Dynamic Neural Networks

Joseph E. Gonzalez
Co-director of the RISE Lab
jegonzal@cs.berkeley.edu

What is the Problem Being Solved?

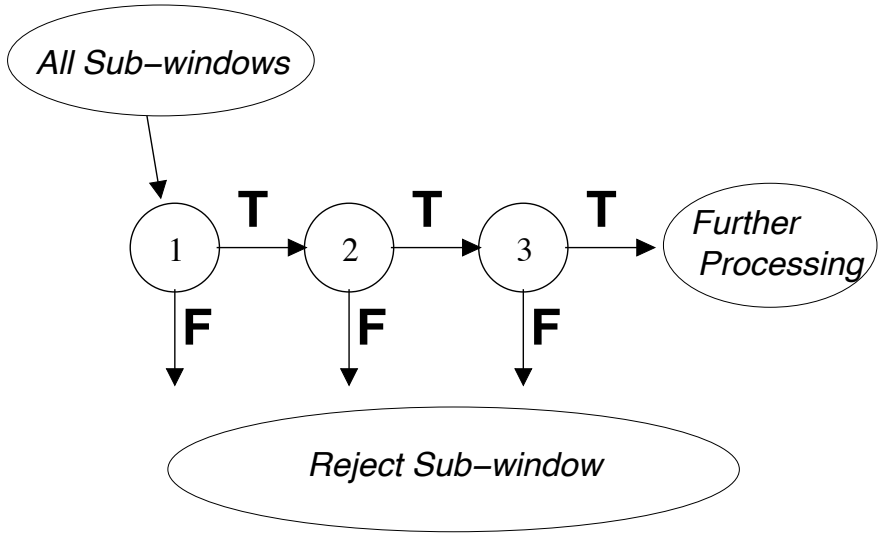
- Neural network computation increasing rapidly
- Larger networks are needed for peak accuracy
- Big Ideas:
  - Adaptively scale computation for a given task
  - Select only the parts of the network needed for a given input

Early Work: Prediction Cascades

- **Viola-Jones** Object Detection Framework (2001):
  - "Rapid Object Detection using a Boosted Cascade of Simple Features" **CVPR'01**
  - Face detection on 384x288 at 15 fps (700MHz Pentium III)

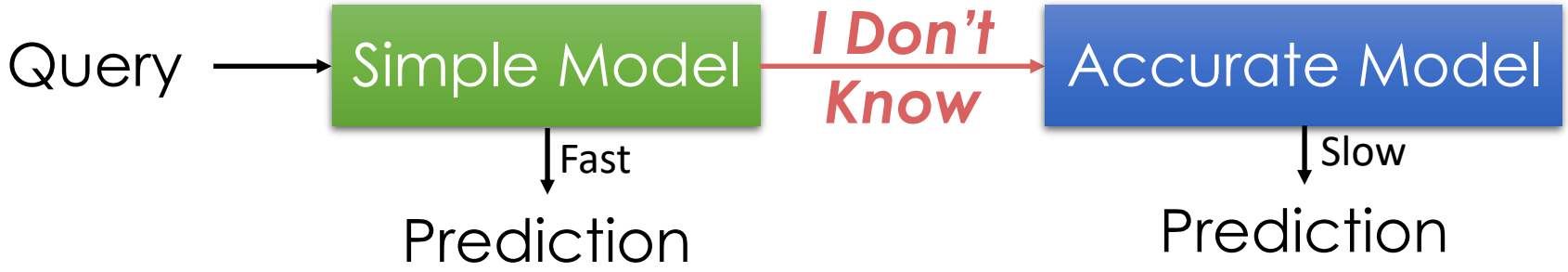
Most parts of the image don't contain a face.

Reject those regions quickly.

