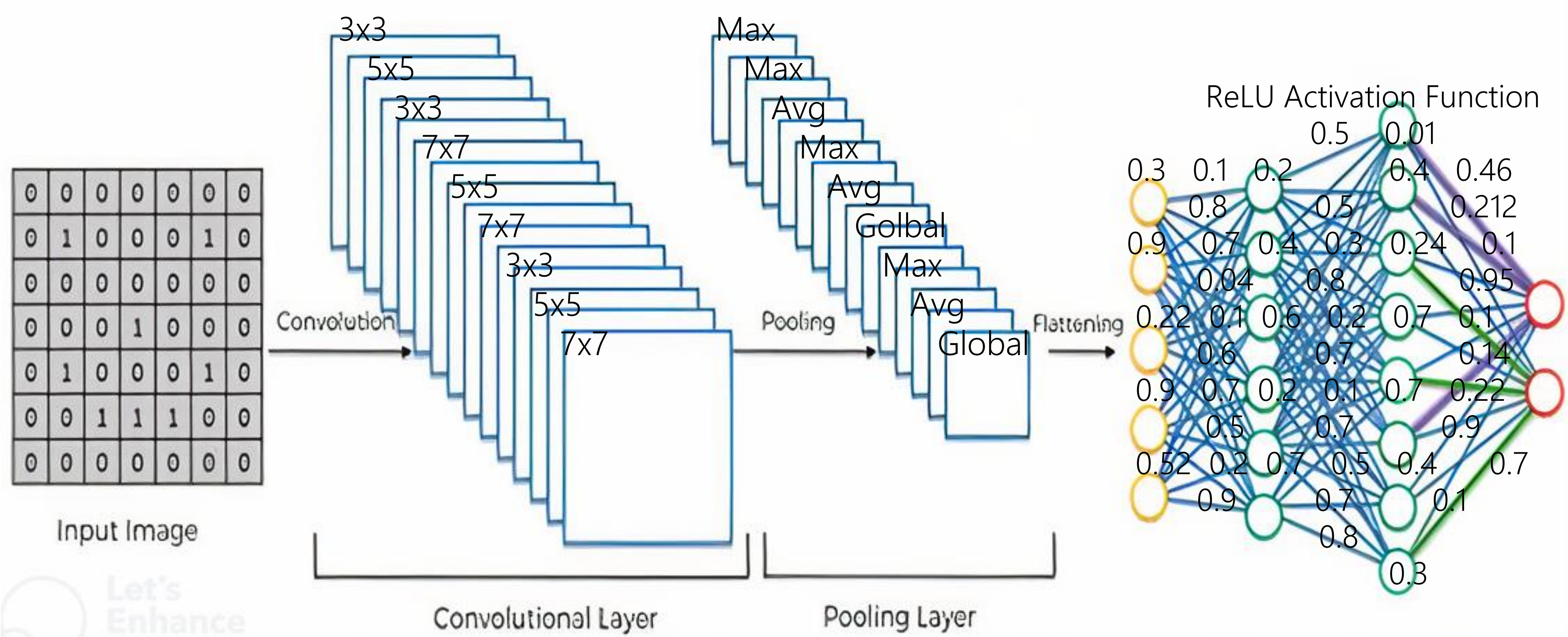
Network Visualizations

How Well do Feature Visualizations Support Causal Understanding of CNN Activations?

Paul Kull

1.1 One Possible Visualization



medium.com on Convolutional Neural Network(CNN) with Practical Implementation

1.2 More Intuitive Visualizations

distill.pub: Feature Visualization –Olah et

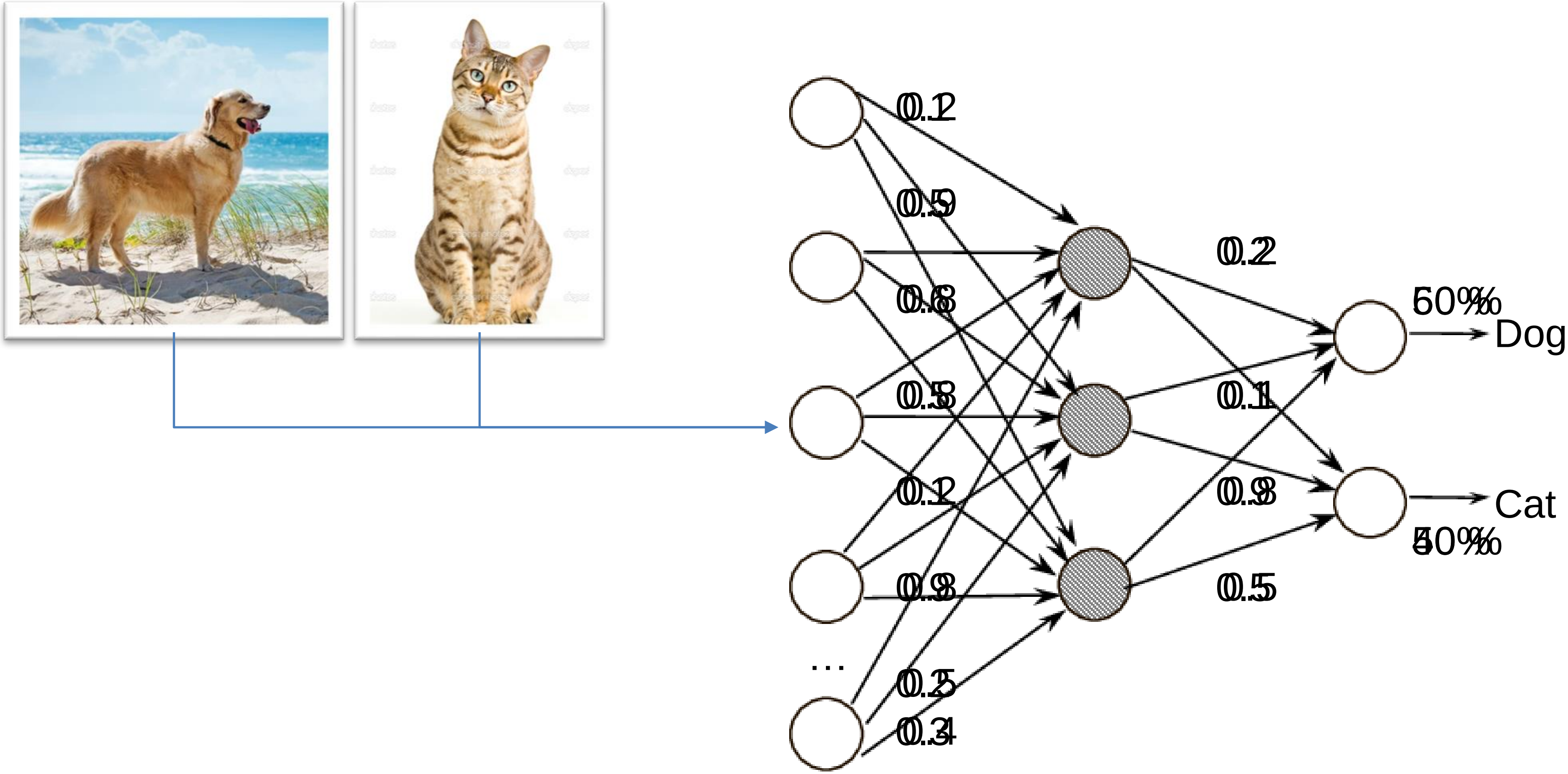
öckermann: Derby Bundesliga Ball
1

distill.pub: Feature Visualization –Olah et al.

