Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Abhinav Gupta, Alexei A. Efros

Presenter: Yiming Pang

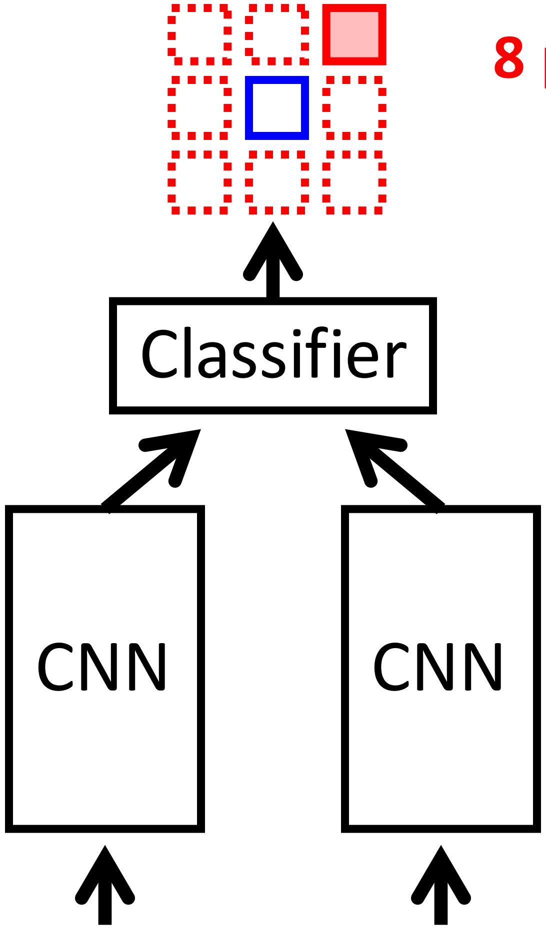
Outline

- Motivation
- Approach
- Experiment
    - Low-level visualization of features
    - Have a deep dream…
    - Apply it to nearest neighbor
- Conclusion

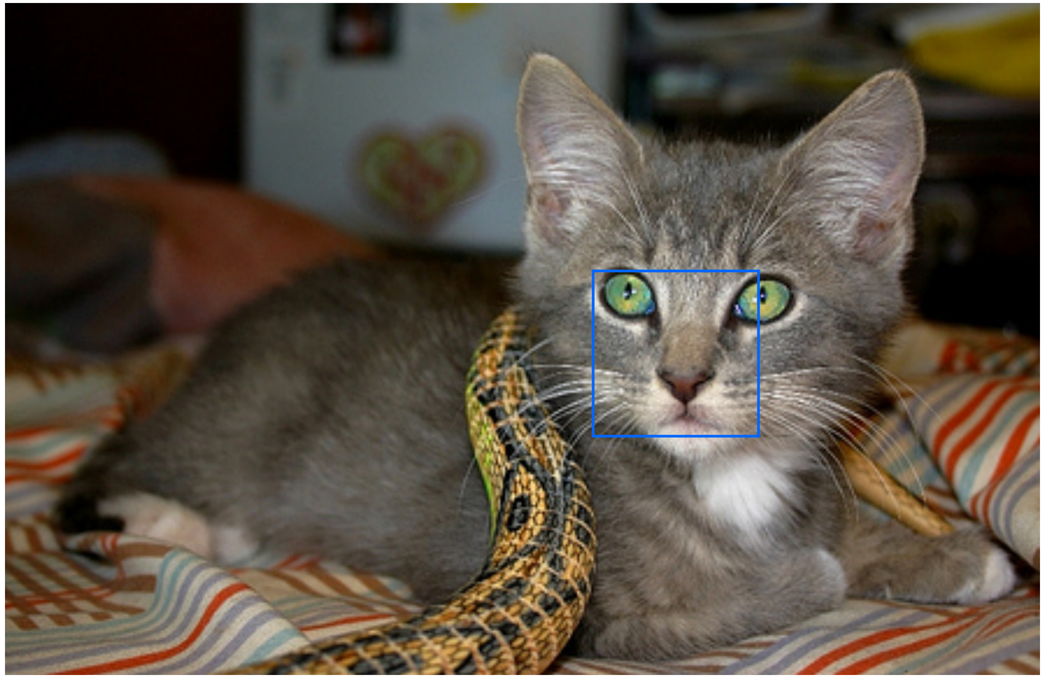
Motivation

- Supervised learning has already shown some promising results…
- with EXPENSIVE labels!

Approach: Make use of Spatial Context



8 possible locations



Randomly Sample Patch

Sample Second Patch

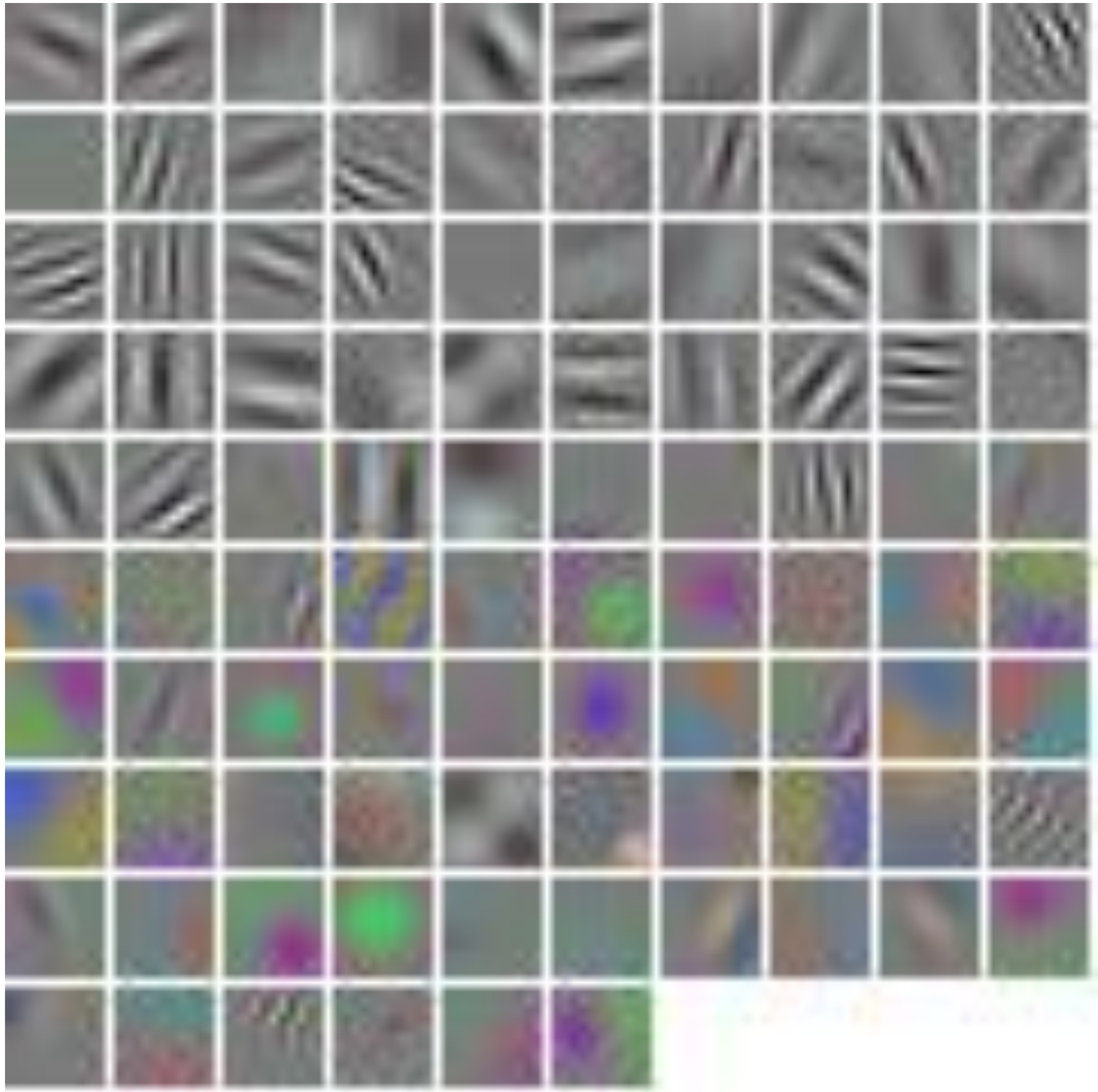
Source: C. Doersch at ICCV 2015

Experiments

- Low-level feature visualization
  - AlexNet
  - Our approach
  - Noroozi and Favaro
  - Wang and Gupta

