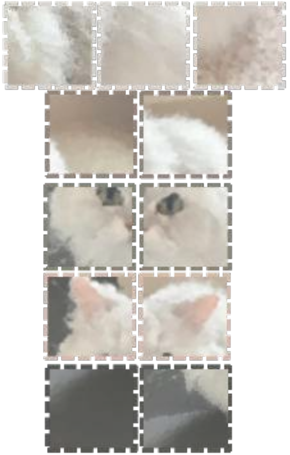
Deep Representations

Recap: Structure of Images

Images

- Reoccuring patterns
- Patterns at various scales

Image: cat



Repeating patterns
with 30 x 30 patches

Recap: Structure of Images

Images

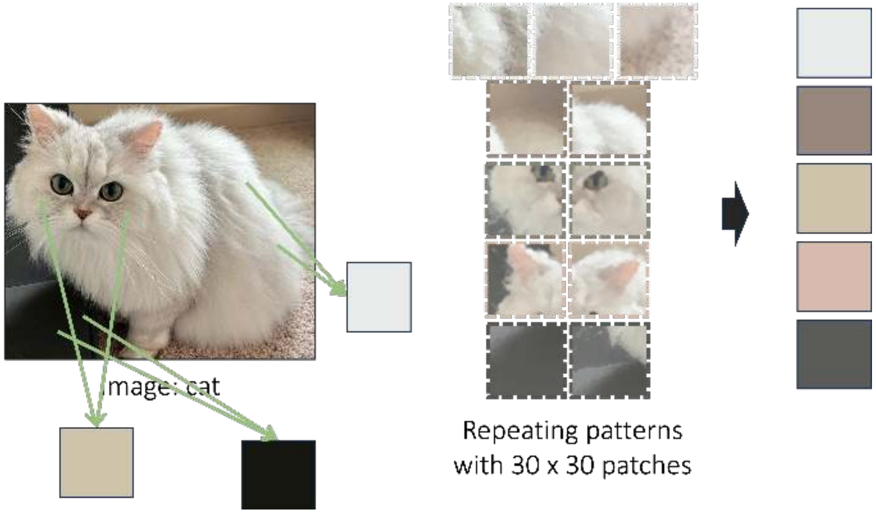
- Reoccuring patterns
- Patterns at various scales

Local Invariance

- Nearby pixels are likely similar values

Semantic Patterns

- Pixels within an "entity" are similar



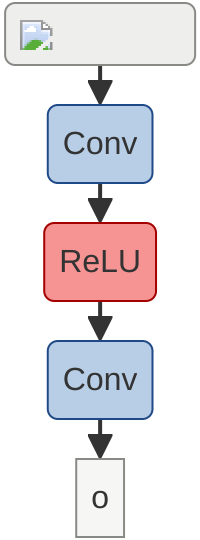
How Do ConvNets See Images?

Option 1: Look at activations [1]

- What patterns most excite a specific activation?

Option 2: What does the network look at for decisions?

- What output is most discriminative? [2]
- What pixels have highest influence on output? [3]