Outline

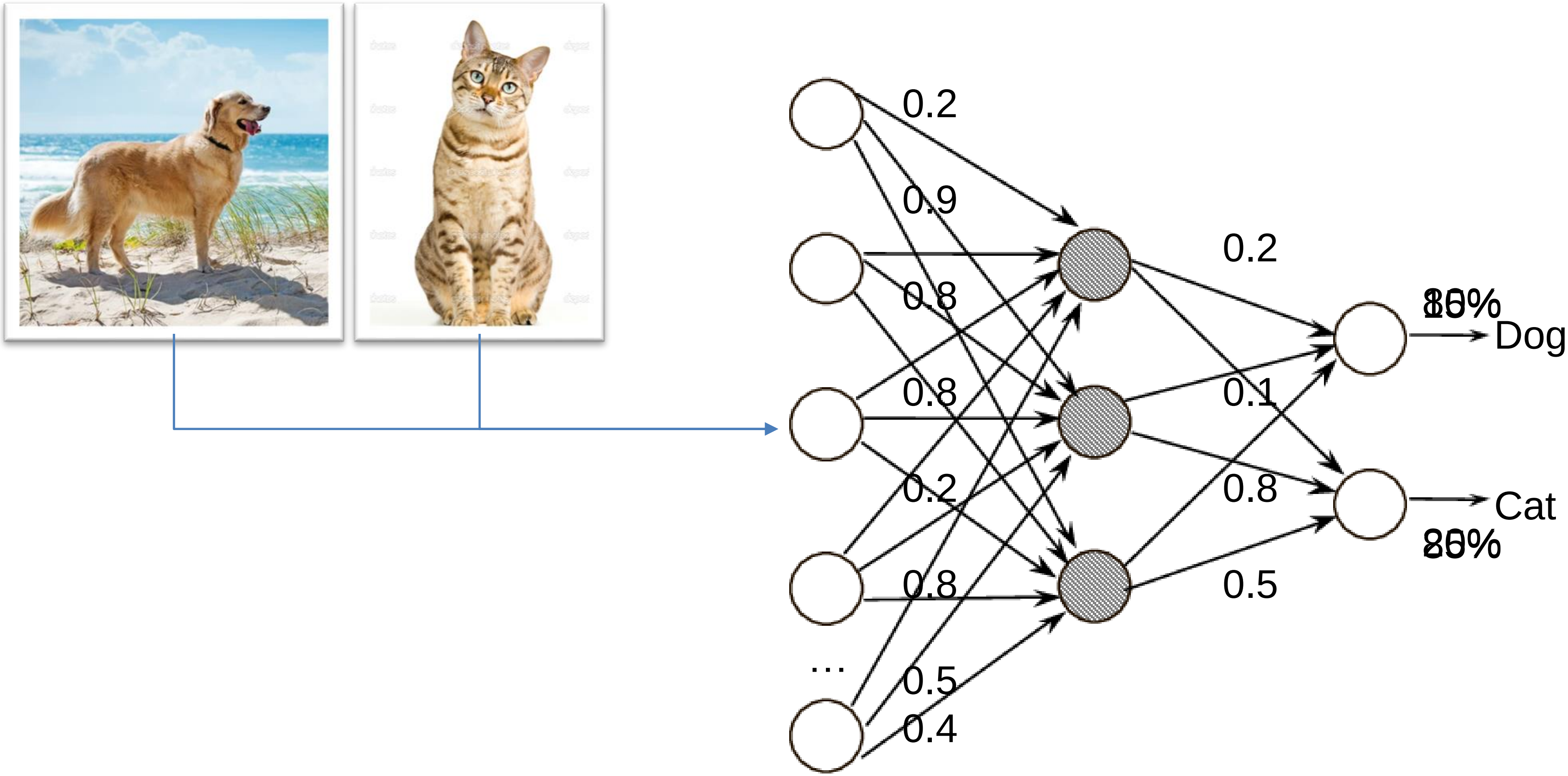
- 1. Introduction
- 2. Image Synthesis by Activation Maximization
- 3. The Experiment
- 4. Results
- 5. Conclusion & Discussion

2.2 Activation Maximization



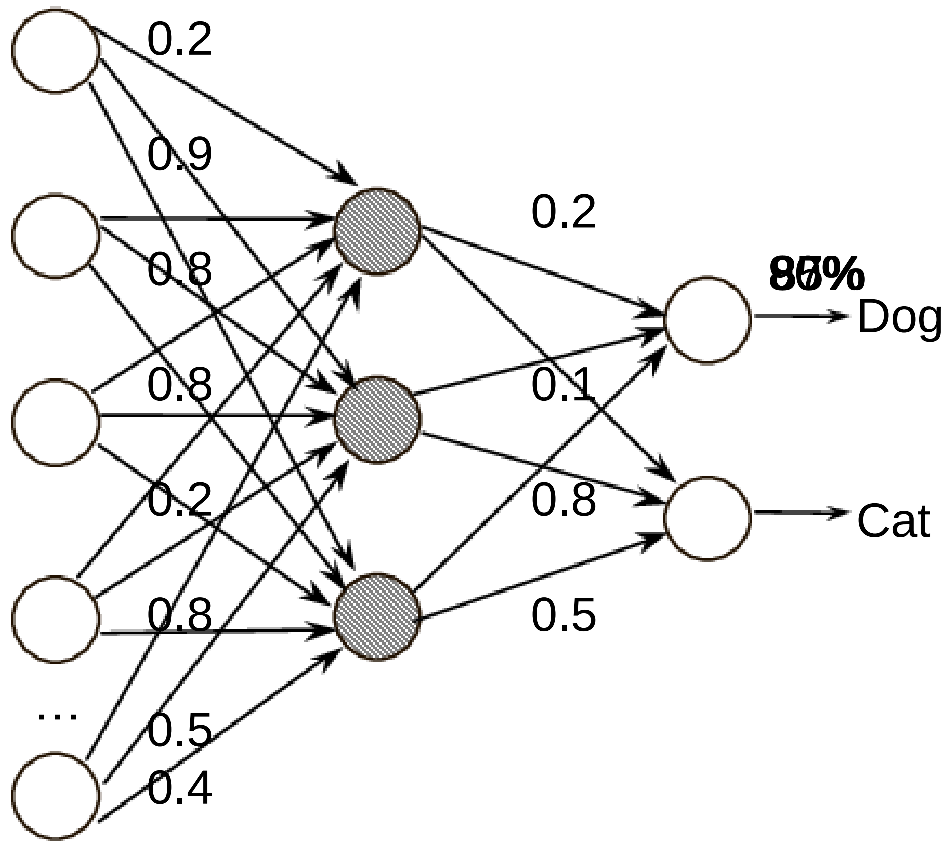
Animal images from depositphotos.com, ANN image from Raymond C Rowe on researchgate.com

2.2 Activation Maximization



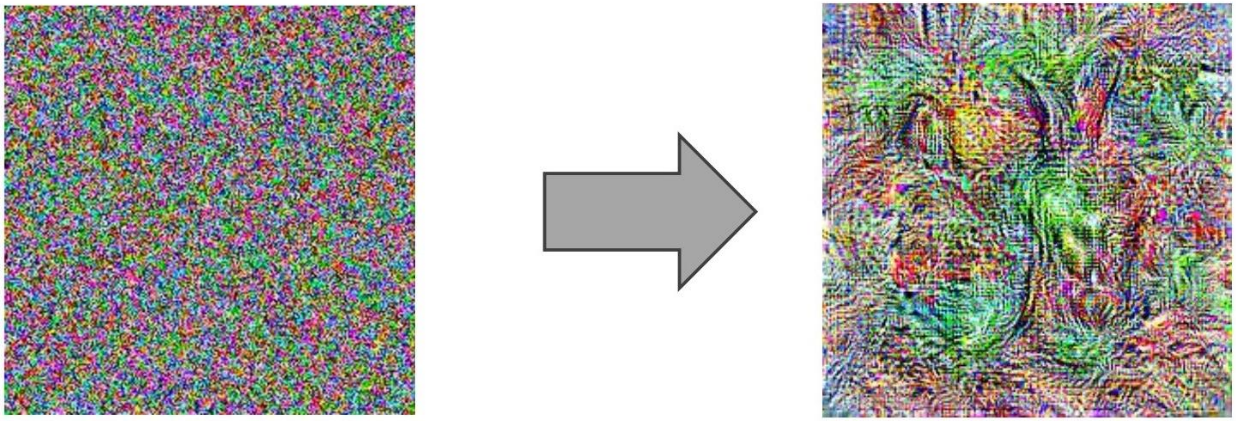
Animal images from depositphotos.com, ANN image from Raymond C Rowe on researchgate.com

2.2 Activation Maximization



Animal images from depositphotos.com, ANN image from Raymond C Rowe on researchgate.com

2.2 Activation Maximization



(a) Random initialization

(b) Synthesized rubbish example

- Synthesizing images via gradient ascent alone is not enough!
- => Use of hand designed prior constraints is necessary