Compare the filters after Conv1

- **AlexNet trained on ImageNet**
  - Large-scale dataset
  - **With labels**
- Interpret the filters:
  - Nice and smooth
  - No noisy patterns
  - 2 separate streams of processing
    - High-frequency grayscale features
    - Low-frequency color features



ImageNet Classification with Deep Convolutional Neural Networks. A. Krizhevsky, I. Sutskever, and G. Hinton. NIPS 2012

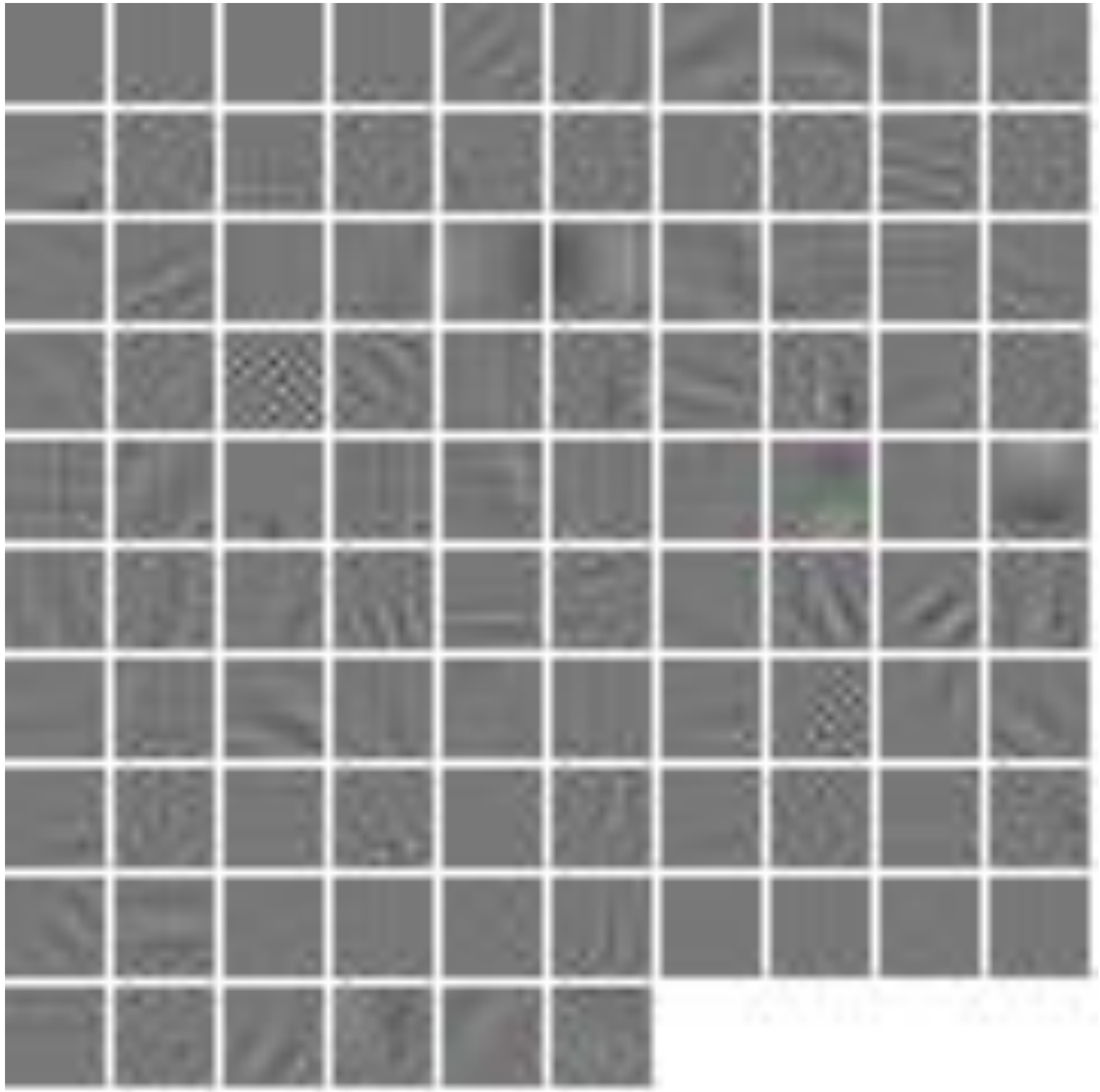
Compare the filters after Conv1

- **Our unsupervised approach**
  - Pre-trained on ImageNet
  - **Without labels**
- Preprocessing with **projection**:
  - Shift green and magenta towards gray
- Interpret the filters
  - Obviously not that good…
  - Noisy patterns exist
  - Due to the projection, some color features are lost

Unsupervised Visual Representation Learning by Context Prediction. C. Doersch, A. Gupta, A. Efros. ICCV 2015.

Compare the filters after Conv1

- **Our unsupervised approach**
  - Pre-trained on ImageNet
  - **Without labels**
- Preprocessing with **color-dropping**:
  - Randomly replace 2 of the 3 color channels with Gaussian noise.
- Interpret the filters
  - Almost no color features
  - More noisy patterns
- ? Somehow it outperforms projection in object detection



Unsupervised Visual Representation Learning by Context Prediction. C. Doersch, A. Gupta, A. Efros. ICCV 2015.

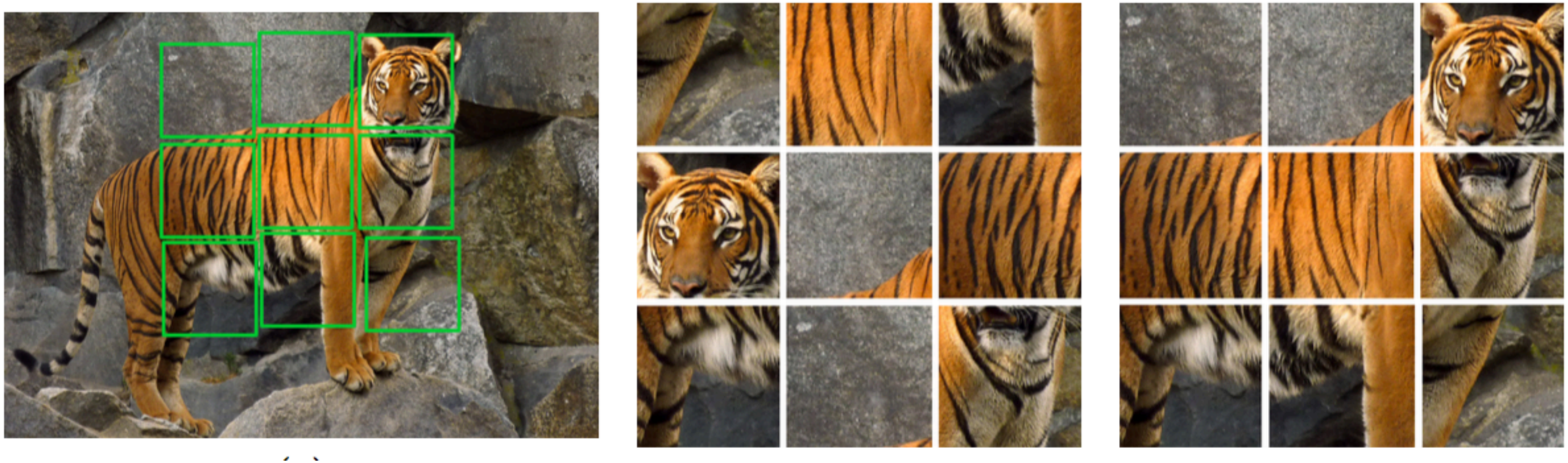
Compare the filters after Conv1

- **Our unsupervised approach**
  - Pre-trained on ImageNet
  - **Without labels**
- **VGG-style network**: high-capacity model (16-layer)
- Interpret the filters
  - Kernel size is 3 (very small)
  - Coarse grained result

Unsupervised Visual Representation Learning by Context Prediction. C. Doersch, A. Gupta, A. Efros. ICCV 2015.