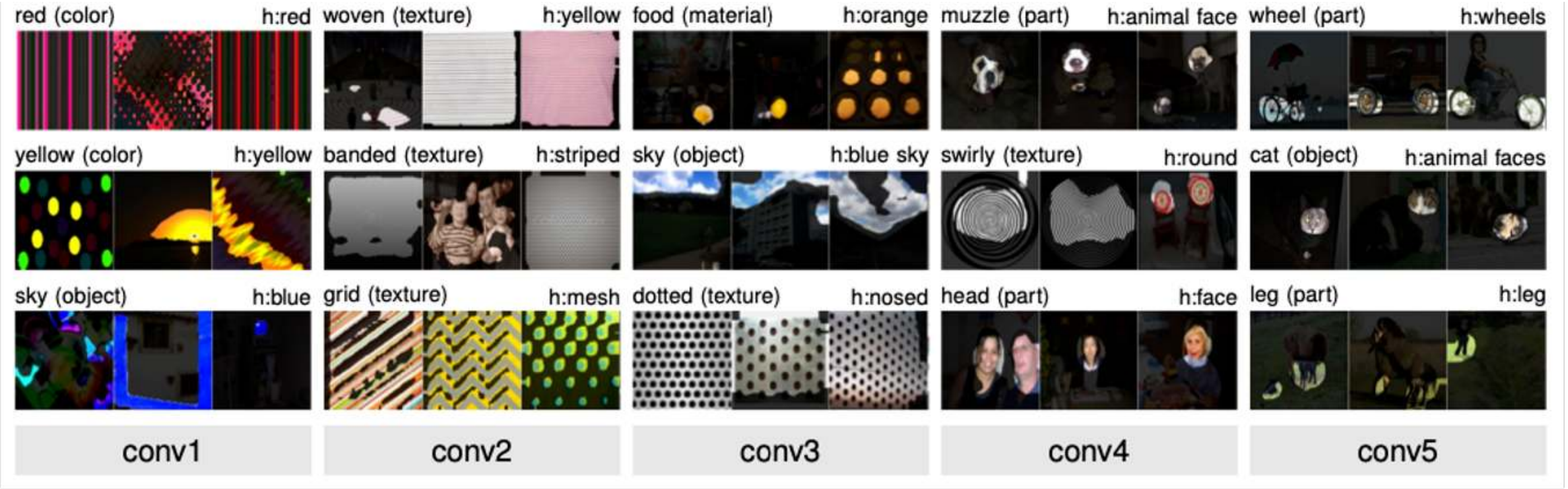


[1] D. Bau et al., "Network Dissection: Quantifying Interpretability of Deep Visual Representations", CVPR 2017 ↩

[2] B. Zhou et al., "Learning Deep Features for Discriminative Localization", CVPR 2016 ↩

[3] R. Selvaraju et al., "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization", CVPR 2017 ↩

Visualizing Activations



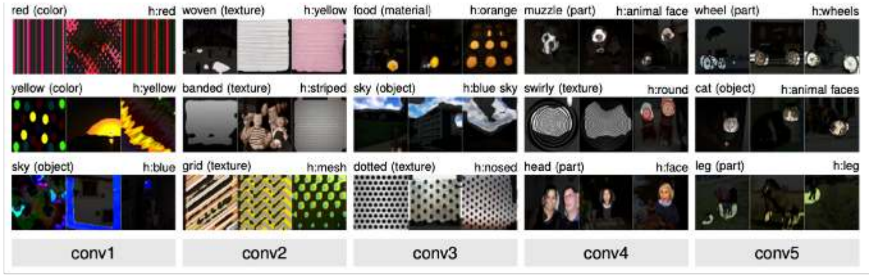
Interpretable hidden units

Visualizing Activations

Record all activations for layer i, channel c

- Pick the top-p percentile ($p \approx 0.005$)
- Look at activation map
- Correlate unit activation with concept / class