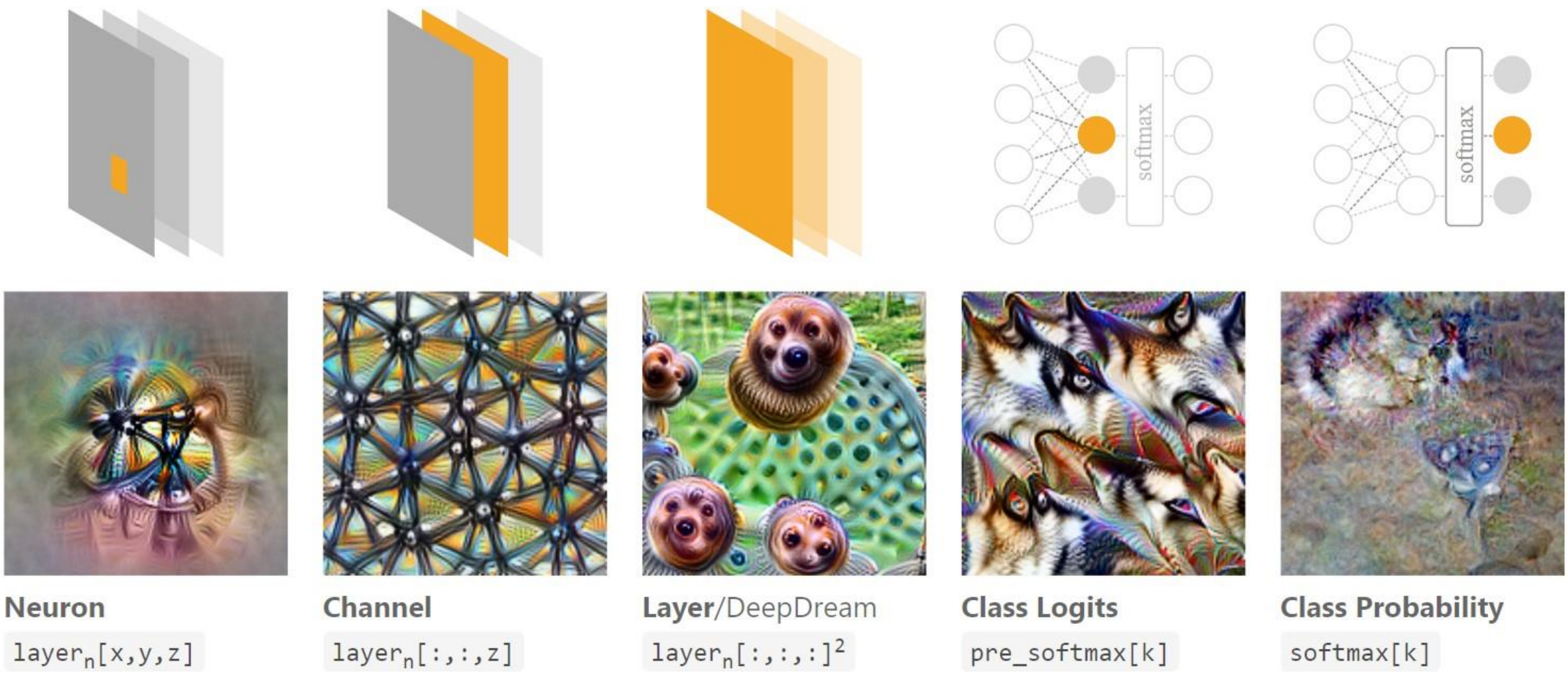
Graphic by Nguyen et al. 2019 Understanding Neural Networks via Feature Visualization: A Survey

2.2 Activation Maximization

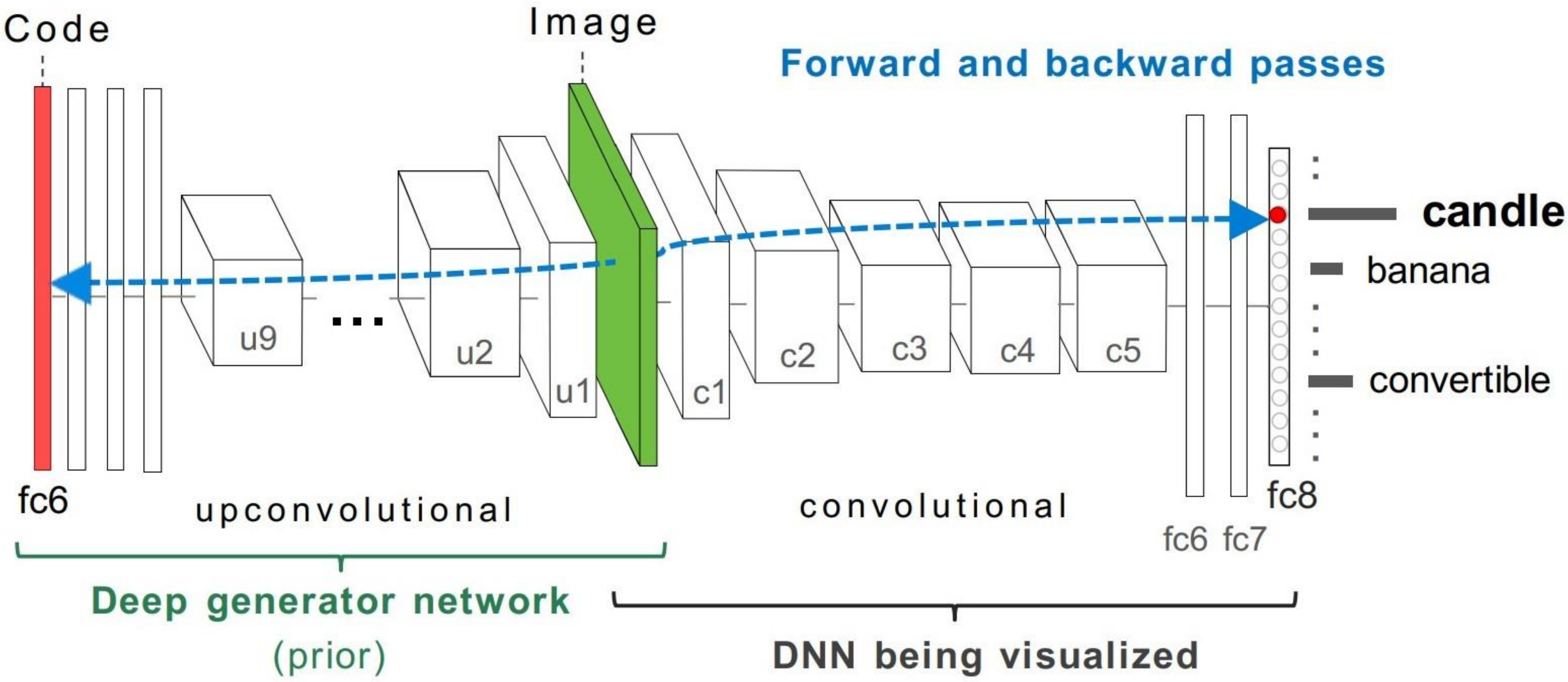
- [Nguyen et al. „Understanding Neural Networks via Feature Visualization: A survey“ (2019)] Priors :
  - Regularization term: $\mathbf{x}^* = \arg\max_{\mathbf{x}}(a(\mathbf{x}) - R(\mathbf{x}))$
  - Penalize high-intensity pixels
  - Penalize high-frequency noise (i.e. smoothing)
  - Penalize the high frequencies in the gradient image
  - Encourage patch-level colour statistics to be more realistic
  - Randomly jitter, rotate or scale the image before each update step
  - These regularizations help improve **local** statistics

