Inference Graphs for CNN Interpretation

Yael Konforti, Alon Shpigler, Boaz Lerner, Aharon Bar Hillel




Ben-Gurion University of the Negev, Beer Sheva, Israel

{yaelkonf, alonshp}@post.bgu.ac.il; {boaz, barhille}@bgu.ac.il

Code: https://github.com/yaelkon/GMM-CNN

Motivation for CNN interpretation

The current state –

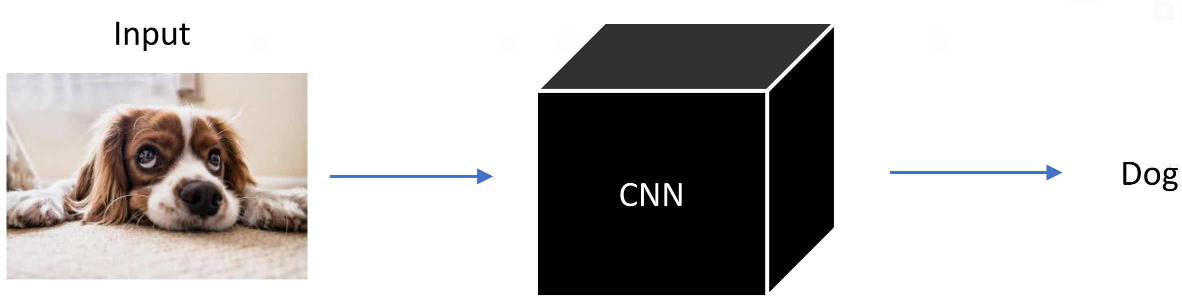
State-of-the-art results for a variety of computer domains.

The problem –

Reasoning of their decision-making process is lacking.

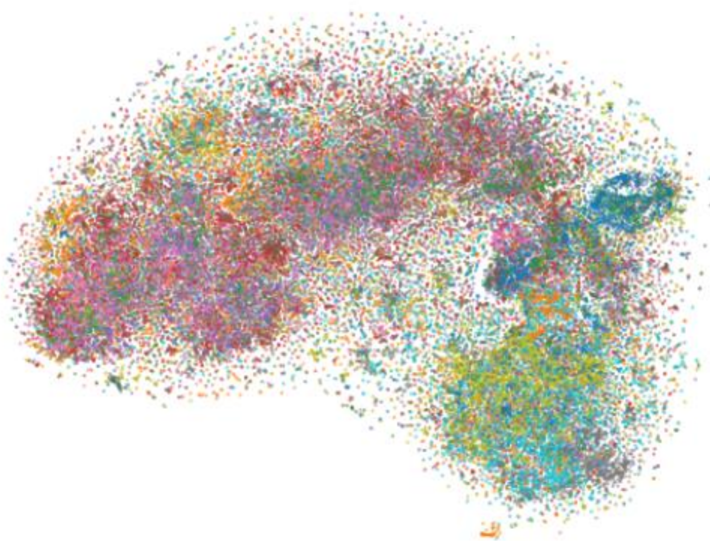
Our main challenge –

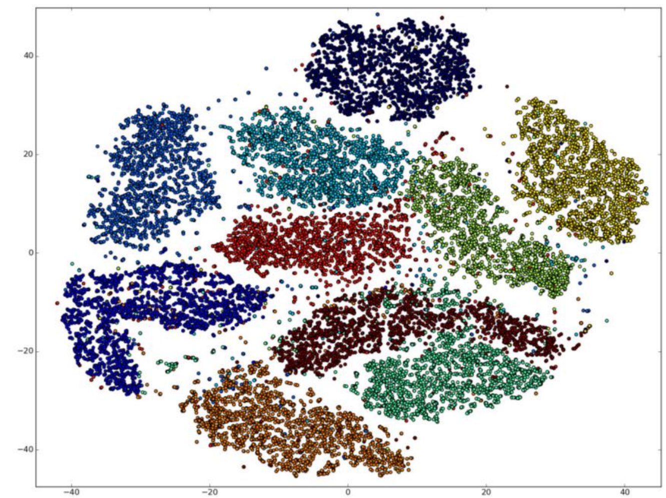
Improving interpretability of the network inference process.



