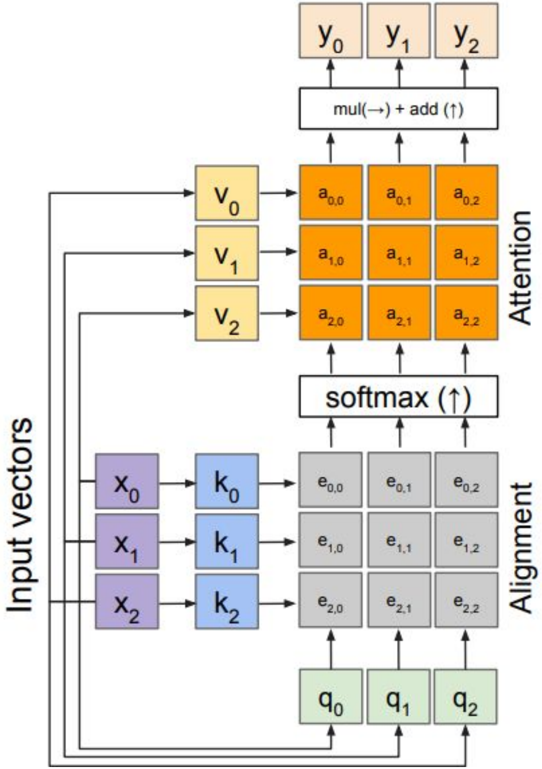
LSTMs and Transformers

CS 231N

Why Transformers?

Self-Attention is a unique and powerful operation!

Recap: What is self attention?



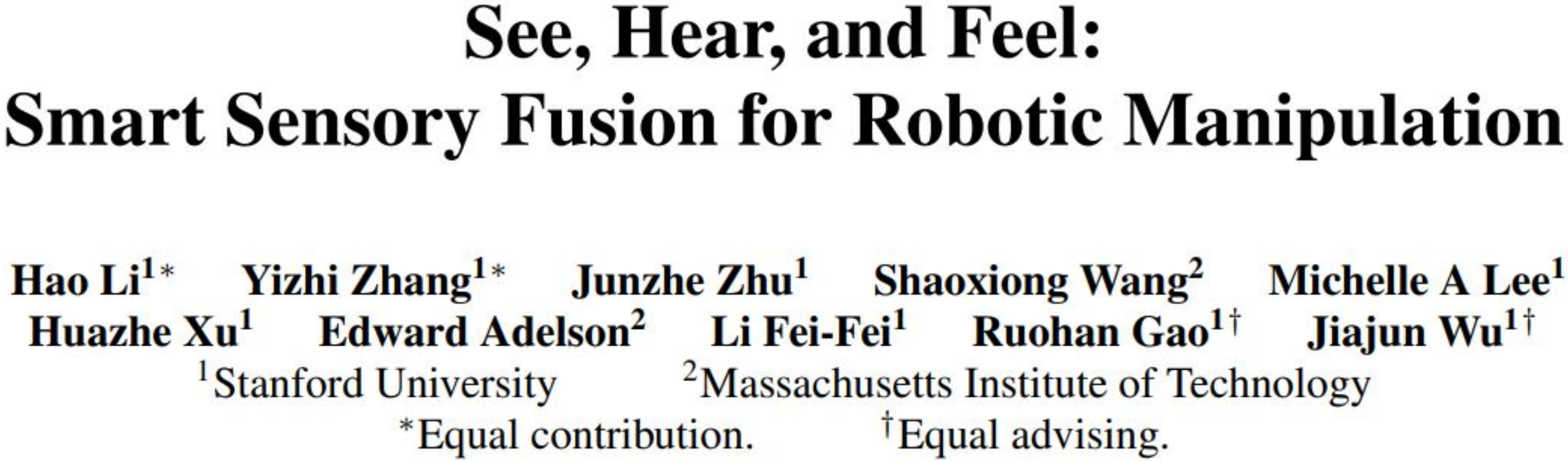
Outputs:
context vectors: **y** (shape: D_v)

Operations:
Key vectors: $\mathbf{k} = \mathbf{x}\mathbf{W}_k$
Value vectors: $\mathbf{v} = \mathbf{x}\mathbf{W}_v$
Query vectors: $\mathbf{q} = \mathbf{x}\mathbf{W}_q$
Alignment: $e_{i,j} = q_j \cdot k_i / \sqrt{D}$
Attention: $\mathbf{a} = \text{softmax}(\mathbf{e})$
Output: $y_j = \sum_i a_{i,j} v_i$

Inputs:
Input vectors: **x** (shape: N x D)

Self-attention allows the network to learn relationships between each segment of the input!