- Priors:
  - **Global** coherence is even harder to achieve

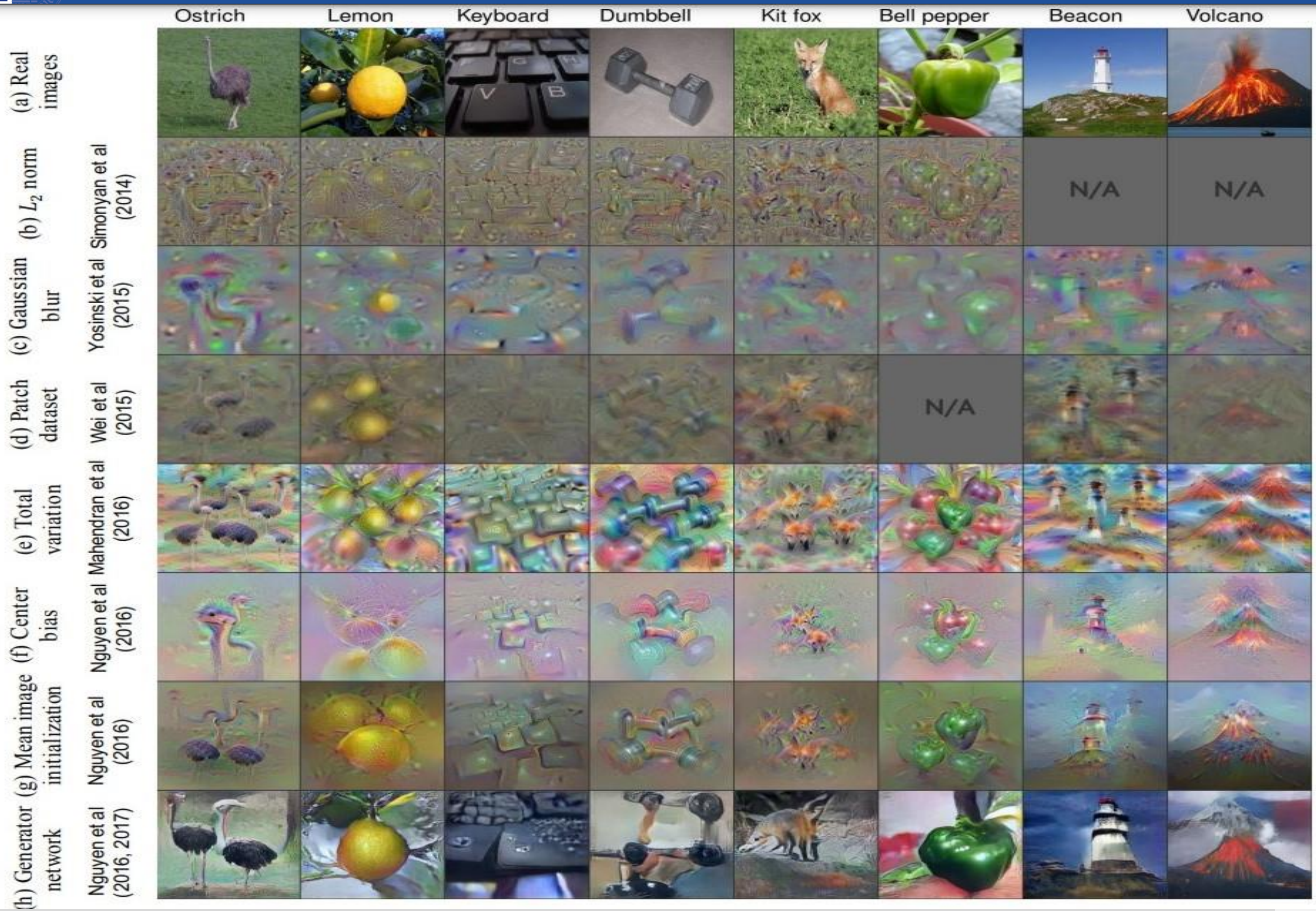
- Diversity:



Deep Generator Networks [Nguyen et Brox et al. „Synthesizing the preferred inputs for neurons in neural networks via deep generator networks“ (2016)]



2.2 Activation Maximization

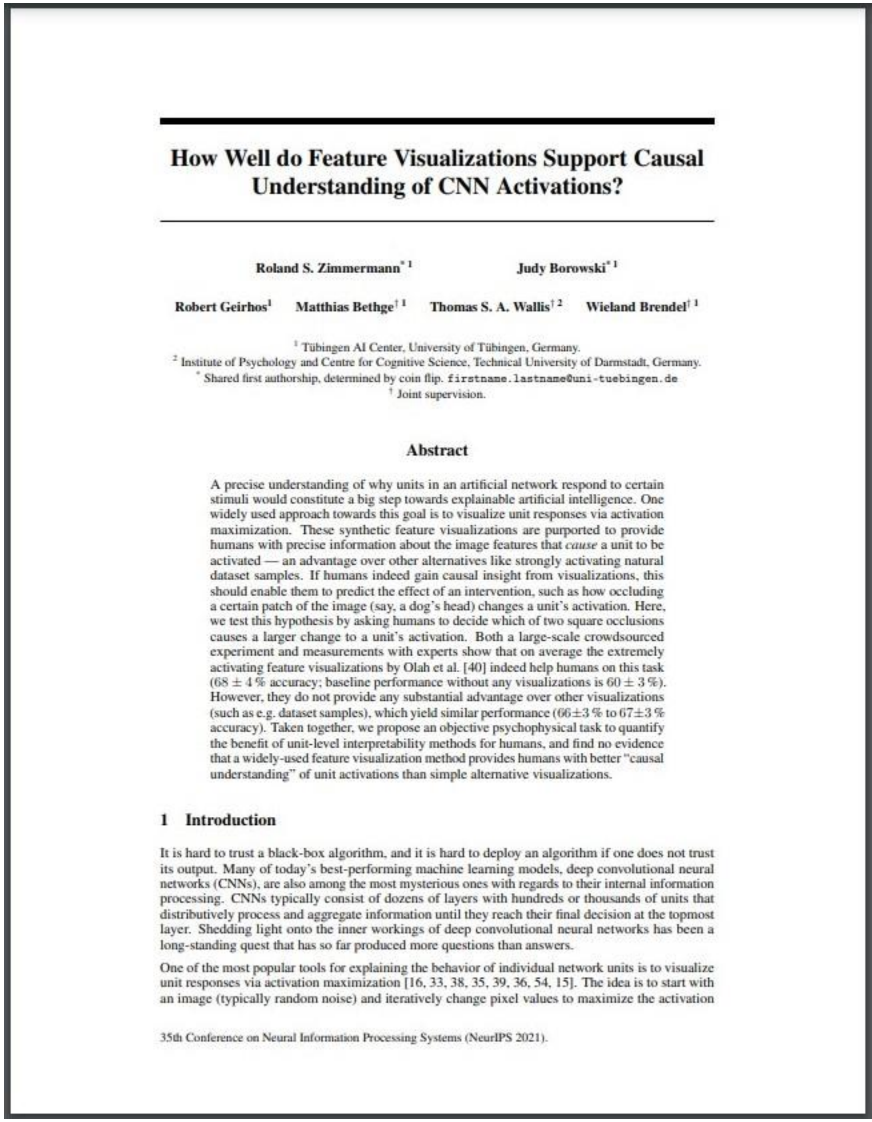


 Graphic by Nguyen et al. 2019 Understanding Neural Networks via Feature Visualization: A Survey

3. The Experiment

- How Well do Feature Visualizations Support Causal Understanding of CNN Activations?

- 2021 Paper by Roland S. Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas S. A. Wallis and Wieland Brendel



How Well do Feature Visualizations Support Causal Understanding of CNN Activations?

Roland S. Zimmermann[*1] Judy Borowski[*1]

Robert Geirhos[1] Matthias Bethge[†1] Thomas S. A. Wallis[†2] Wieland Brendel[†1]

[1] Tübingen AI Center, University of Tübingen, Germany.
[2] Institute of Psychology and Centre for Cognitive Science, Technical University of Darmstadt, Germany.
[*] Shared first authorship, determined by coin flip. `firstname.lastname@uni-tuebingen.de`
[†] Joint supervision.

Abstract

A precise understanding of why units in an artificial network respond to certain stimuli would constitute a big step towards explainable artificial intelligence. One widely used approach towards this goal is to visualize unit responses via activation maximization. These synthetic feature visualizations are purported to provide humans with precise information about the image features that *cause* a unit to be activated — an advantage over other alternatives like strongly activating natural dataset samples. If humans indeed gain causal insight from visualizations, this should enable them to predict the effect of an intervention, such as how occluding a certain patch of the image (say, a dog's head) changes a unit's activation. Here, we test this hypothesis by asking humans to decide which of two square occlusions causes a larger change to a unit's activation. Both a large-scale crowdsourced experiment and measurements with experts show that on average the extremely activating feature visualizations by Olah et al. [40] indeed help humans on this task (68 ± 4 % accuracy; baseline performance without any visualizations is 60 ± 3 %). However, they do not provide any substantial advantage over other visualizations (such as e.g. dataset samples), which yield similar performance (66 ± 3 % to 67 ± 3 % accuracy). Taken together, we propose an objective psychophysical task to quantify the benefit of unit-level interpretability methods for humans, and find no evidence that a widely-used feature visualization method provides humans with better "causal understanding" of unit activations than simple alternative visualizations.