Our challenges

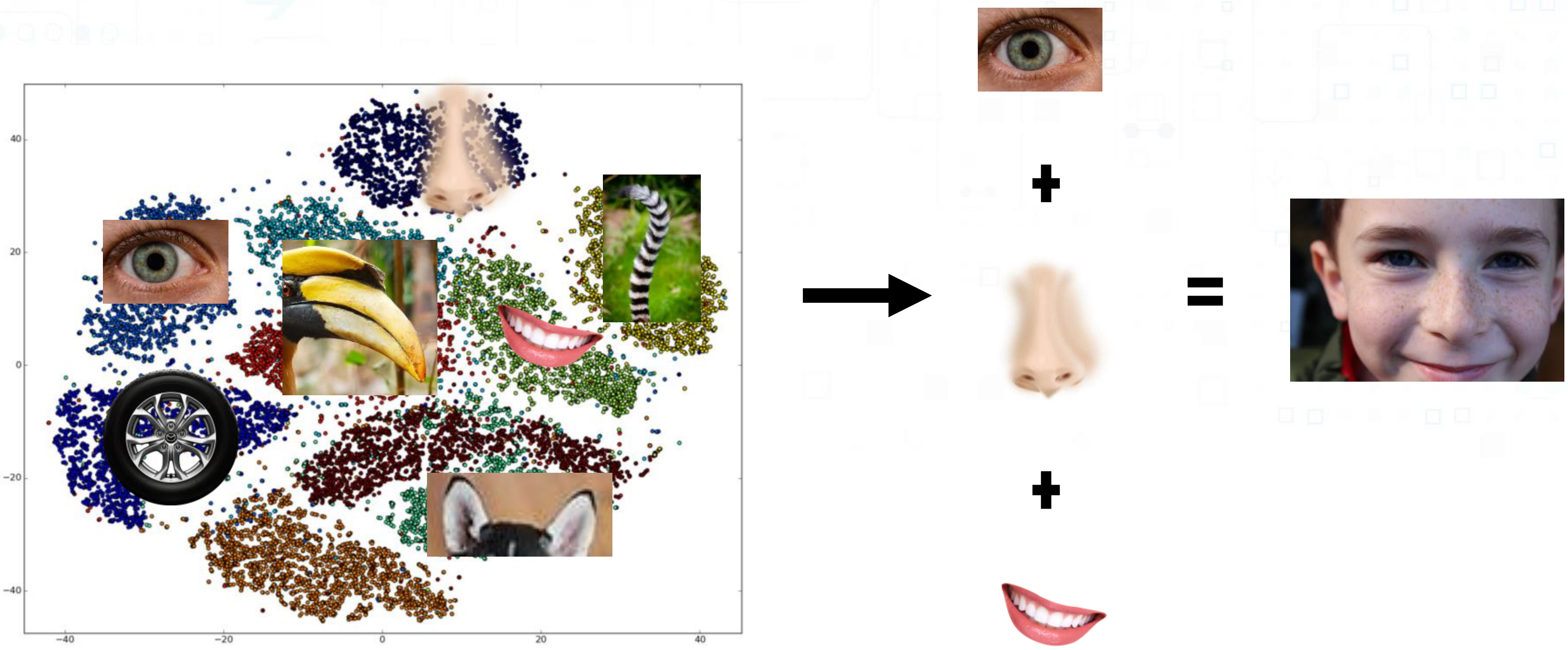
Can we convert distributed representations into a human-oriented language?





Our challenges

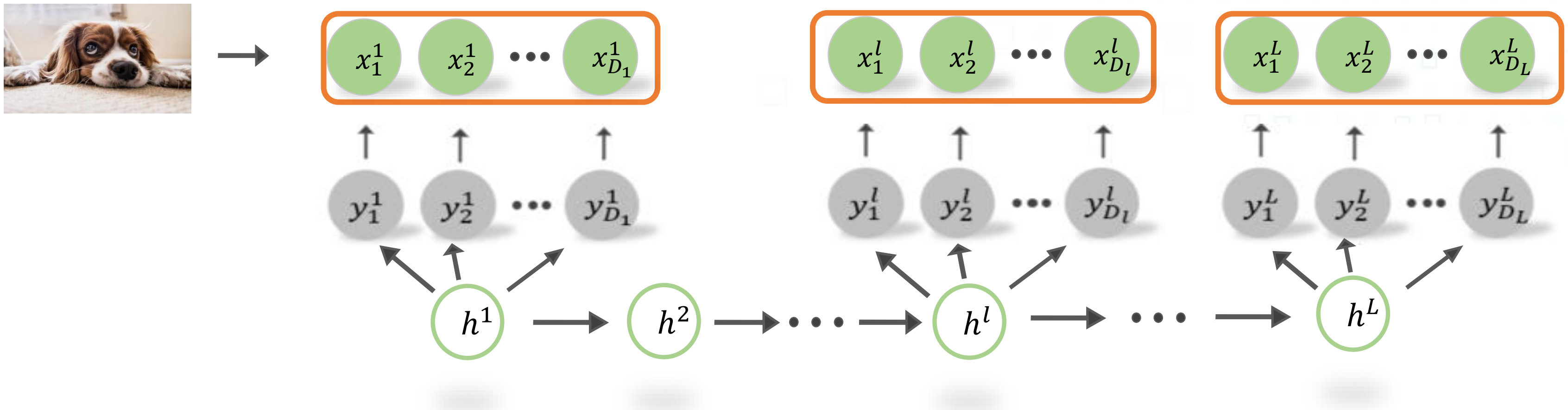
Can we learn a dictionary of visual words and model their interrelationships?