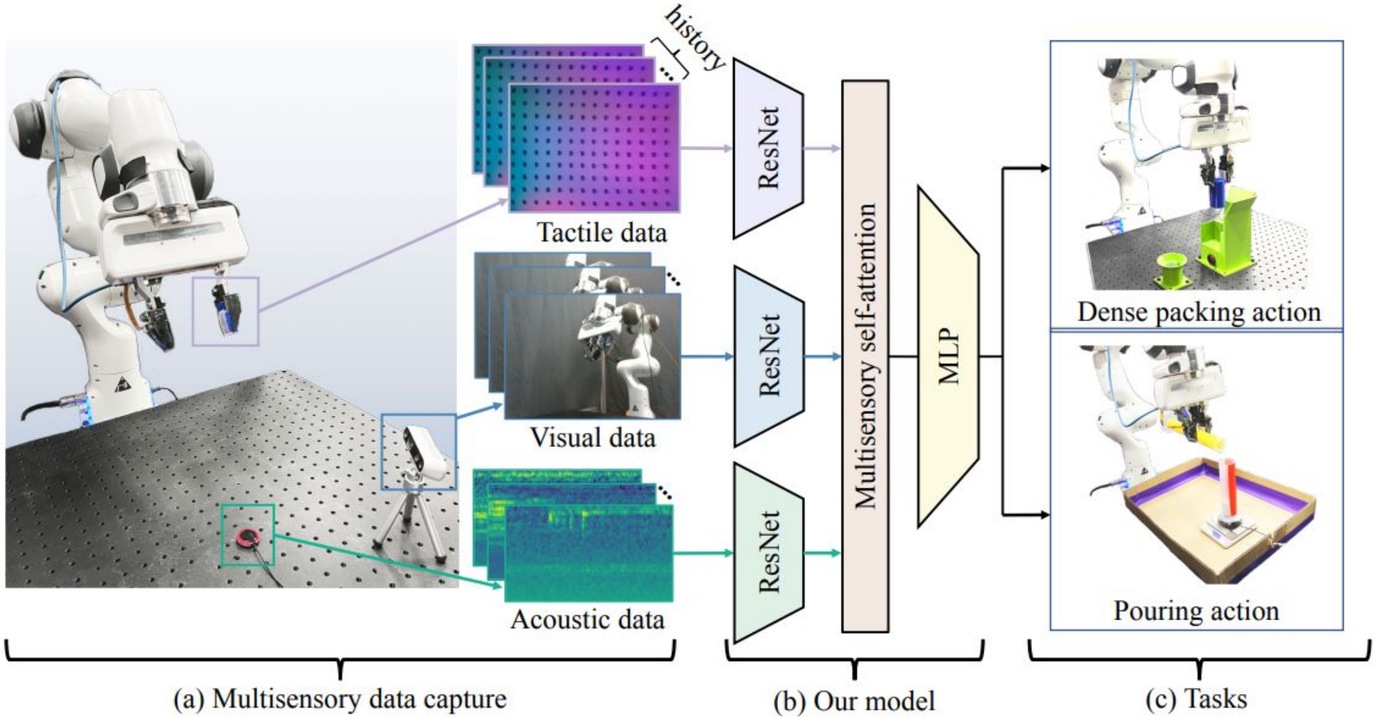
Interesting Application



See, Hear, and Feel: Smart Sensory Fusion for Robotic Manipulation

Hao Li[1]* **Yizhi Zhang**[1]* **Junzhe Zhu**[1] **Shaoxiong Wang**[2] **Michelle A Lee**[1]
Huazhe Xu[1] **Edward Adelson**[2] **Li Fei-Fei**[1] **Ruohan Gao**[1]† **Jiajun Wu**[1]†
[1]Stanford University [2]Massachusetts Institute of Technology
*Equal contribution. †Equal advising.