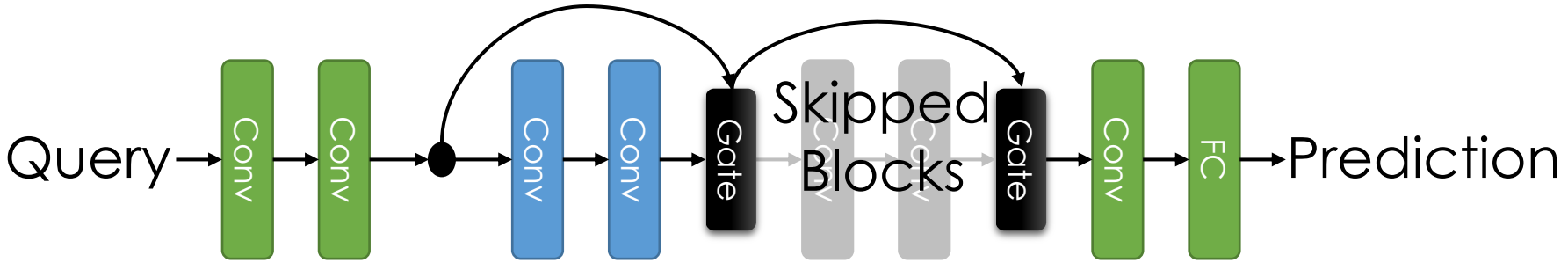


Dynamic Networks for **fast** and **accurate** inference

IDK Cascades: Using the fastest model possible [UAI'18]



SkipNet: dynamic execution within a model [ECCV'18]



Task Aware Feature Embeddings
[CVPR'19]