Generative modeling of MLPs

Trained MLP inference process as a Hidden Markov Model (HMM):

- **Single MLP layer distribution**: a mixture model of multivariate Gaussians (GMM).
- **Connections between layer representations**: conditional probability tables between GMM components of adjacent layers.
- **Post-Relu activations**: rectified Gaussians via additional hidden variables.
- **Optimization:** online Expectation Maximization (EM).

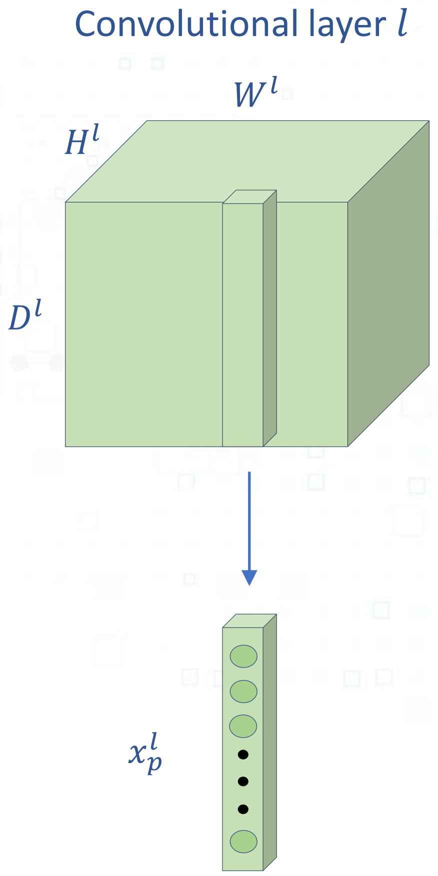


Layers Dictionaries for CNNs

- Each spatial location example vector x_p^l , located at $p = (i, j) \in \{\{1, \dots, H^l\} \times \{1, \dots, W^l\}\}$, is described as arising from a GMM of K^l components.

- Example x_p^l is assigned to cluster C_k^l iff $P(h_p^l = k | x_p^l)$ is the highest.

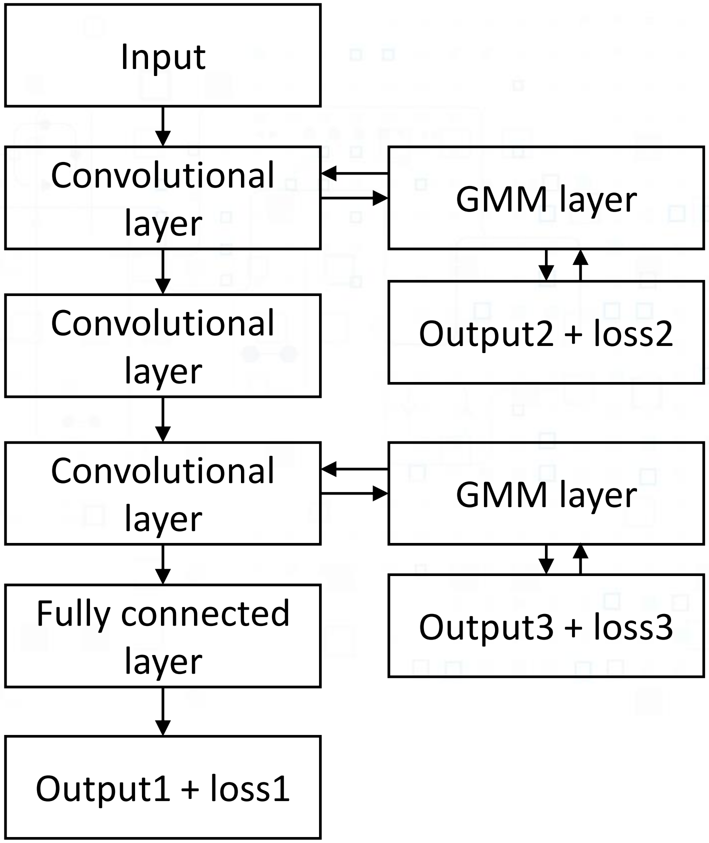
- Each cluster C_k^l represents a **visual word** (edge, texture, body part etc.), together forming the layer **dictionary**.