Overall Setup



Multisensory Self-Attention

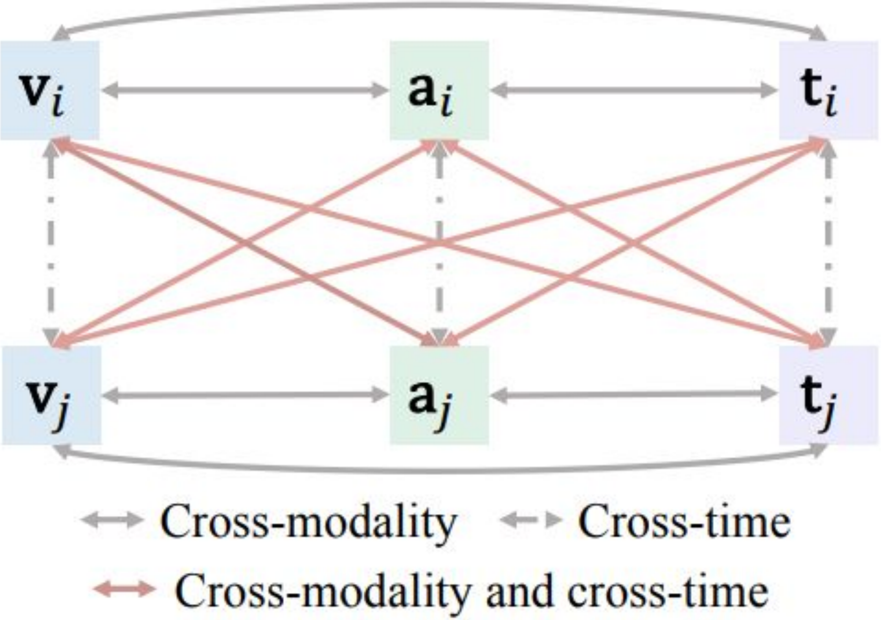
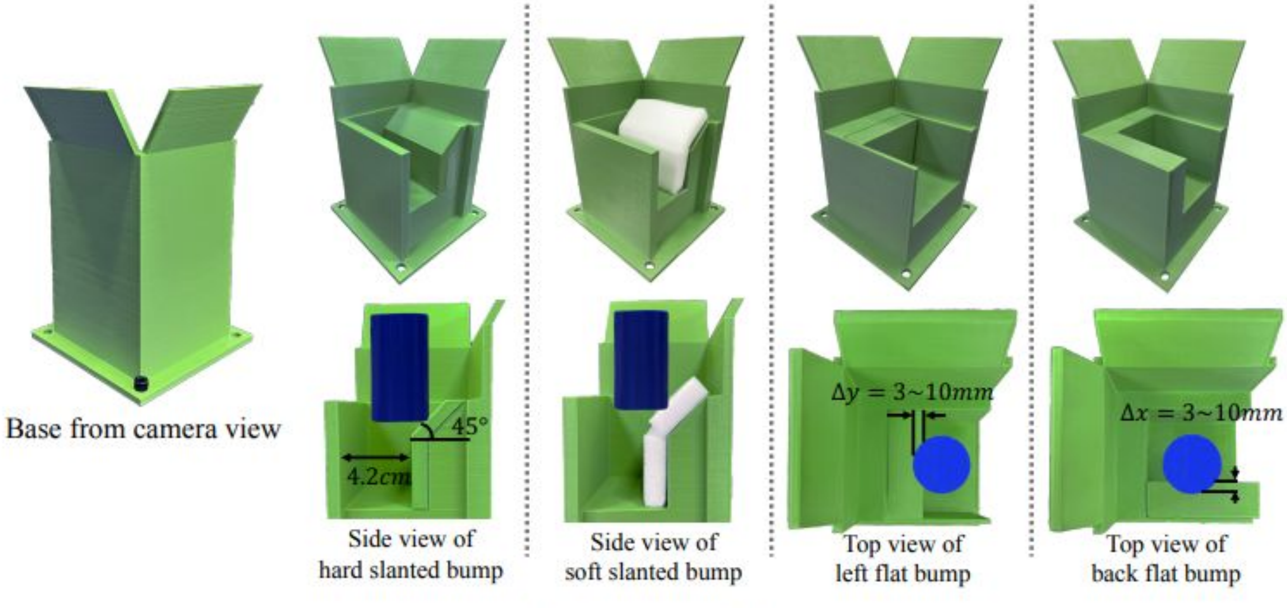


Figure 3: Multisensory self-attention.