Compare with other models

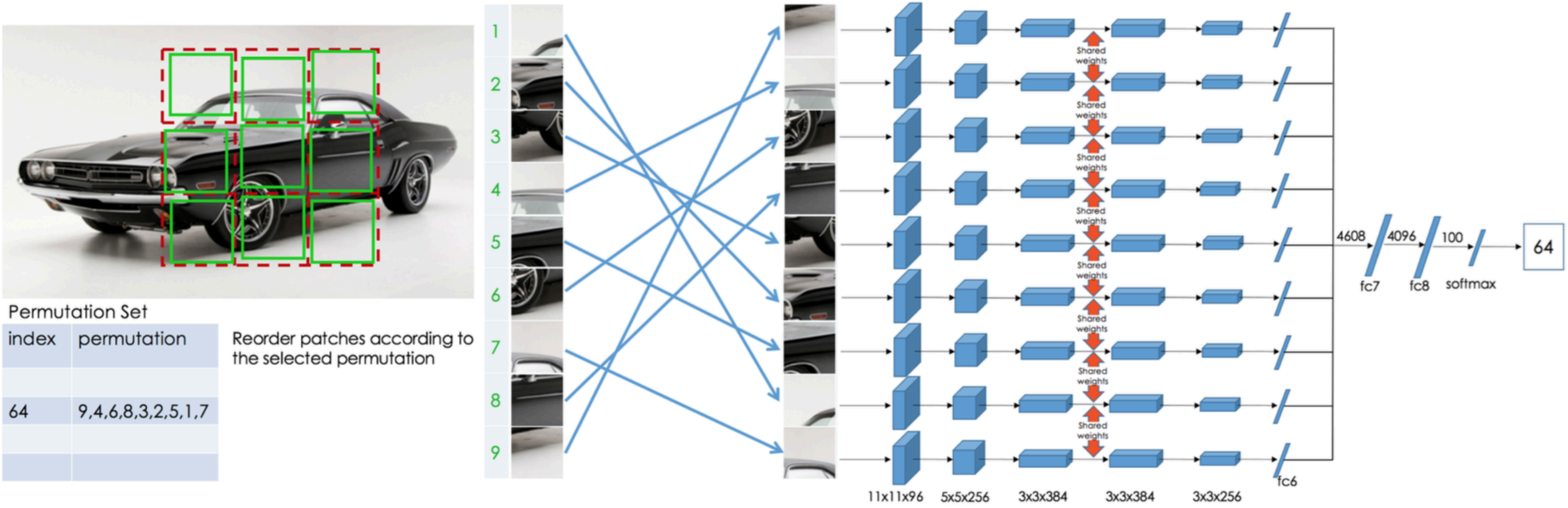
- Instead of just playing with 2 adjacent patches…



Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles M. Noroozi and P. Favaro

Solving Jigsaw Puzzels

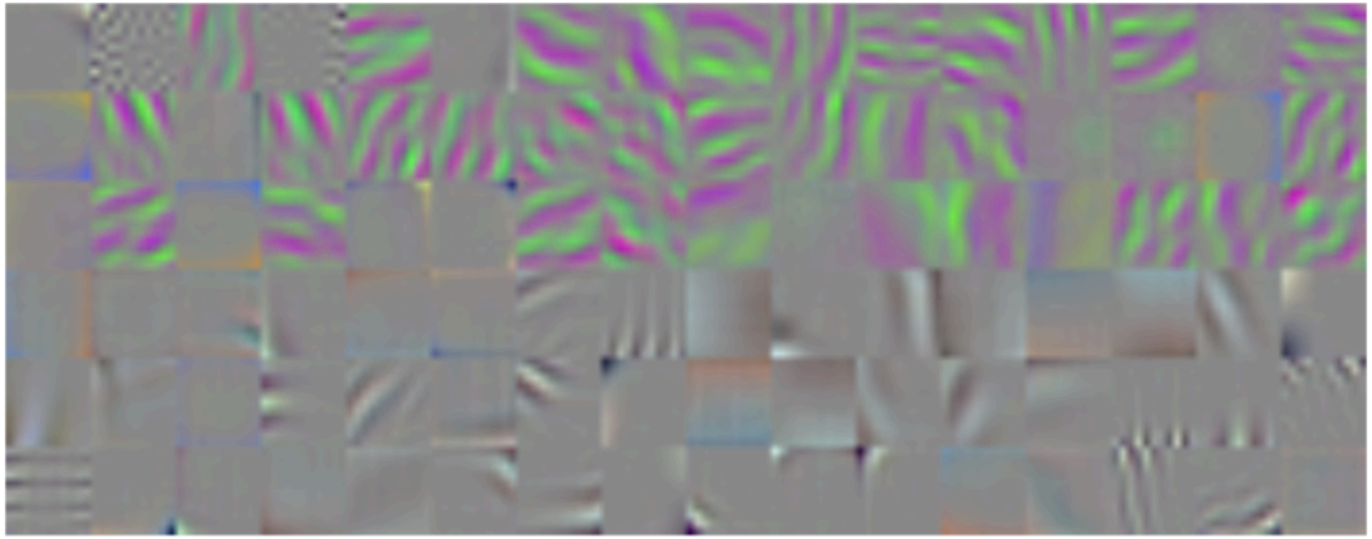
- 2 stacks -> 9 stacks



Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles M. Noroozi and P. Favaro

Filters after Conv1 by the "Jigsaw" approach

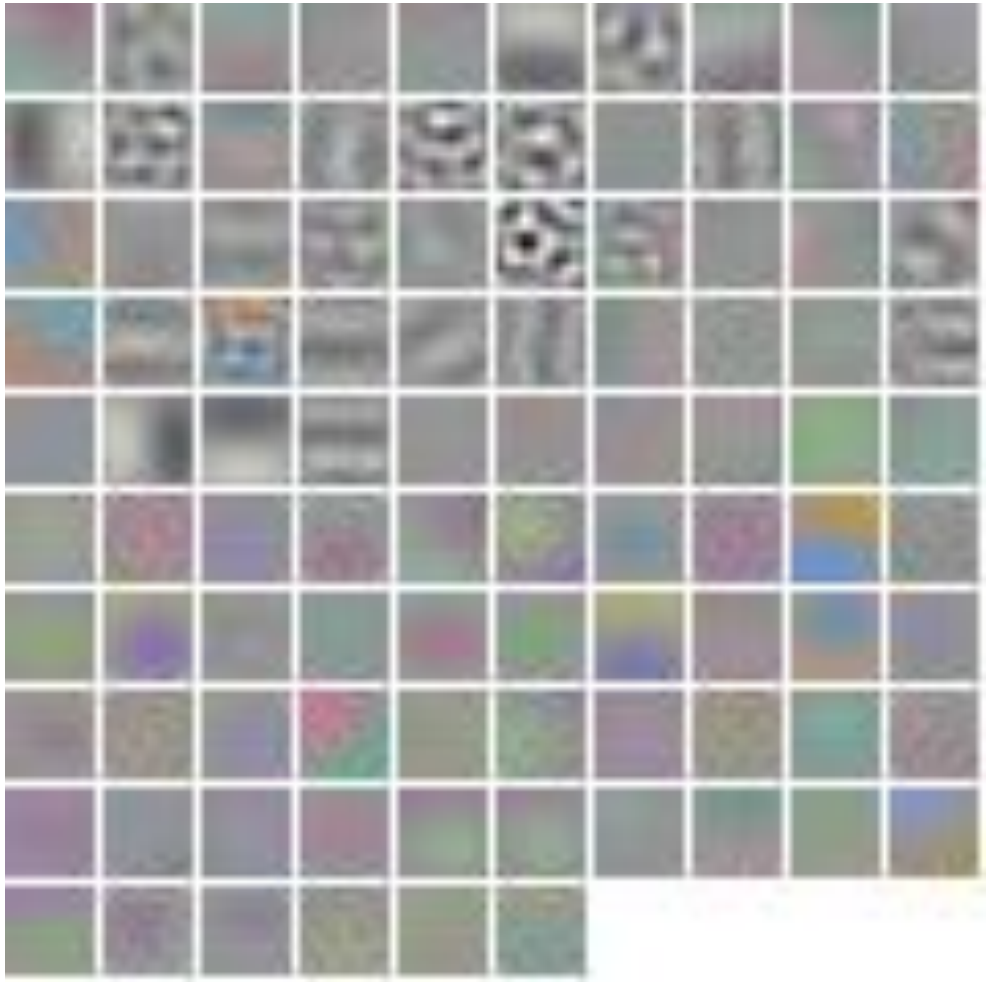
- Unsupervised learning
- Trained on ImageNet
- Compared with Doersch's approach, filters are more smooth with less noisy patterns



Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles M. Noroozi and P. Favaro

Results from other unsupervised methods

- No ImageNet, just 100K unlabeled videos and the VOC 2012 dataset.
- Leverage the fact visual tracking provides the supervision.
- Trained with RGB images



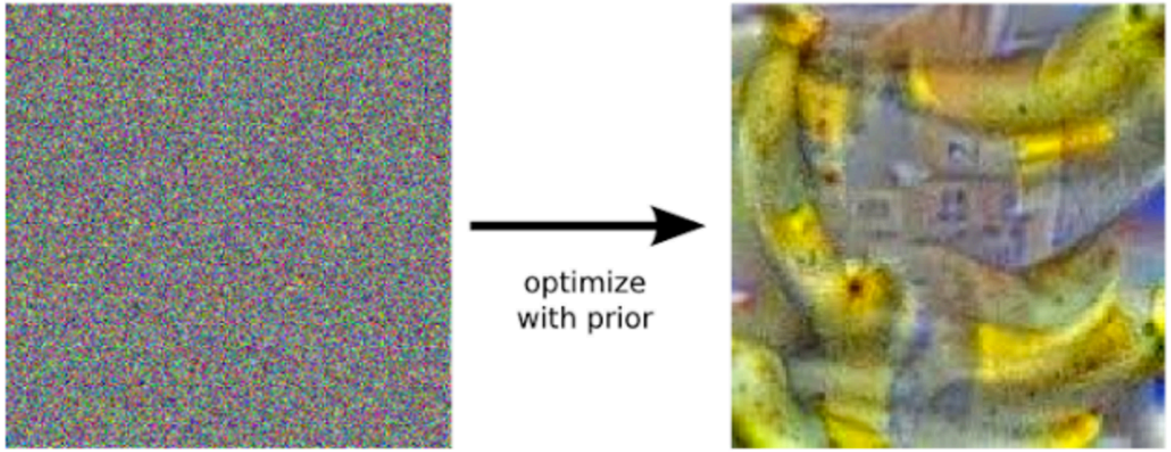
Unsupervised Learning of Visual Representations using Videos X. Wang and A. Gupta (ICCV 2015)

Experiments

- Low-level feature visualization
  - AlexNet
  - Our approach
  - Noroozi and Favaro
  - Wang and Gupta
- Have a deep dream…

Going Deeper into Neural Network

- We understand little of why certain models work and others don't.
- We want to understand what exactly goes on at each layer.
- To visualize this procedure:
  - Turn the network upside down and ask it to enhance an input image in such way as to elicit a particular interpretation.



https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

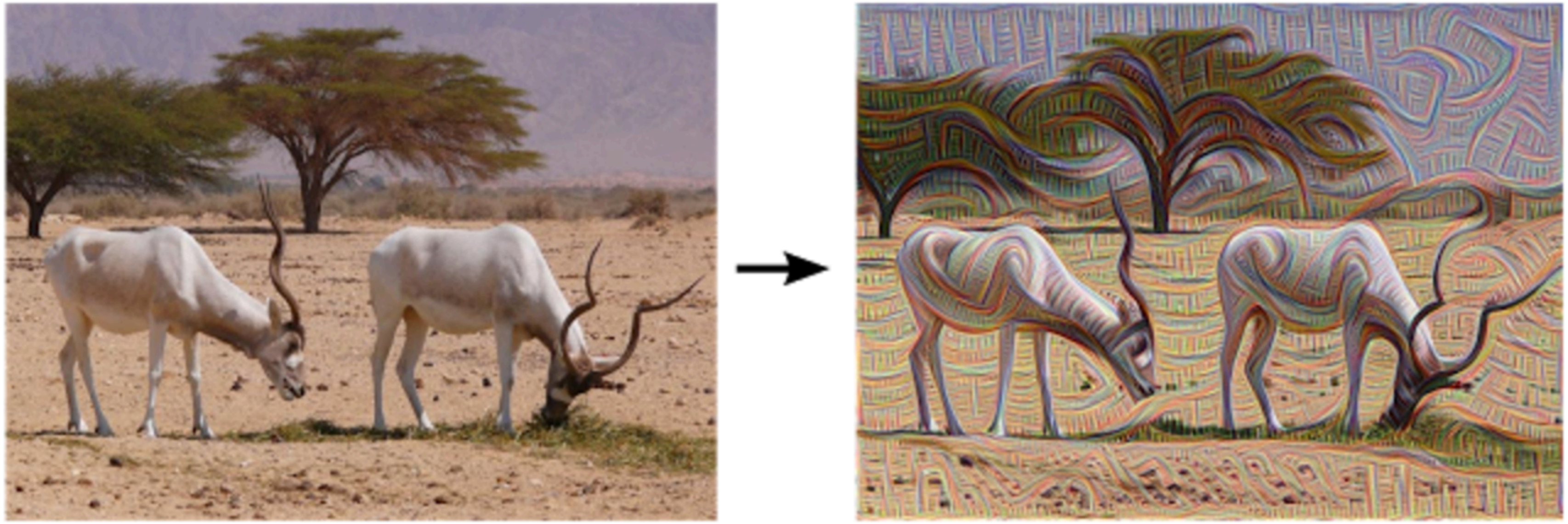
Going Deeper into Neural Network(cont)

- Interesting examples:

https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Going Deeper into Neural Network(cont)

- Enhance the learning result:
  - Feed in an arbitrary image
  - Whatever you see there, just show me more!



https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

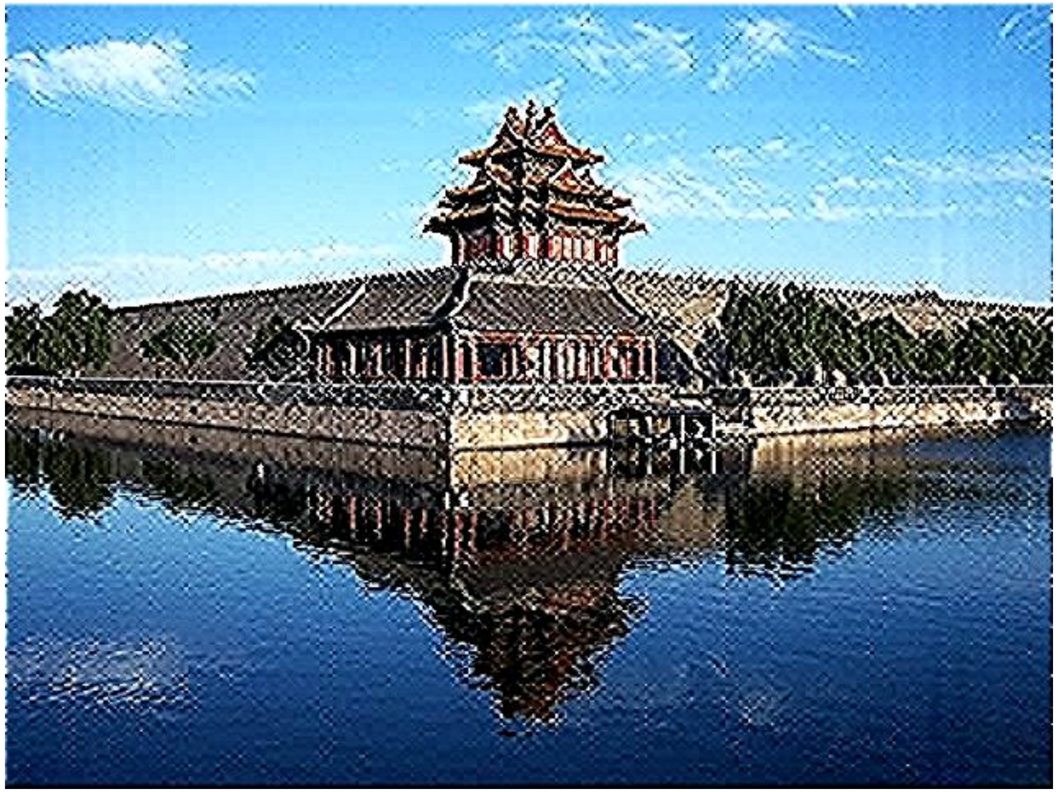
What does the network see:

- Original image:

Supervised AlexNet vs. Unsupervised VGG(ours)

- conv1 vs. conv1_1

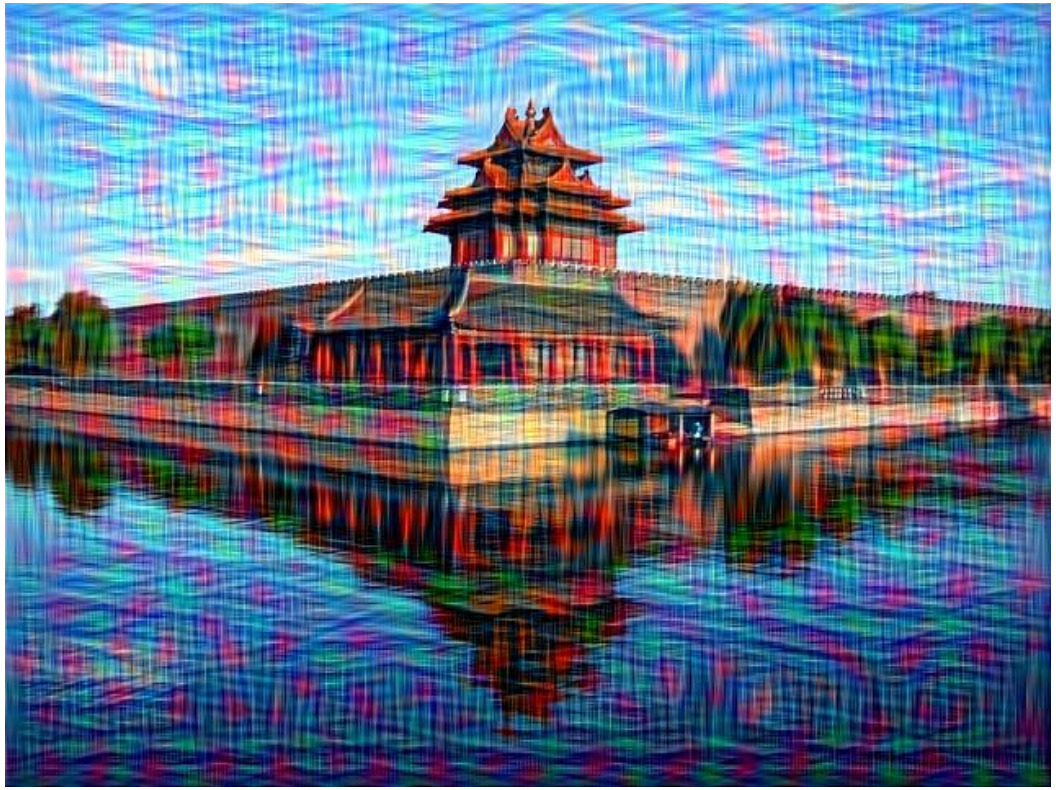
Most on color contrast and the contour

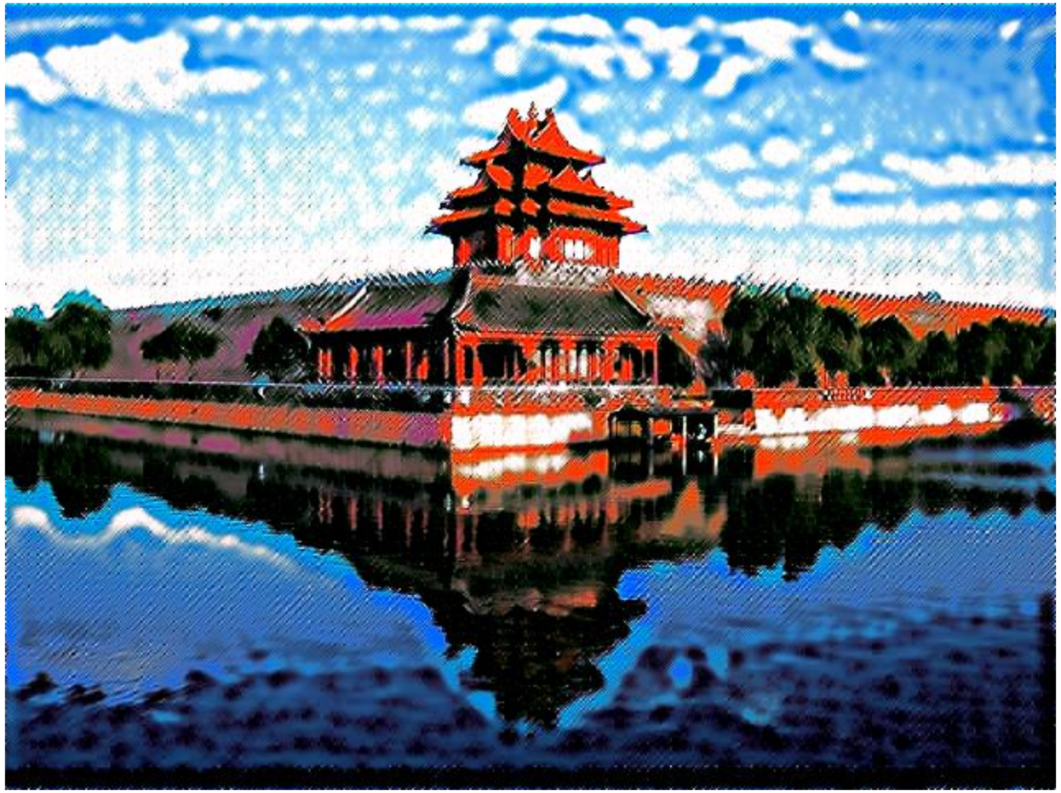


More “fragmented” on edges

Supervised AlexNet vs. Unsupervised VGG(ours)