1 Introduction

It is hard to trust a black-box algorithm, and it is hard to deploy an algorithm if one does not trust its output. Many of today's best-performing machine learning models, deep convolutional neural networks (CNNs), are also among the most mysterious ones with regards to their internal information processing. CNNs typically consist of dozens of layers with hundreds or thousands of units that distributively process and aggregate information until they reach their final decision at the topmost layer. Shedding light onto the inner workings of deep convolutional neural networks has been a long-standing quest that has so far produced more questions than answers.

One of the most popular tools for explaining the behavior of individual network units is to visualize unit responses via activation maximization [16, 33, 38, 35, 39, 36, 54, 15]. The idea is to start with an image (typically random noise) and iteratively change pixel values to maximize the activation

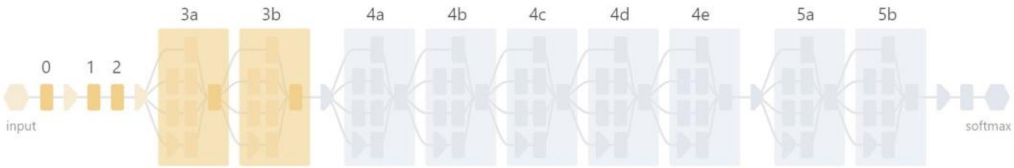
35th Conference on Neural Information Processing Systems (NeurIPS 2021).

3. The Experiment

- 3.1 Technical Facts

3.1 Technical Facts

- Inception V1 Network



Olah et al. “an overview of early vision in InceptionV1”

- Trained on ImageNet

- Query and Natural Images are selected from a random subset of 599.552 images from ImageNet ILSVRC 2012 dataset

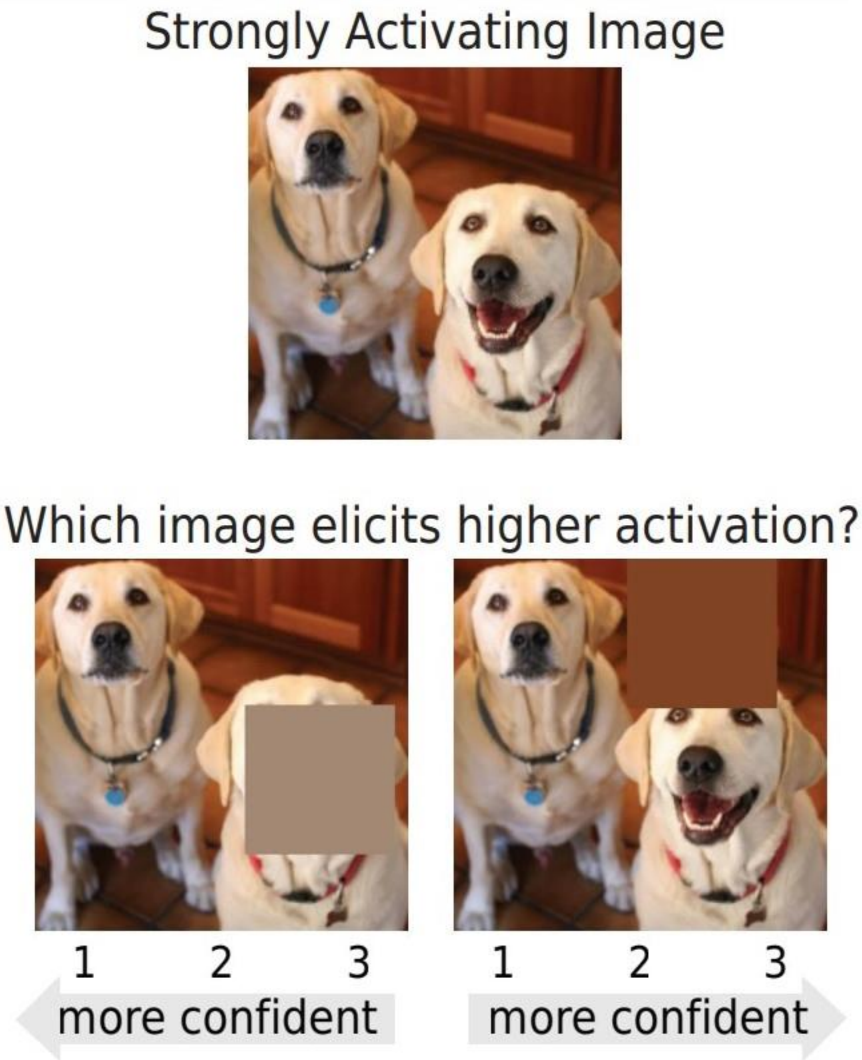
- Units are sampled from 9 layers and 2 Inception module branches (3 × 3 and POOL)