Layers Dictionaries for CNNs

Train a CNN-GMM model:

1. Get a pre-trained CNN.
2. For every modeled layer l, append to it a GMM layer with the GMM parameters $\Theta^l = \{\pi_k^l, \mu_k^l, \Sigma_k^l\}_{k=1}^{K^l}$ as learning weights.
3. Train all GMM layers and estimate their parameters independently, using SGD.



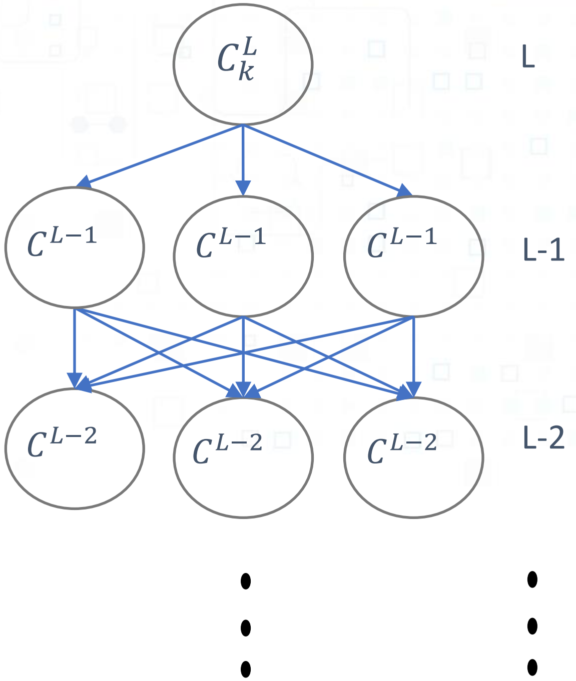
Graph Node Selection Algorithm

Consider a graph in which visual words $\{C_k^l\}_{l=1,k=1}^{L,K^l}$ are the nodes, and transition probabilities between visual words of consecutive layers quantify edges.

- Given a selected subset of images to be explained $\Omega = \{I_n\}_{n=1}^{N}$ (e.g., class or a single image), a specific subgraph can have high explanatory value.

How can we find the most explanatory visual words?

- **Node selection algorithm** – iterative algorithm starting from "explaining" the classification decision node, then explaining layers backward, until outputting a subgraph.