Dense Packing



Base from camera view

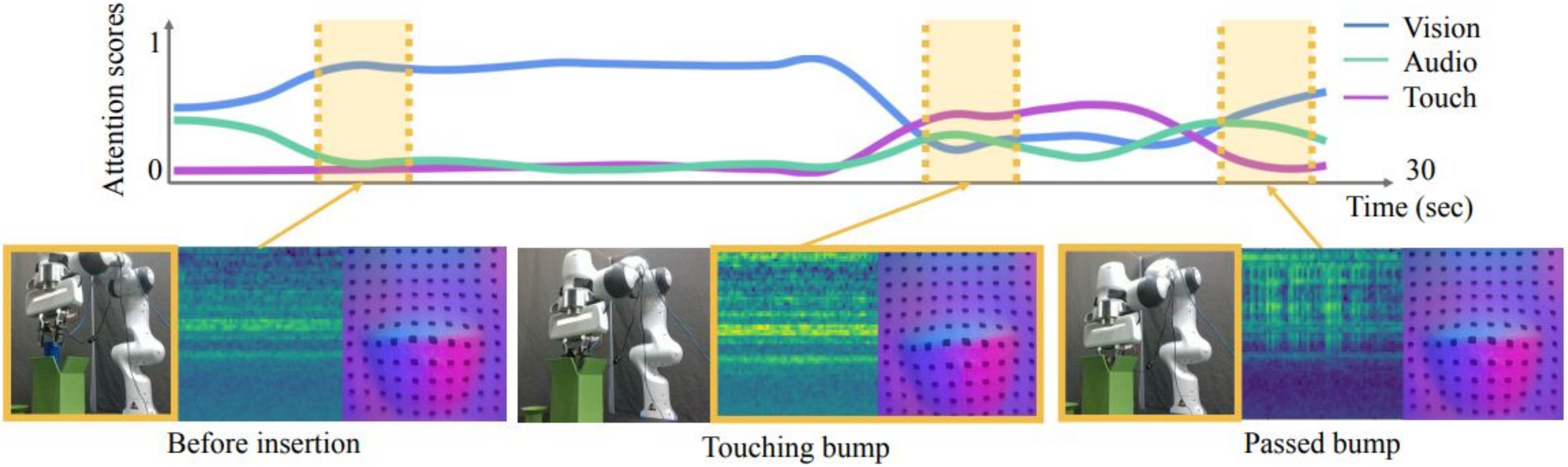
Side view of hard slanted bump

Side view of soft slanted bump

Top view of left flat bump

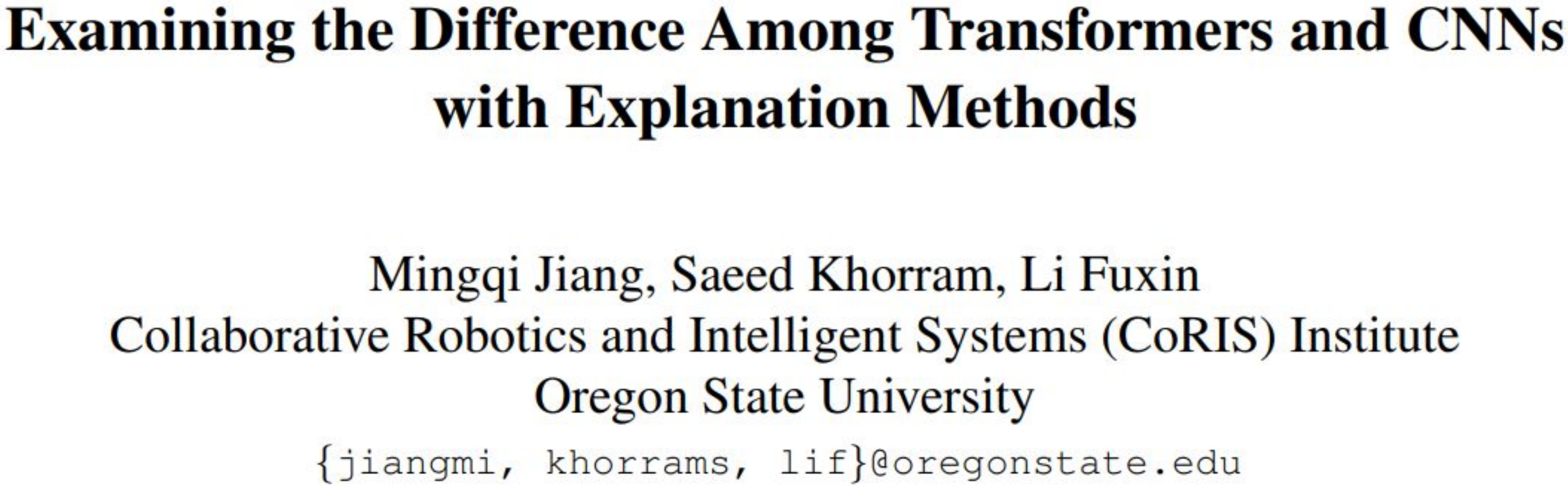
Top view of back flat bump

Tasks Evaluated: Dense Packing



Real World Example:

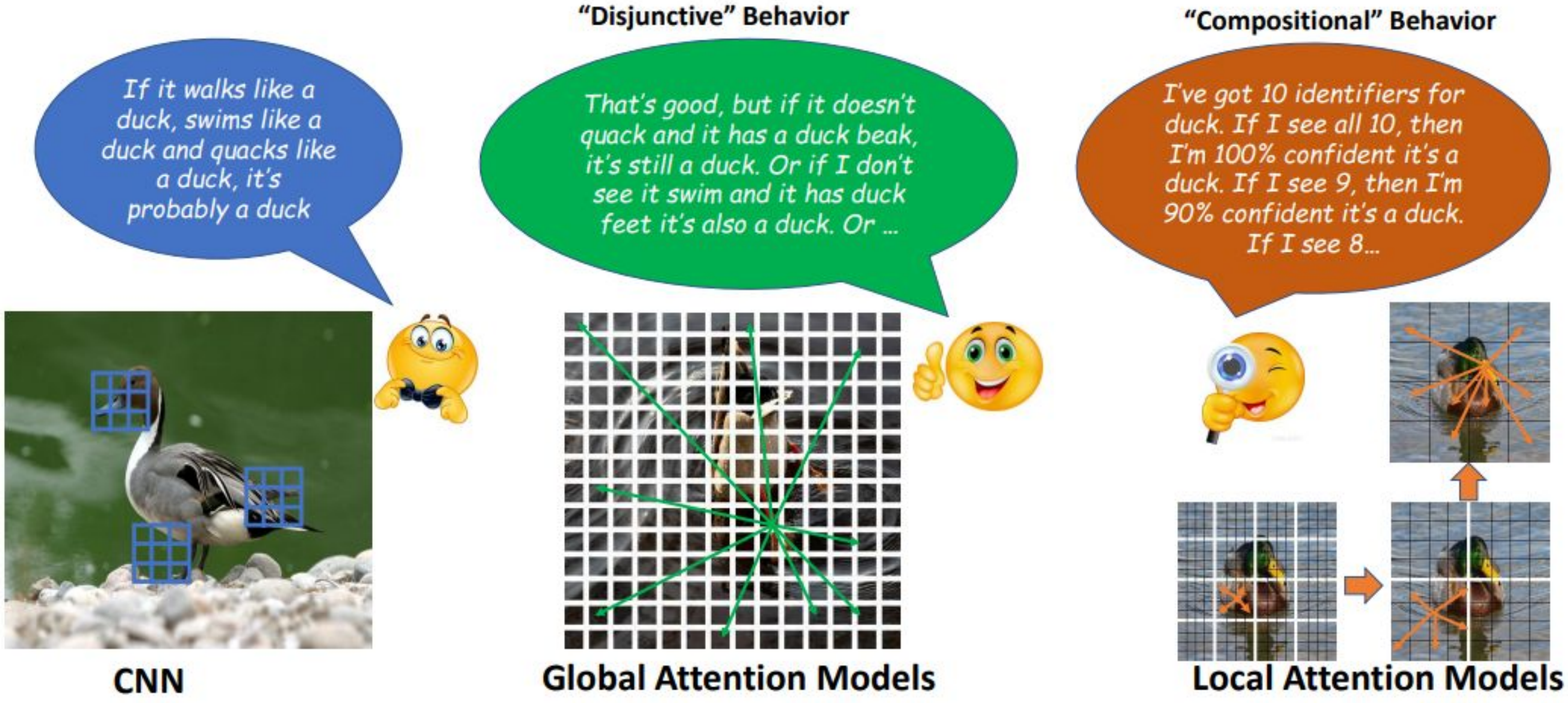
CNN vs Transformer



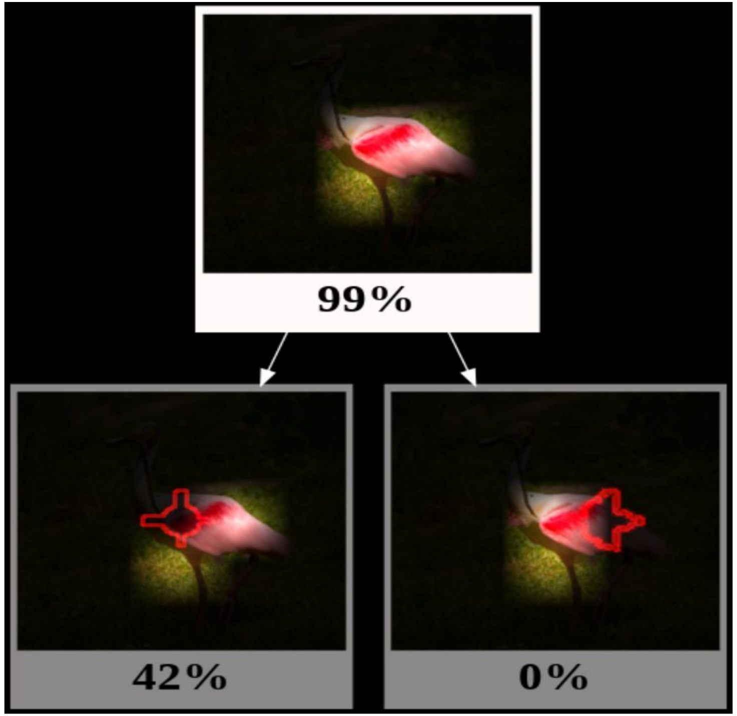
Examining the Difference Among Transformers and CNNs with Explanation Methods

Mingqi Jiang, Saeed Khorram, Li Fuxin
Collaborative Robotics and Intelligent Systems (CoRIS) Institute
Oregon State University
{jiangmi, khorrams, lif}@oregonstate.edu

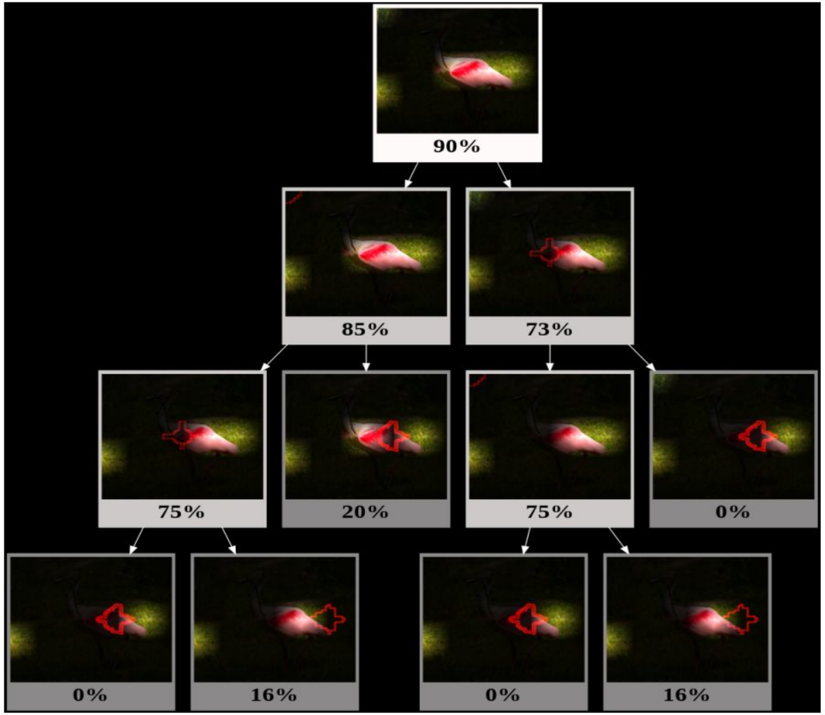
CNN vs Transformer


"Disjunctive" Behavior
"Compositional" Behavior
If it walks like a duck, swims like a duck and quacks like a duck, it's probably a duck
That's good, but if it doesn't quack and it has a duck beak, it's still a duck. Or if I don't see it swim and it has duck feet it's also a duck. Or ...
I've got 10 identifiers for duck. If I see all 10, then I'm 100% confident it's a duck. If I see 9, then I'm 90% confident it's a duck. If I see 8...
CNN
Global Attention Models
Local Attention Models

Results



(b) ResNet-50-C2



(d) Swin-T