3. The Experiment

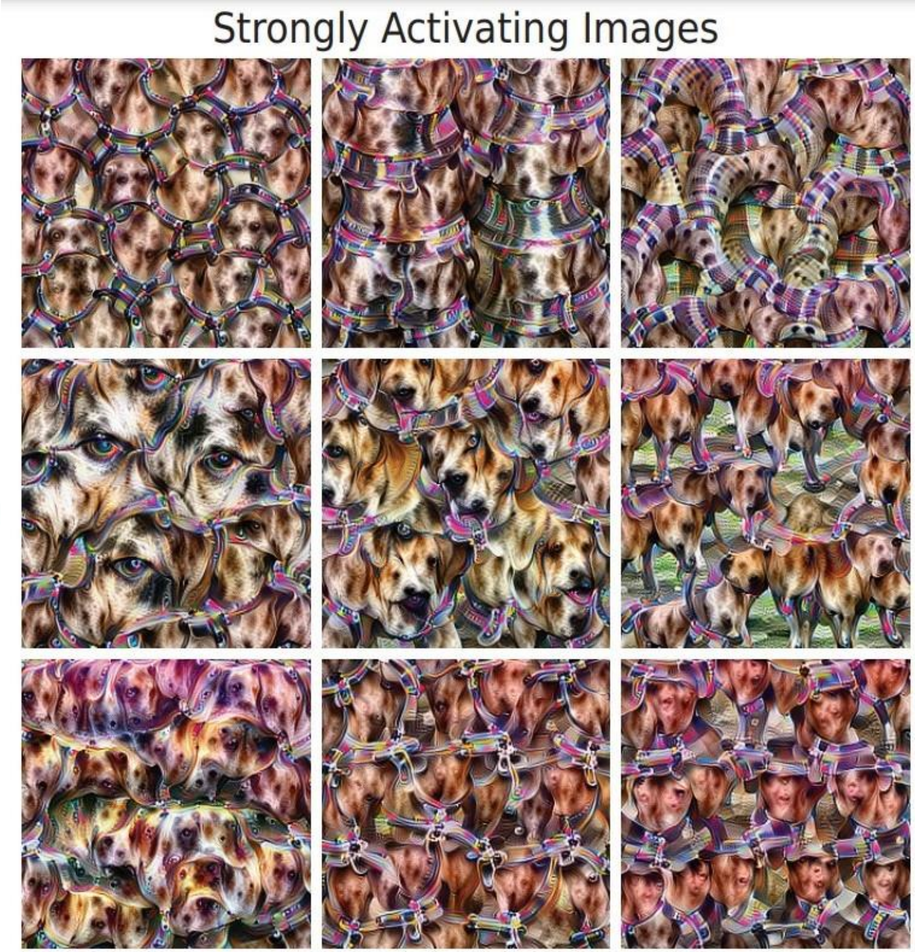
- 3.1 Technical Facts

- 3.2 The Task

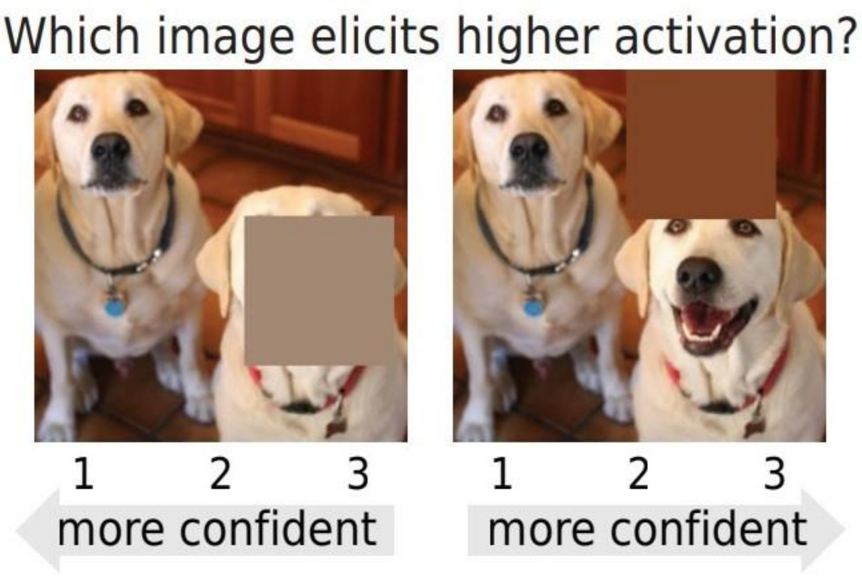
3.2 The Task



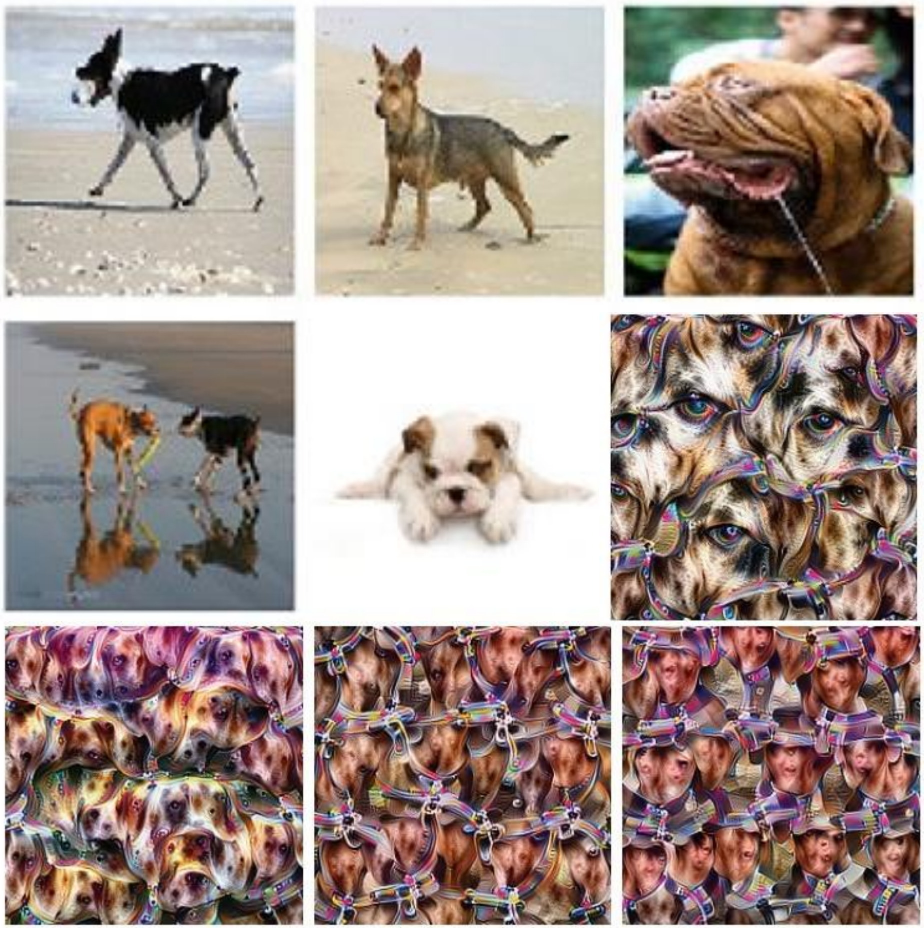
Synthetic reference image class



3.2 The Task

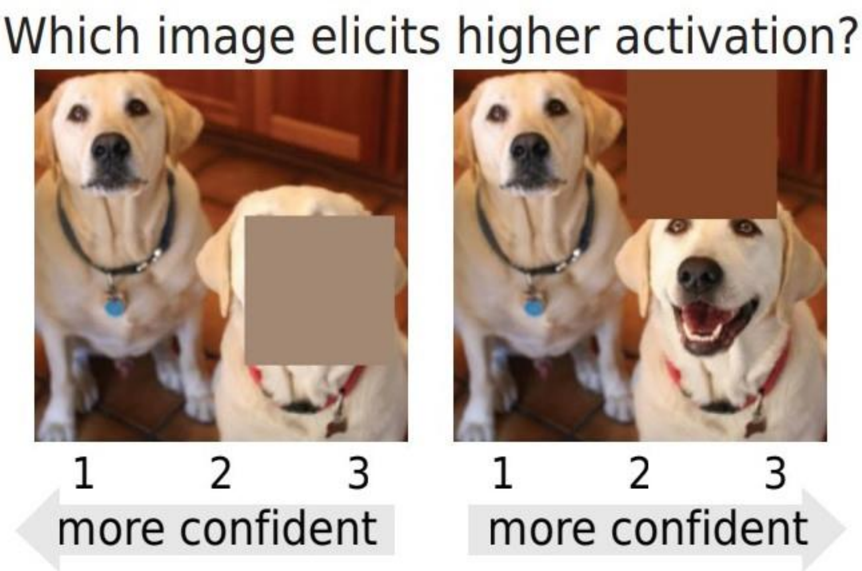


Mixed reference image class



Muhammad Atta Othman Ahmed: An efficient deep convolutional nn for visual image classification dogs
Pictures from ImageNet

3.2 The Task



Non reference image class

Blur reference image class

Images from focusedcollection.com and depositphotos.com, blurred with befunky online photo editor

3. The Experiment

- 3.1 Technical Facts
- 3.2 The Task
- 3.3 The Setup
  - 3.3.1 Experiment-Design
  - 3.3.2 Ensuring High Quality Data

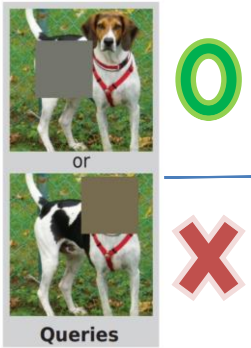
3.3.1 The Experiment-Design

For each class of reference images…

… data is collected from 50 MTurk participants.

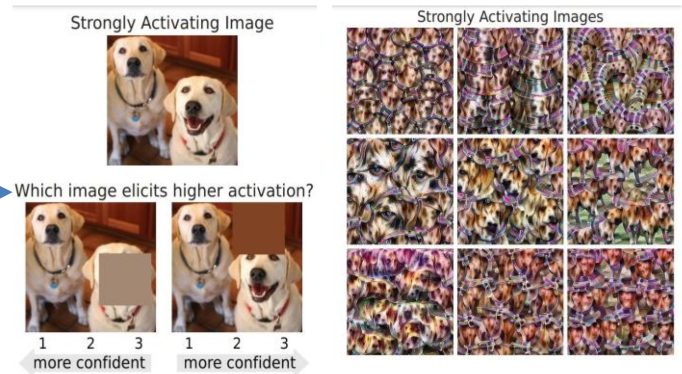
These 50 each do:

An Instruction Trial



4 Practice Trials

18 Main Trials with 3 Catch Trials



3.3.1 The Experiment-Design

For the Next Class of Reference Images…

… Data is collected from a different Subject

Between-Subject Design

…

3.3.2 Ensuring High Quality Data

- Exclusion Criteria:
  - Time to read instructions
  - Time for whole experiment
  - Performance Threshold for Catch Trials
  - Answer Variability

- Small financial compensation

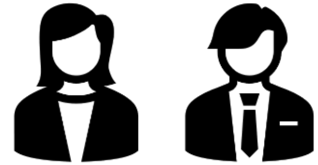
- Participants only from English speaking countries to ensure that instructions are understood

3. The Experiment

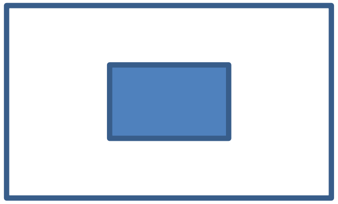
- 3.1 Technical Facts
- 3.2 The Task
- 3.3 The Execution Design
- 3.4 Baselines