Graph Node Selection Algorithm

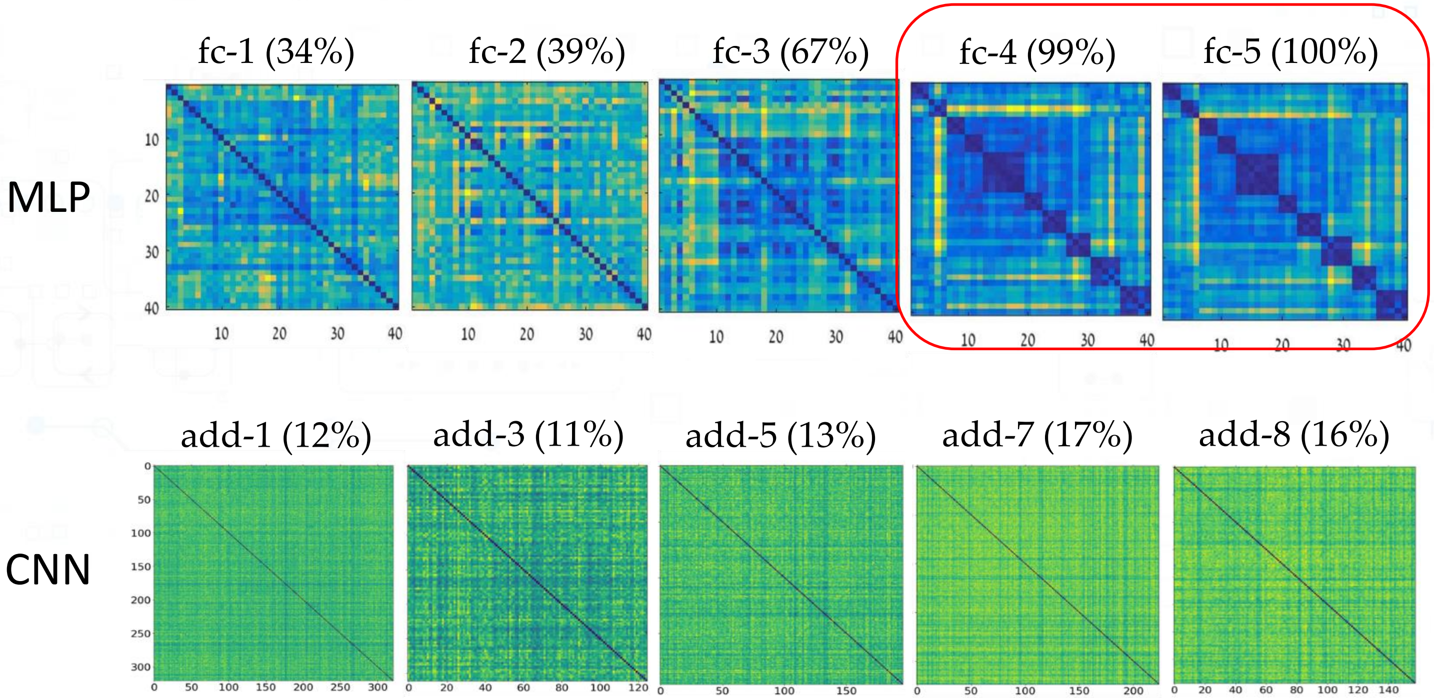
Given an instance of a single visual word $h_p^l = s$ to "explain", we look for the visual words T in its receptive field $R(p)$ most contributing to its likelihood:

$$\max_{T,|T|=Z} \log P\left(h_p^l = s | h_q^{l'}: q \in R(p), h_q^{l'} \in T\right)$$

Using location independence assumptions, a 't –explains–s' score is derived:

$$S^{l'}(s,t) = \sum_{t=1}^{K^{l'}} \underbrace{\left|\left\{q: h_q^{l'} = t, q \in R(p)\right\}\right|}_{\text{the number of times visual word } t \text{ appears in the receptive field of location } p} \log \underbrace{\frac{P\left(h_q^{l'} = t | h_p^l = s, q \in R(p)\right)}{P\left(h_q^{l'} = t\right)}}_{\text{how likely it is to see word } t \text{ in the receptive field of } s \text{ compared to seeing it in general}}$$