Feature Network

FF Net

FF Net

FF Net

Baby

Params

Params

Params

X

X

X

Emb. Network

FC Layer

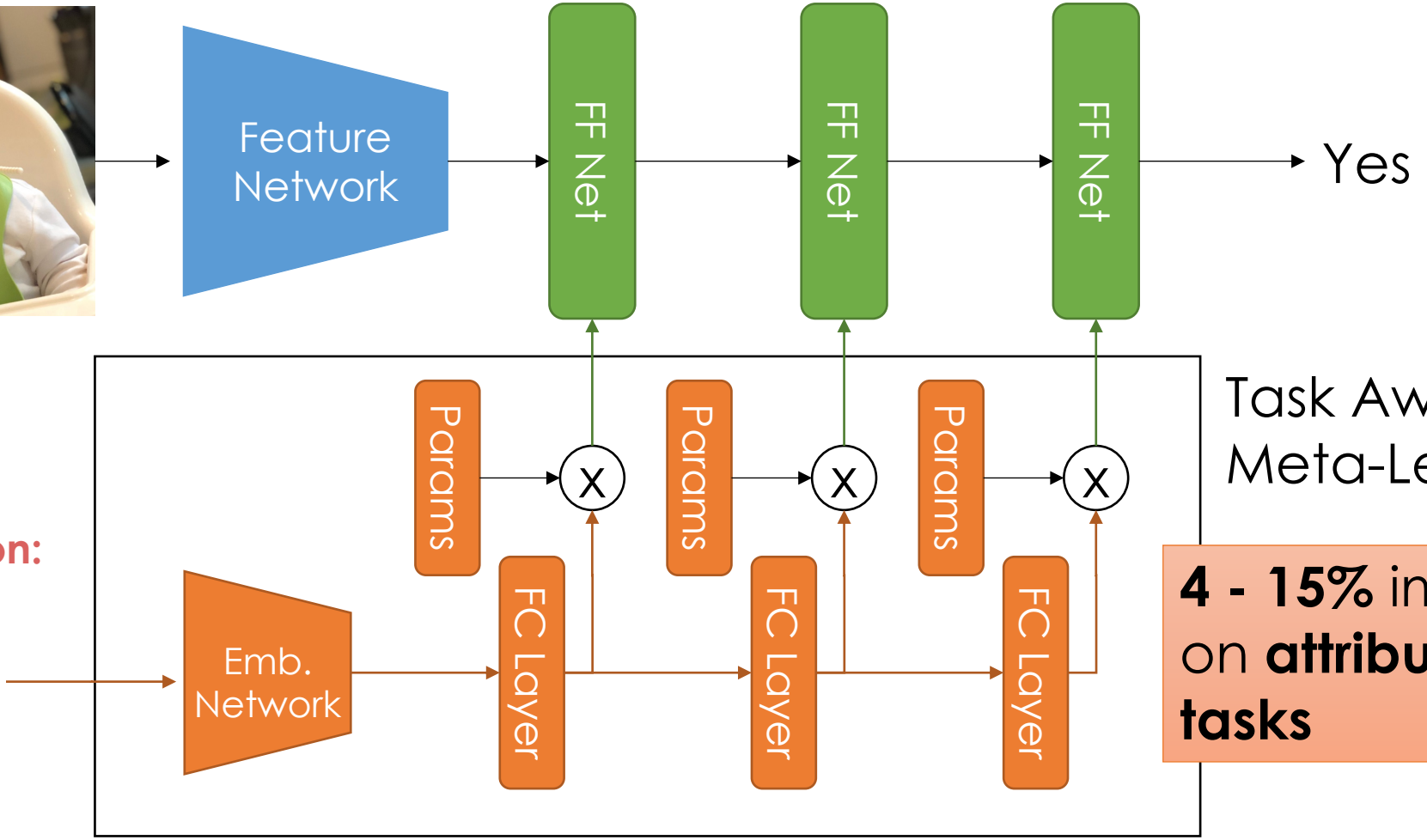
FC Layer

FC Layer

Task Aware Meta-Learner

More **accurate** and **efficient** than existing dynamic pruning networks

Task Aware Feature Embeddings
[CVPR'19]



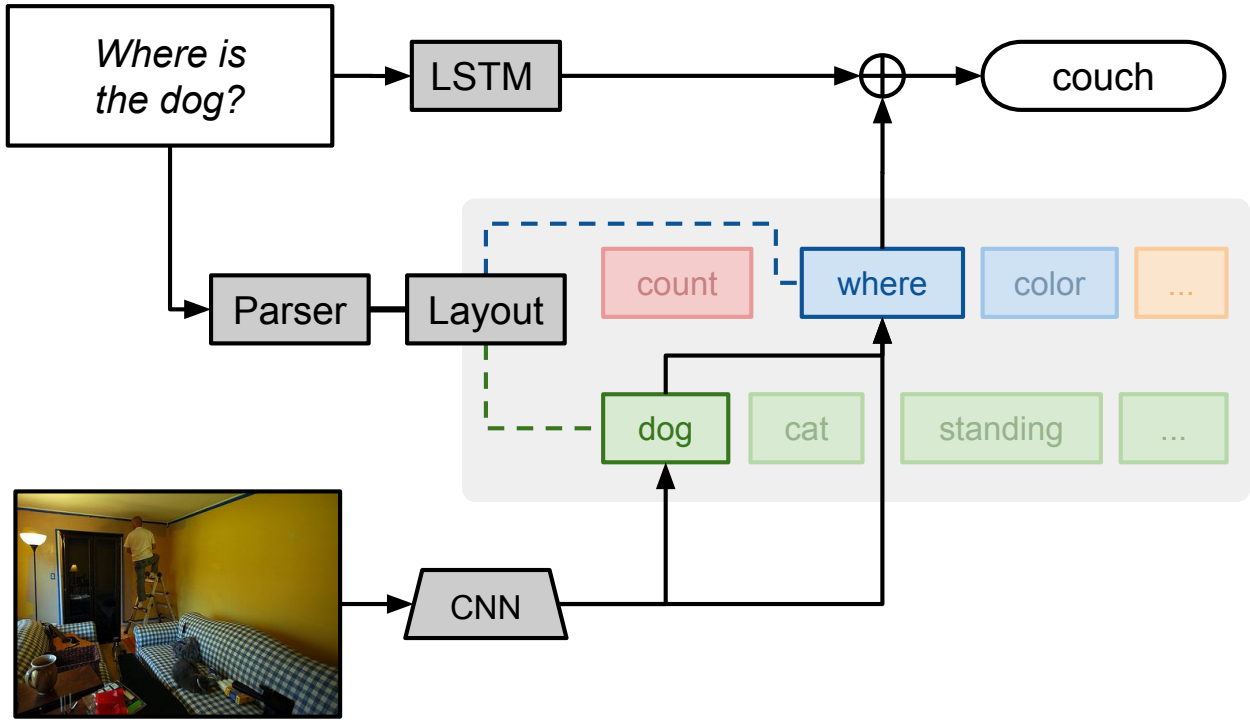
Task Description:

"Smiling Baby"

4 - 15% improvement on **attribute-object tasks**

Neural Modular Networks

Jacob Andreas et al., "*Deep Compositional Question Answering with Neural Module Networks*"



Trends Today

- Multi-task Learning to solve many problems
  - Zero-shot learning
- Adjust network architecture for a given query
  - Neural Modular Networks
  - Capsule Networks
- Language models … more on this in future lectures
  - Why are these dynamic? How does computation change with input?

Dynamic Networks → Systems Issues

- Reduce computation but do they reduce runtime?
  - Limitations in existing evaluations?
- Implications on hardware executions?
- Challenges in expressing dynamic computation graphs…
- Likely to be the future of network design?
  - Modularity …