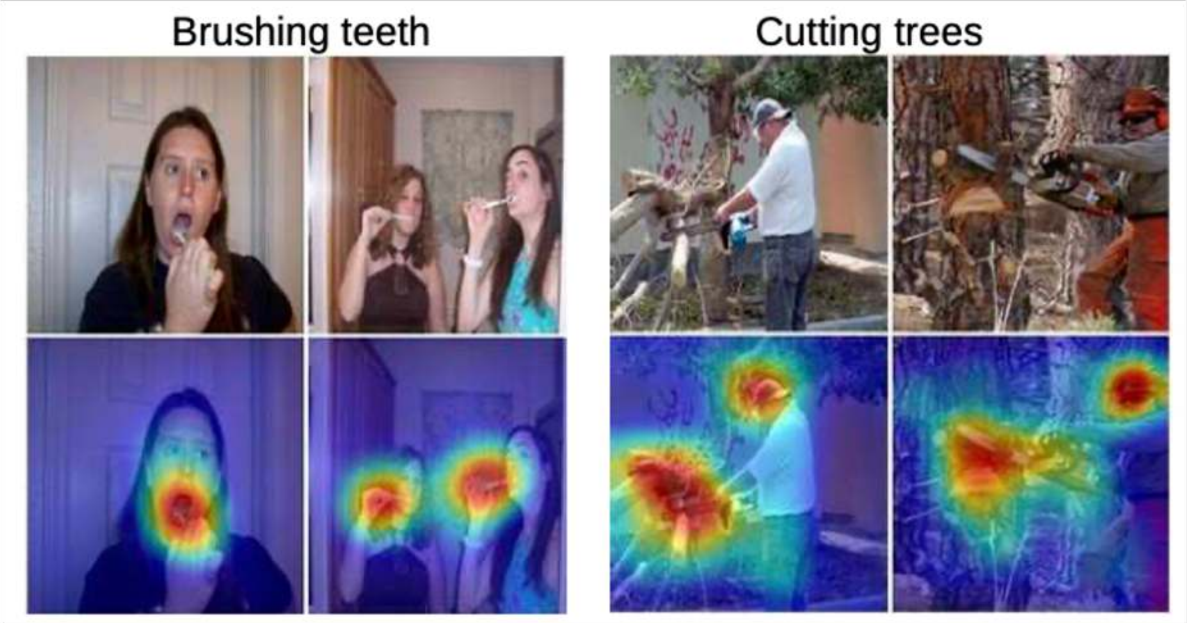
For each concept / class show top input patches



Interpretable hidden units

D. Bau et al., "Network Dissection: Quantifying Interpretability of Deep Visual Representations", CVPR 2017

What Does the Network Look at for Decisions?

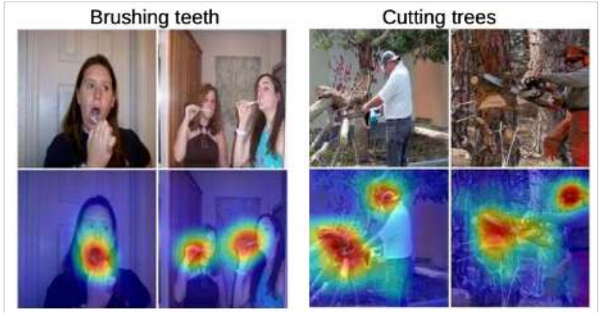


B. Zhou et al., "Learning Deep Features for Discriminative Localization", CVPR 2016 ↩

What Does the Network Look at for Decisions?

Visualize Class Activation Map (CAM)[1]**:**

1. Make network fully convolutional
   - `Global Avg Pool -> Linear` classifier
   - Convert to `1x1 Conv -> Global Avg Pool`
   - For other structures... give up
2. Visualize heatmap of activations



[1] B. Zhou et al., "Learning Deep Features for Discriminative Localization", CVPR 2016 ↩️

What Does the Network Look at for Decisions?

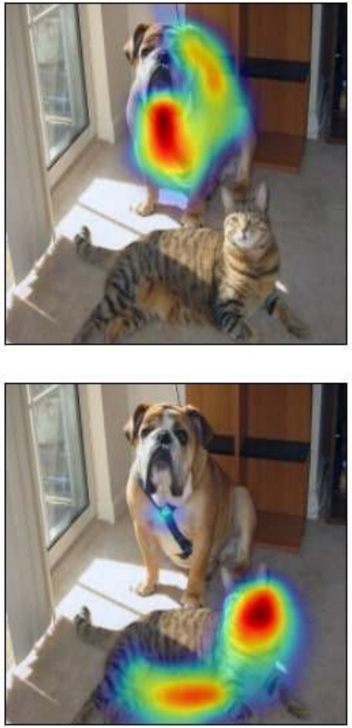
Grad-CAM[1]:

- Generalization of CAM to arbitrary networks

1. Compute class-specific gradient

- Measure influence of each input onto activation
- Average over all spatial locations

2. Visualize heatmap of gradients



[1] R. Selvaraju et al., "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization", CVPR 2017 ↩

Deep Representations - TL;DR

CAM and Grad-CAM help you understand what a network looks at

Network dissection organizes internal activations along categories/concepts