- conv2 vs. conv2_1



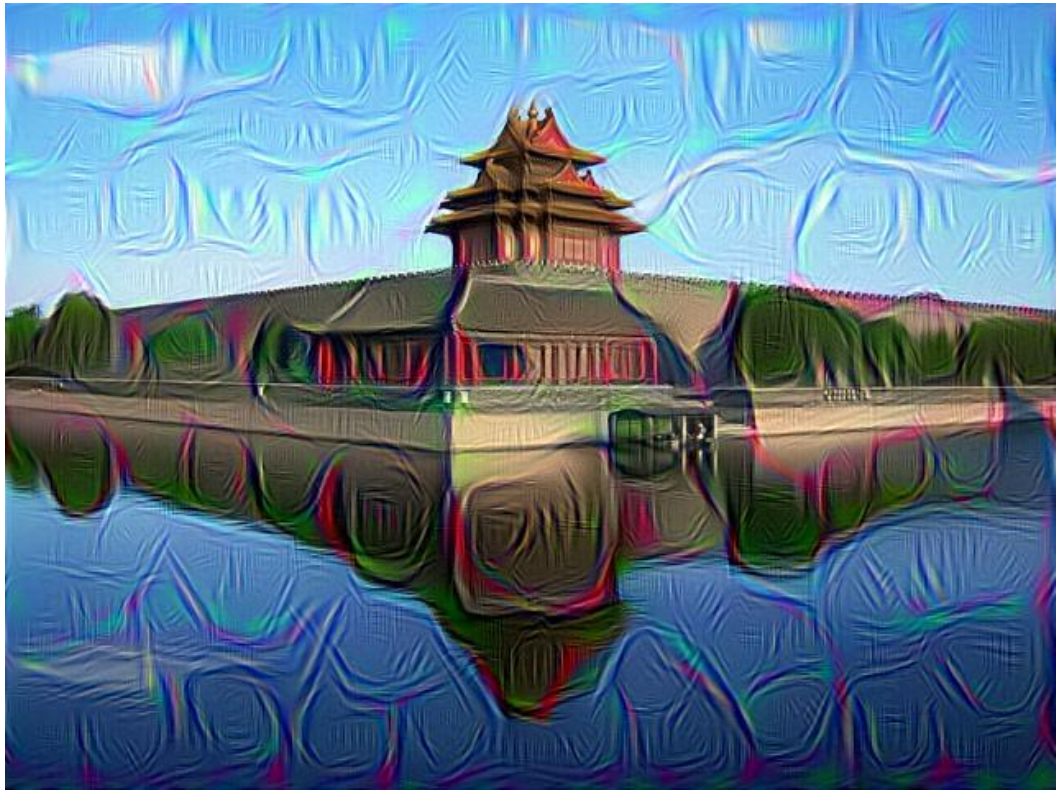


Compared to conv1, this is obviously more “fine-grained”, but still on gradient, as I understand…

Compared to the nice tiny fragments on conv1, this is more “chunked” due to more features focus on the relative position for PATCHES.

Supervised AlexNet vs. Unsupervised VGG(ours)

- conv3 vs. conv3_1

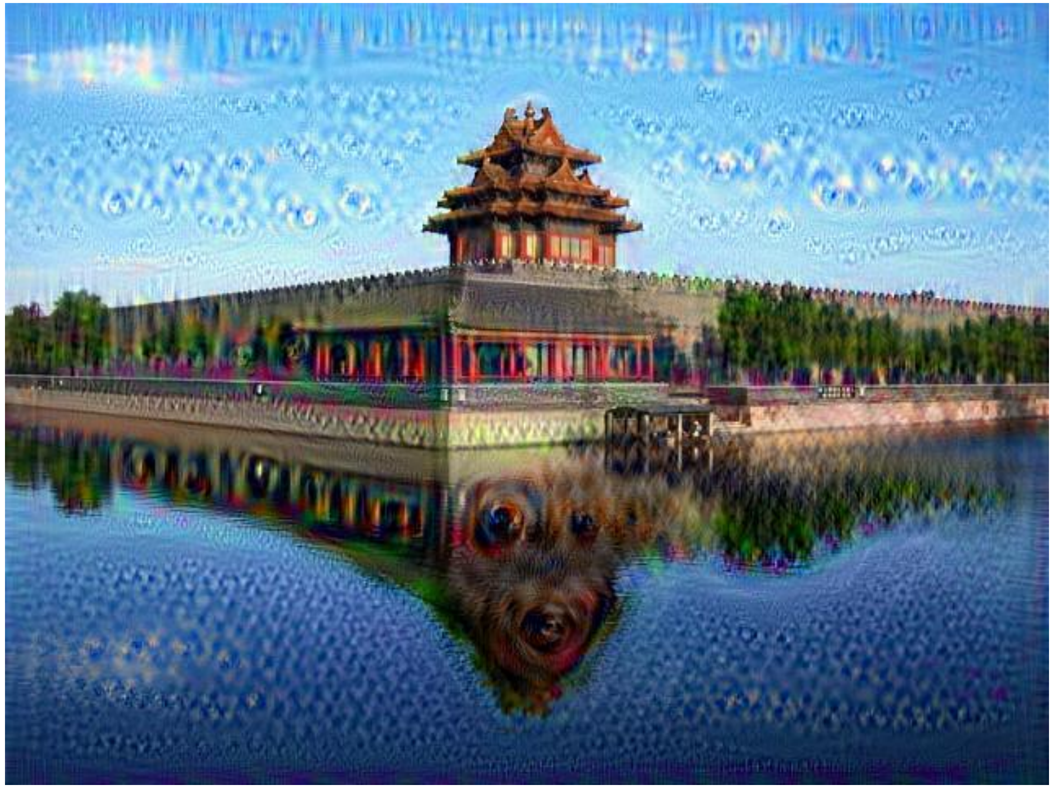


More sophisticated features in image, start to showing some contours indicated by the features.

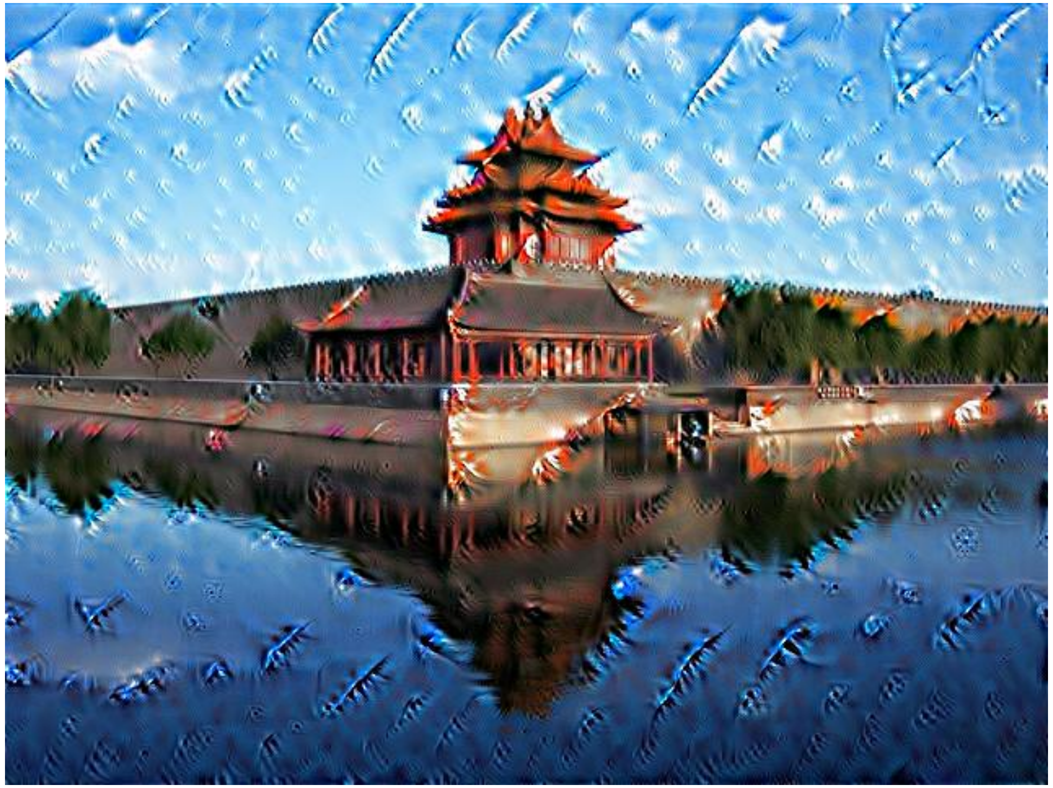
It seems like to be on the opposite direction…
Coarser-grained and the image seems to be divided into tiny patches. We can actually tell some patterns here(like the cloud and sky)

Supervised AlexNet vs. Unsupervised VGG(ours)

- conv4 vs. conv4_1



Some objects start to showing up in the image.