3.4 Baselines

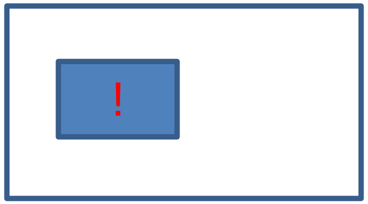
- 1. Expert Baseline



- 2. Center Baseline



- 3. Primary Object Baseline



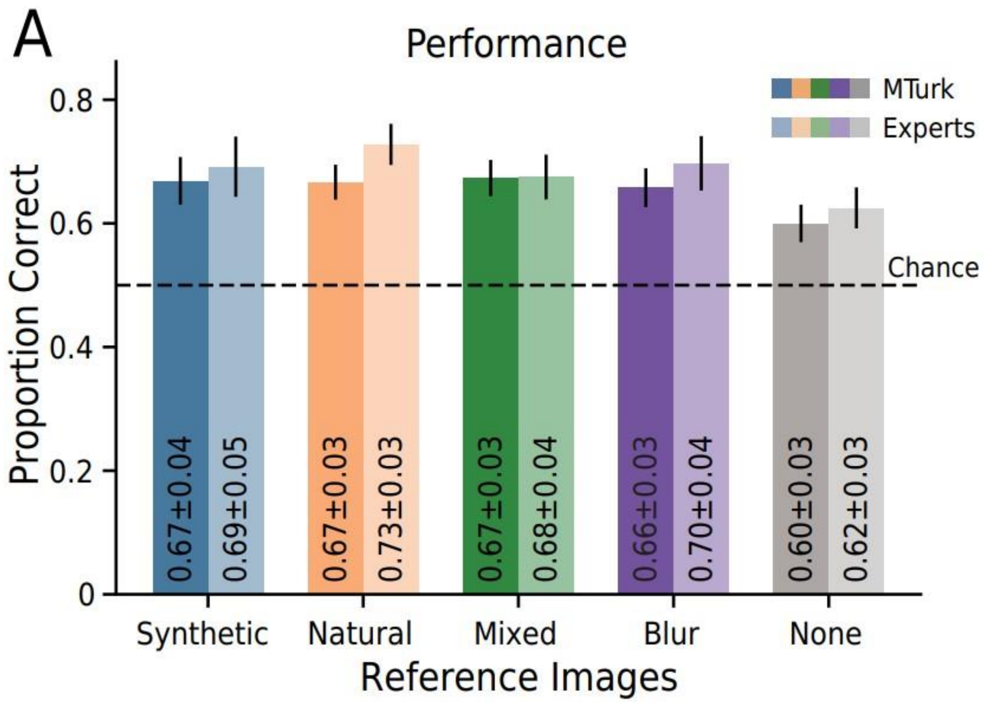
- 4. Variance Baseline

- 5. Saliency Baseline

4. Results

- 4.1 Reference Image Comparison

4.1 Reference Image Comparison



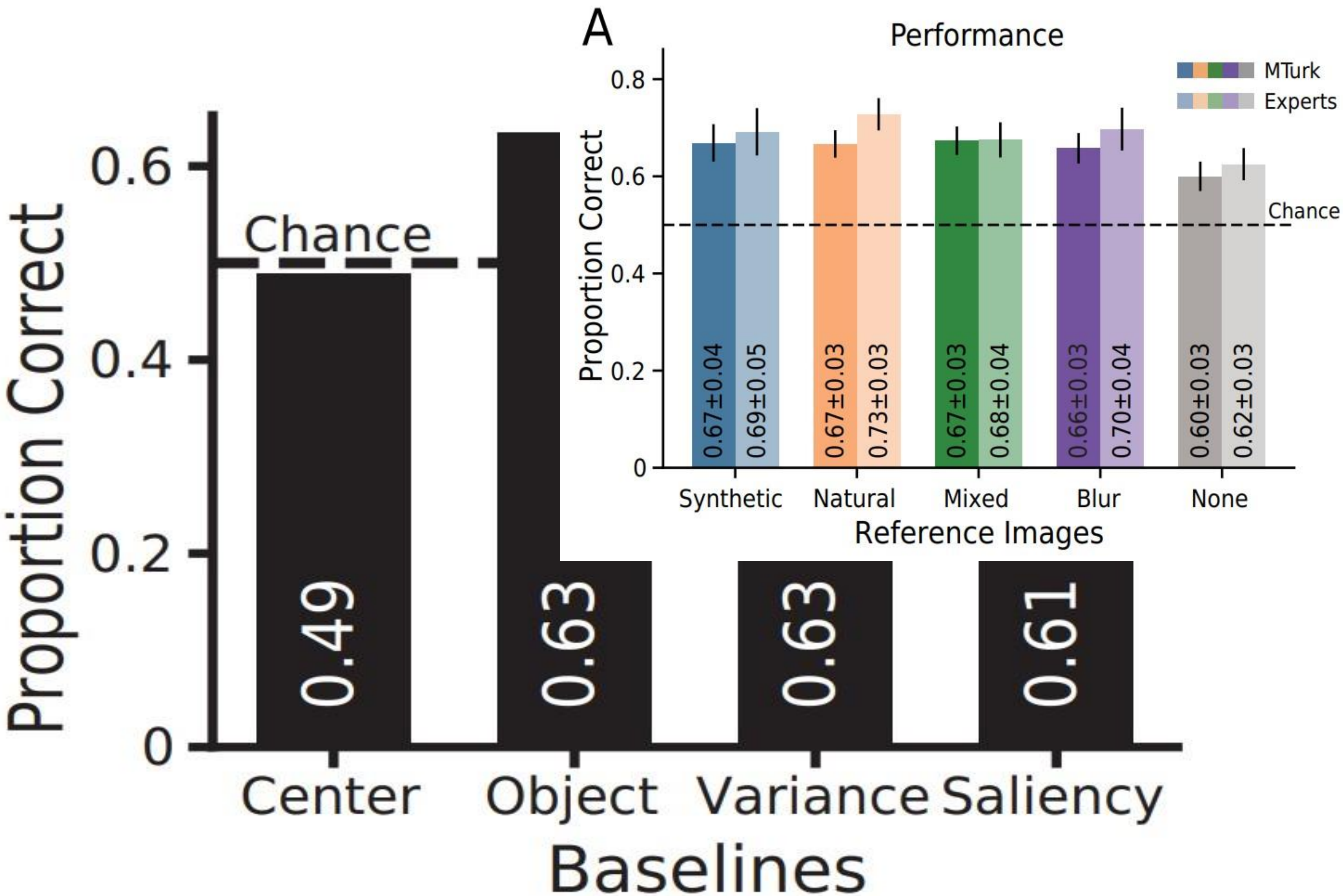
-> Sigificant difference between None-Group and the others
-> No significant performance difference between the different types of visualization

