specialized models have accuracy comparable to oracle models?
- Can we produce specialized models fast enough during runtime?
- How to determine when to switch specialized models without any ground truth data?

Specialized models have comparable accuracy under skewed distributions

Model	FLOPS	CPU lat.	GPU lat.
GoogLeNet (oracle)	3.17G	779 ms	11.0 ms
Compact CNN	0.82G	218 ms	4.4 ms

Object recognition (1000 classes)



Producing specialized models can be fast

- We pre-train the compact models on the full, unskewed datasets during development time.
- At the test time, fix the lower layers and only re-train the top fully connected layer of the compact model.
- Cache feature vectors of compact models for all inputs in the training datasets.

Generate the specialized model ~10 seconds.

Bandit-style algorithm to determine when to switch specialized models

- Oracle Bandit Problem
  - Exploration: use **the oracle model** to estimate the distribution.
  - Exploitation: use **a specialized model** to accelerate the recognition
- Windowed ε-Greedy (WEG) Algorithm
  - Adaptively select the windows size for sampling.
  - Produce a specialized model when a skew is detected.
  - Use heuristics to detect skew changes while “exploiting” specialized models.

Evaluation

video	length (min)	oracle		WEG	
		acc. (%)	GPU lat. (ms)	acc. (%)	GPU lat. (ms)
Friends	24	93.2	28.97	93.5	7.0 **(x4.1)**
Good Will Hunting	14	97.6	28.84	95.1	3.7 **(x7.8)**
Ellen Show	11	98.6	29.26	94.6	4.7 **(x6.2)**
The Departed	9	93.9	29.18	93.5	6.9 **(x4.2)**
Ocean’s Eleven / Twelve	6	97.9	28.97	96.0	12.3 **(x2.4)**