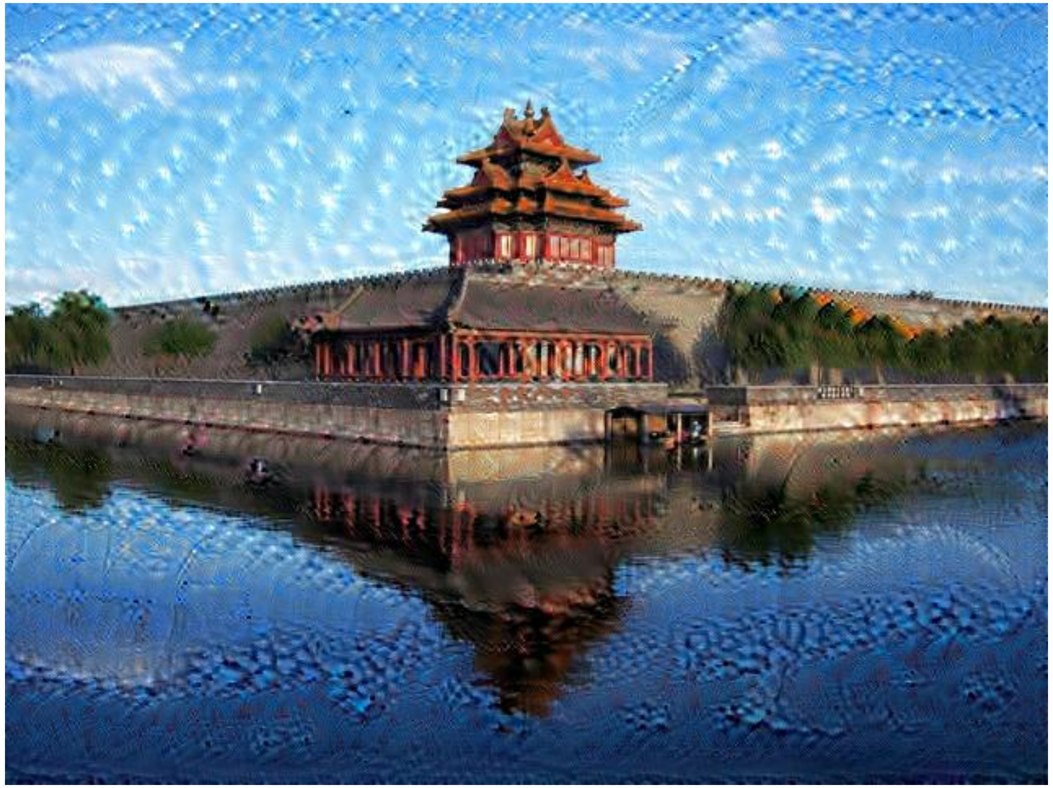


Features start to “converge”

Supervised AlexNet vs. Unsupervised VGG(ours)

- conv5 vs. conv5_1

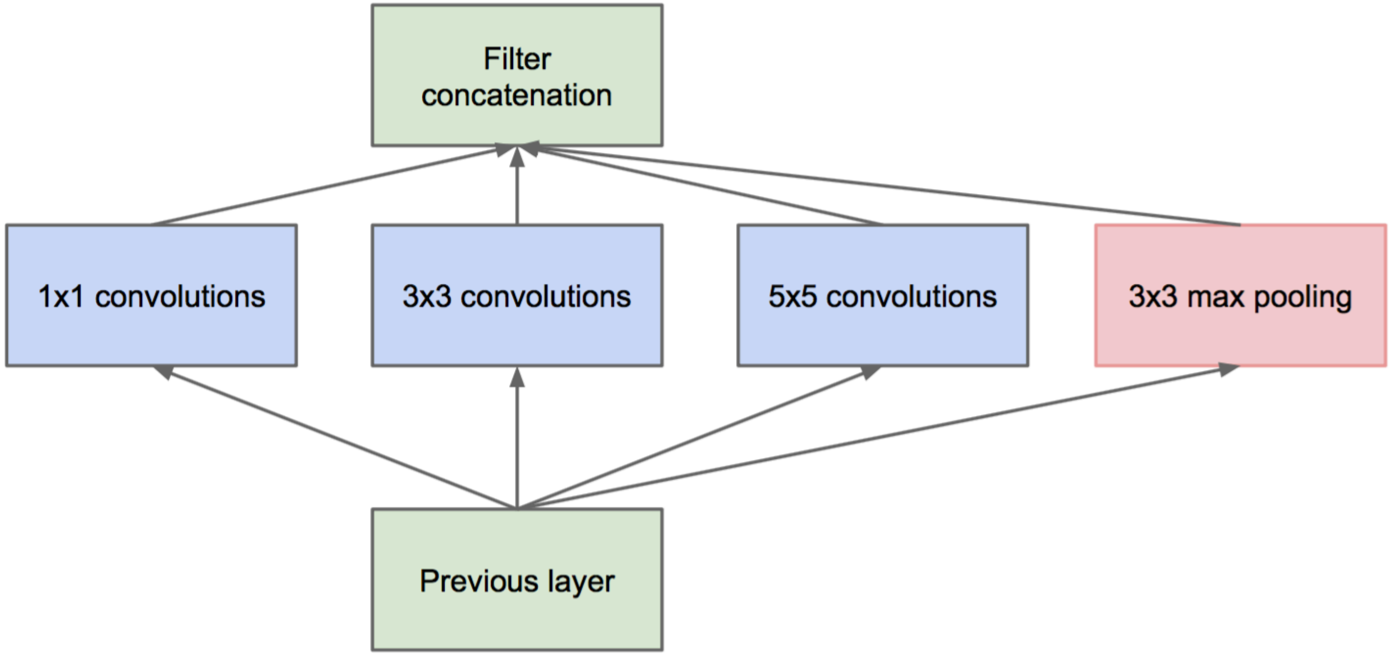
This is how the machine interpret image…



Although starting late, the final results are quite similar to those of the supervised approach.