Cluster Similarity Across Layers



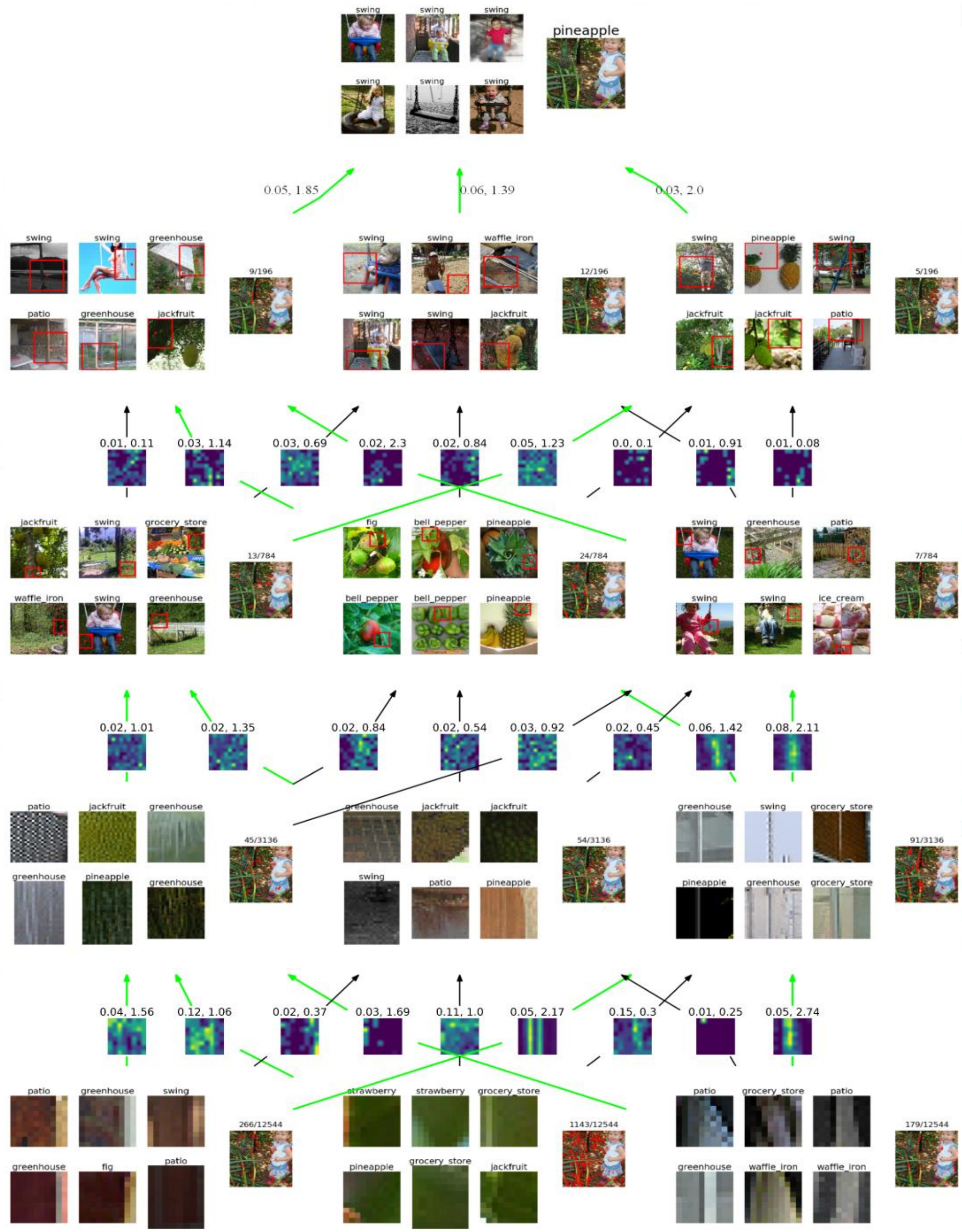
increasing similarity between clusters representing the same class.

the clusters stay local and diverse, even at the uppermost layers.