4. Results

- 4.1 Reference Image Comparison

- 4.3 Comparison with the Baselines

4.3 Comparison with the Baselines

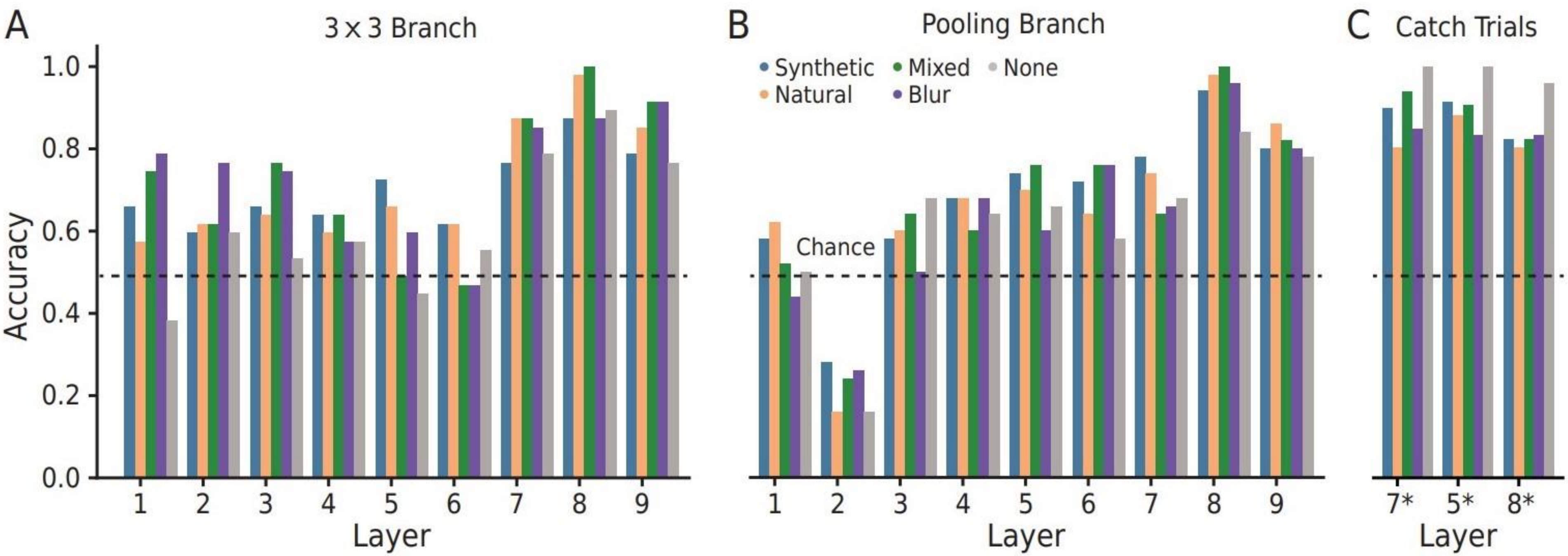


4. Results

- 4.1 Reference Image Comparison
- 4.3 Comparison with the Baselines
- 4.4 Performance Variation

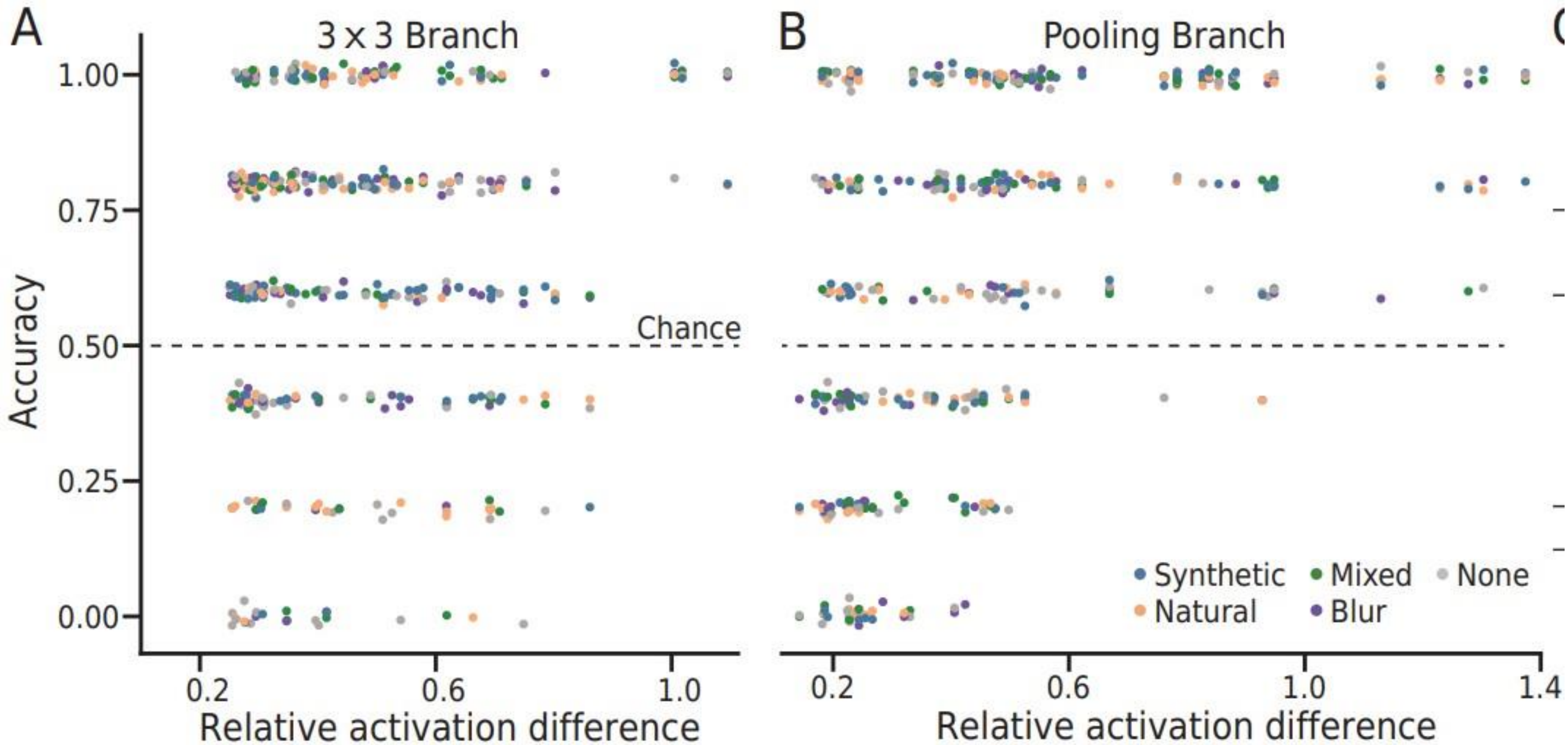
4.4 Performance Variation

- Type of visualization not very important for performance
- Systematic performance difference across different units



from layer 8 and 2 of the POOL branch, respectively.

4.4 Performance Variation



5. Conclusion

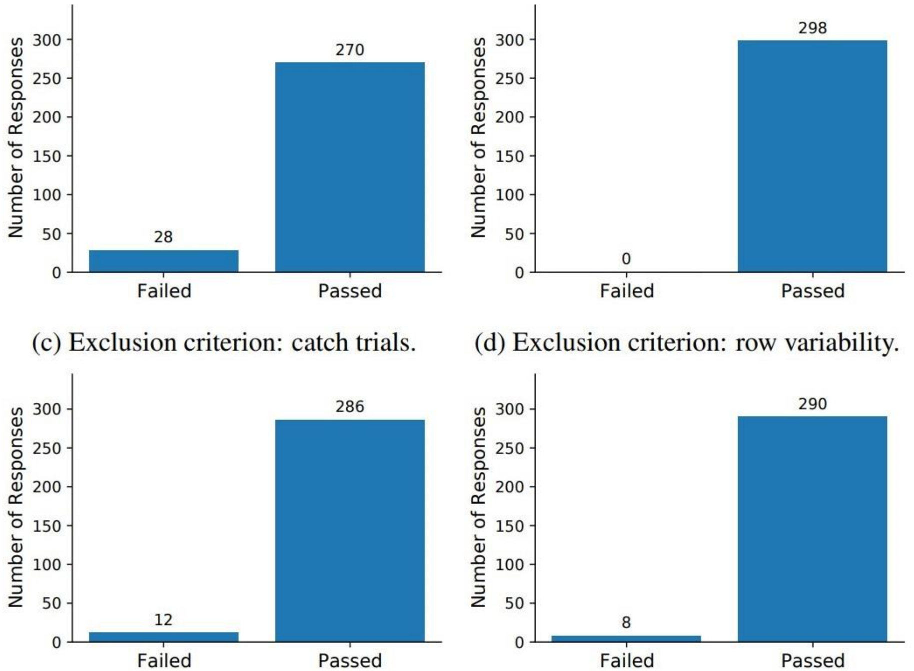
- Humans are better able to understand and predict behaviour of a CNN when provided visualization
- Images synthesized by Activation-Maximization are NOT more helpful than other kinds of visualizations
- Experiment is limited: e.g. fixed size and shape of occlusion patch
- More visualization methods could be added in the future

Discussion

- Olah et al. „Feature Visualization“
- Nguyen et al. „Understanding Neural Networks via Feature Visualization: A survey“
- Synthesizing the preferred inputs for neurons in neural networks via deep generator networks

High Quality Data!

- Performance on average very similar to the performance of the experts
- 260 of 298 participants passed the Exclusion Criteria:



(c) Exclusion criterion: catch trials. (d) Exclusion criterion: row variability.

(e) Exclusion criterion: instruction time. (f) Exclusion criterion: total response time.

- Trial-by-Trial Responses are more similar than chance would predict
- Reasonable Reaction Time

Comparison with the Baselines