Deeper Inception

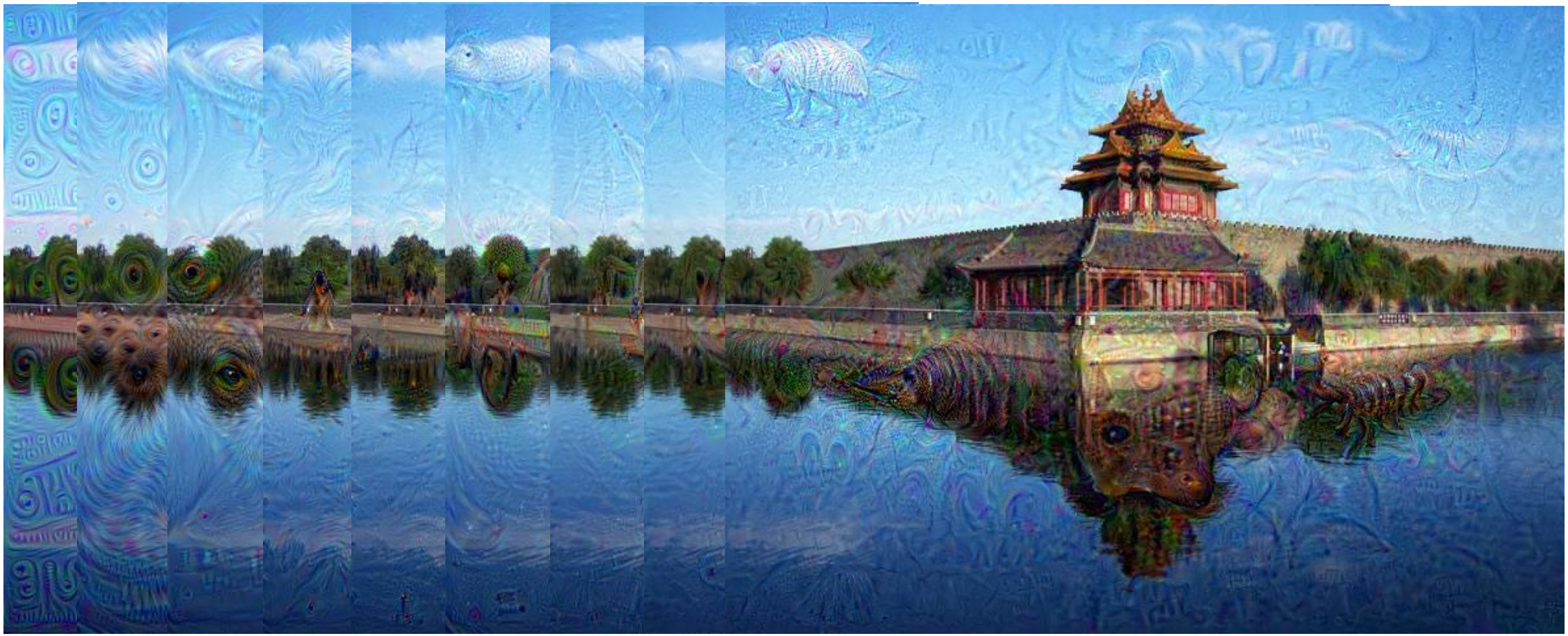
- GoogleNet



Going Deeper with Convolutions C. Szegedy et. al CVPR 2015

GoogleNet Layer by Layer

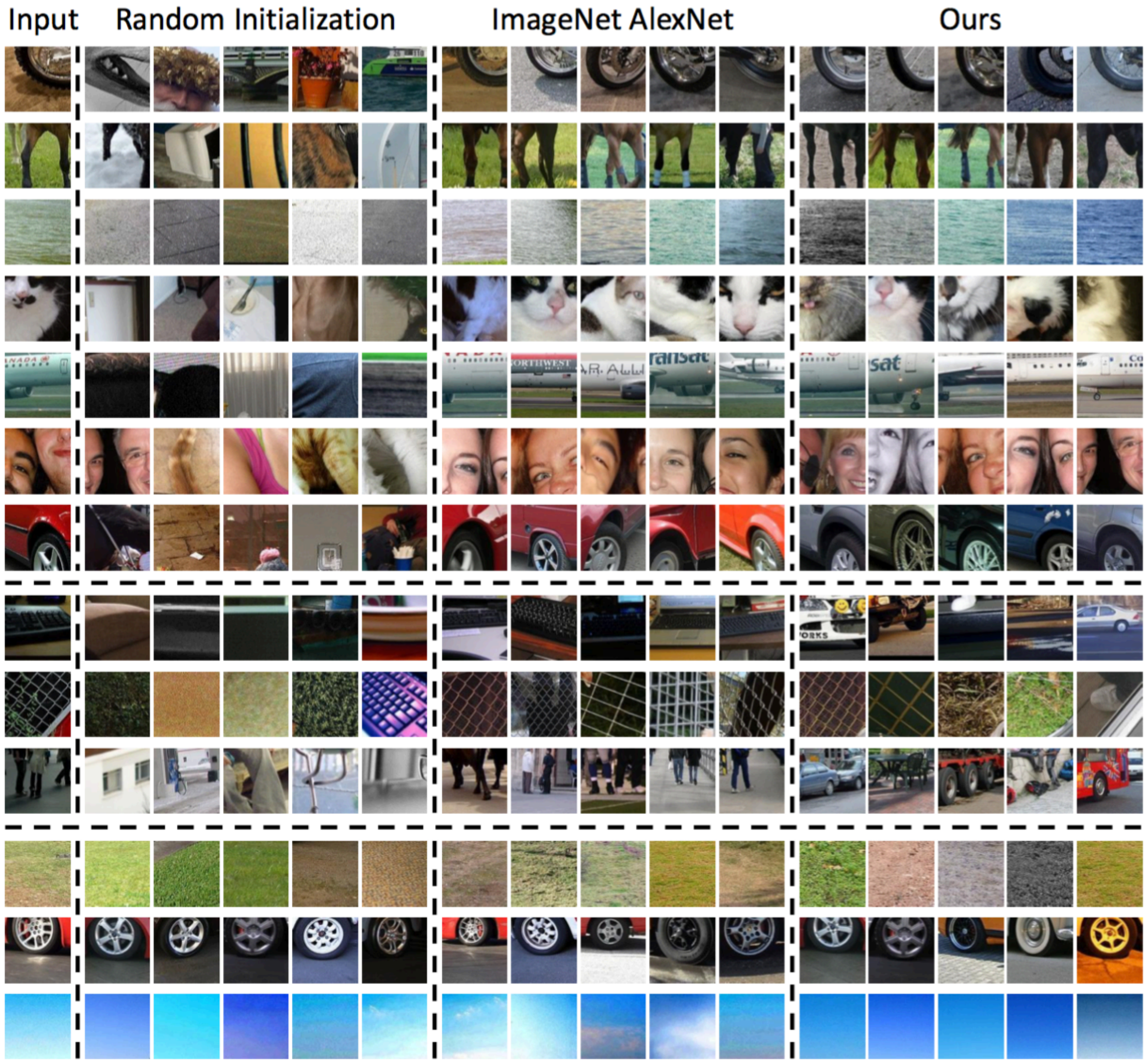
As you go deeper to the network…..



Experiments

- Low-level feature visualization
  - AlexNet
  - Our approach
  - Noroozi and Favaro
  - Wang and Gupta
- Have a deep dream…
- How well can the features do? – nearest neighbor

Results from the paper



The semantic meaning makes this approach different

Having a tire on the bonnet forms a very strange layout, different from normal car image.

AlexNet: More on the image structure, like the round structure of the light and tire

Our approach: It somehow get some “semantic” sense: a tire near the car

The semantic meaning makes this approach different

Some animal's leg near a ladder structure.

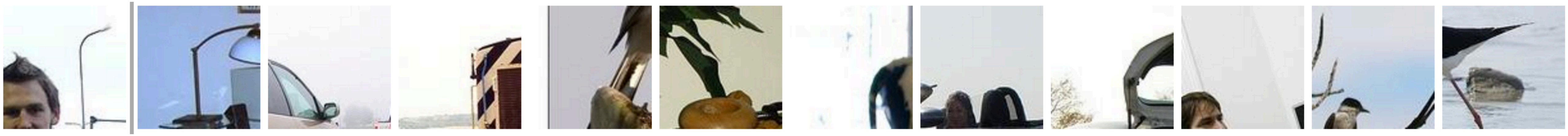
AlexNet: All the results do not make any sense due to there is no salient feature for the query patch.

Our approach: The first result is very similar to the query patch. A "leg"(maybe just some random white bar) and a "ladder"(although it's just weeds forms a ladder shape)

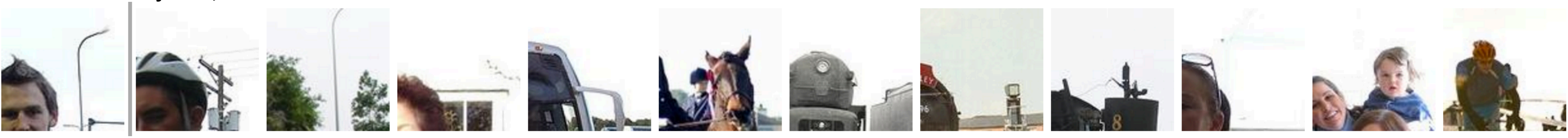
The semantic meaning makes this approach different

A man near a street lights.

AlexNet: The first result shows a very similar street light, all other results are not quite relevant



Our approach: The first result shows exactly the same thing. Other results show a relative position of a human face and other objects, more or less.



Beyond semantics

- Should this be recognized as a car or teeth?

Beyond semantics

- Supervised AlexNet vs. Unsupervised VGG

Distance:
Supervised Model: 0.6221
Our Approach: 0.4360

Distance:
Supervised Model: 0.9296
Our Approach: 0.3306

Supervised model thinks it more of a car meanwhile our unsupervised approach thinks it more of teeth.

Supervised model more on geometry, shapes; our approach more on the contents.

Conclusion

- Show me what you have learned
  - Low-level feature visualization
- How to understand what you have learned
  - Amplify the features obtained by the network at specific layer
- How can that help us
  - Show the features’ “high-level” performance.

- Q&A