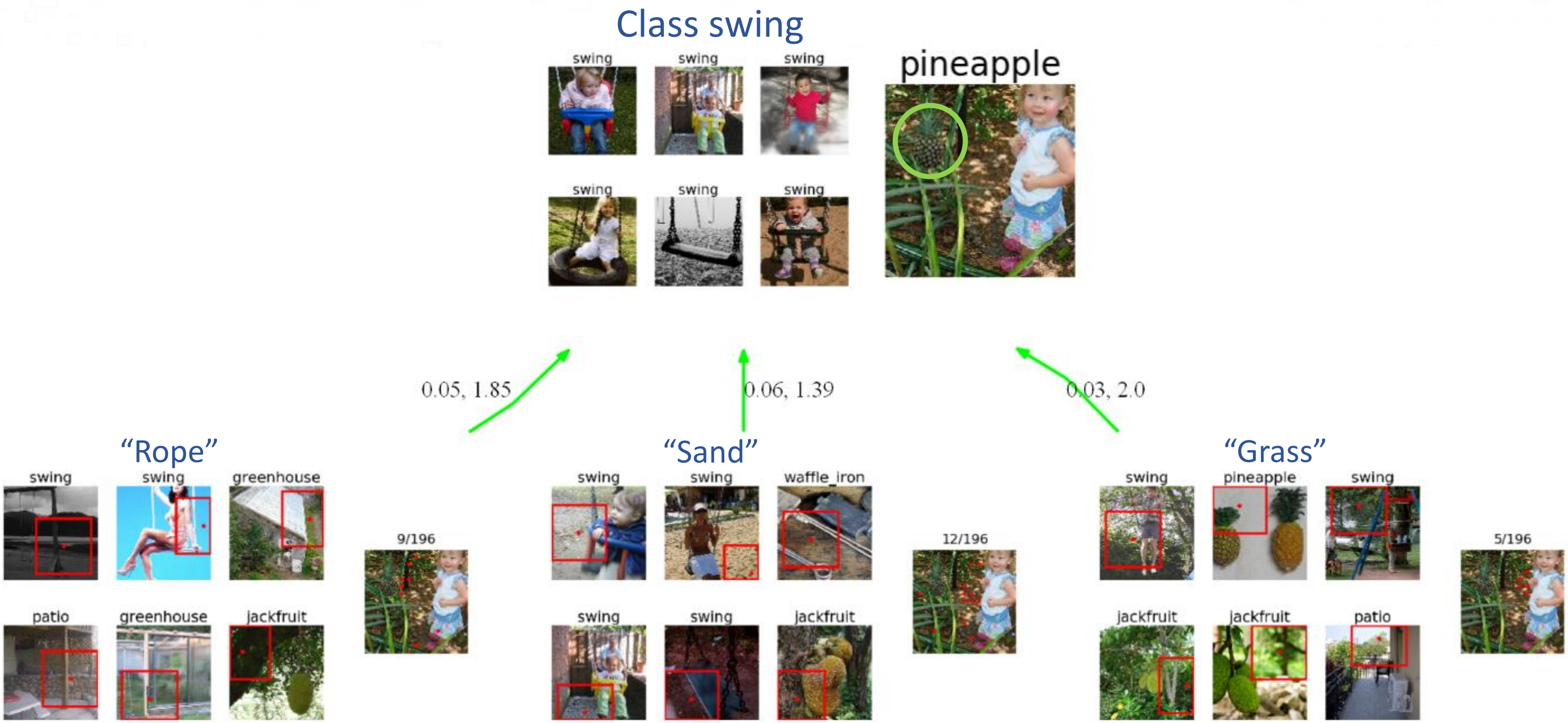
VGG-16 Image Inference Graph

Ground truth: Pineapple

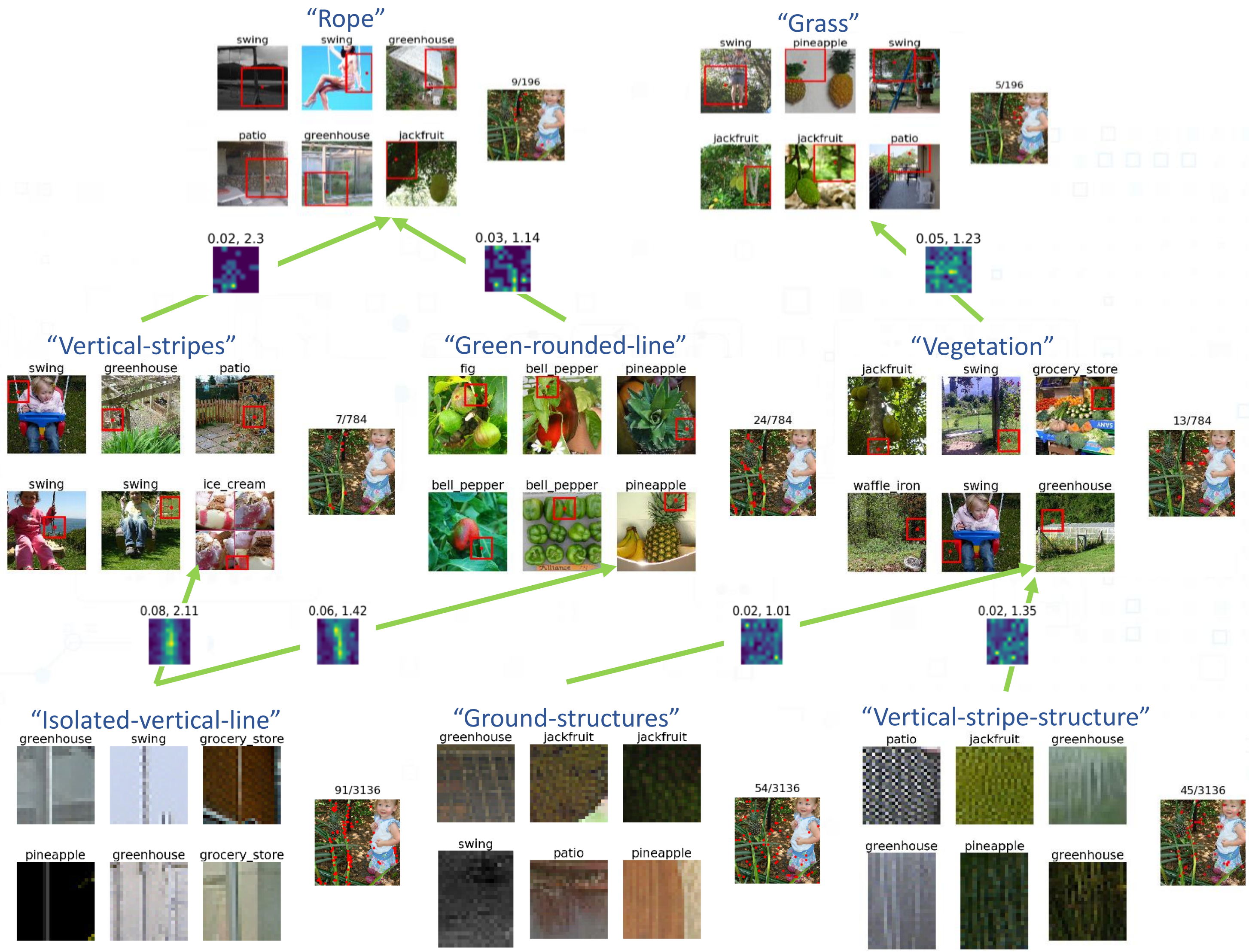
Network prediction: Swing



VGG-16 Image Inference Graph



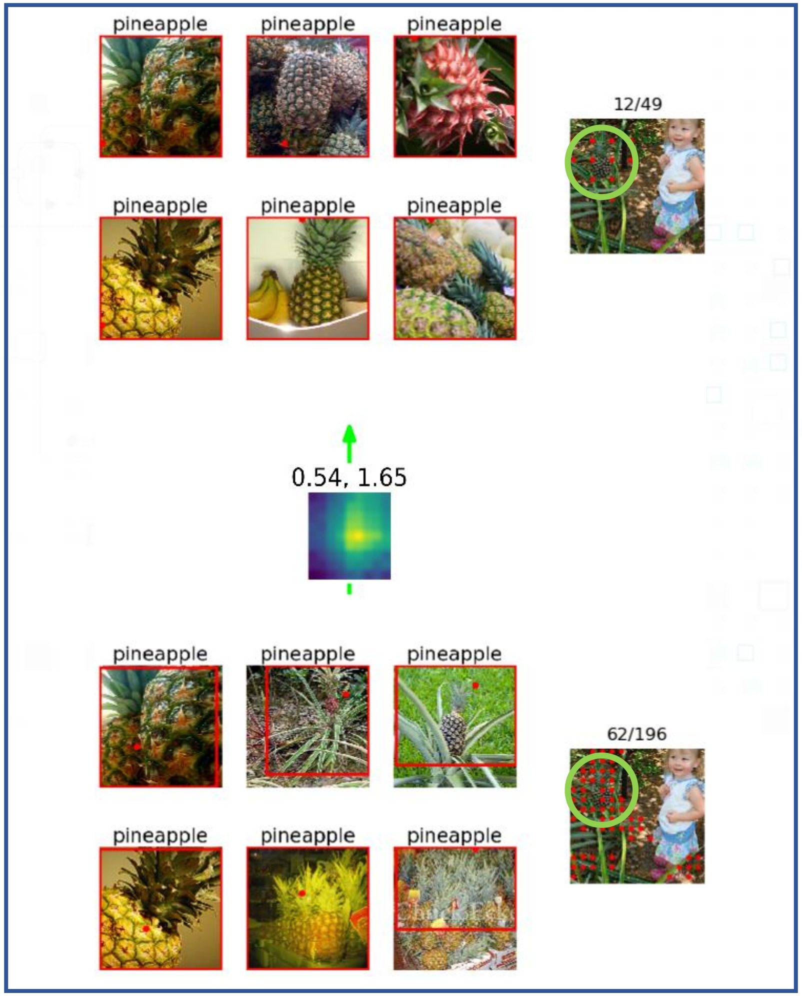
ECCV'20 ONLINE
23-28 AUGUST 2020



ResNet50 Image Inference Graph

Ground truth: Pineapple

Network prediction: Pineapple



Inference Graphs for CNN Interpretation

Yael Konforti, Alon Shpigler, Boaz Lerner, Aharon Bar Hillel




Ben-Gurion University of the Negev, Beer Sheva, Israel

{yaelkonf, alonshp}@post.bgu.ac.il; {boaz, barhille}@bgu.ac.il

Code: https://github.com/yaelkon/GMM-CNN