Deep Neural Networks Predict Category Typicality Ratings for Images

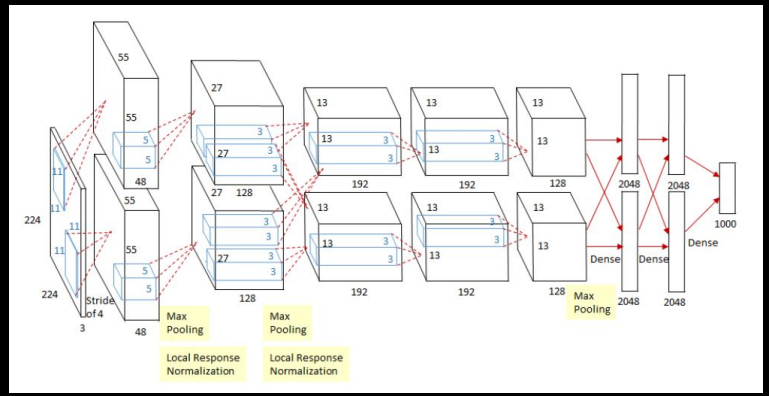
Lake et al. 2015

Orhan Soyuhos

Introduction

Background

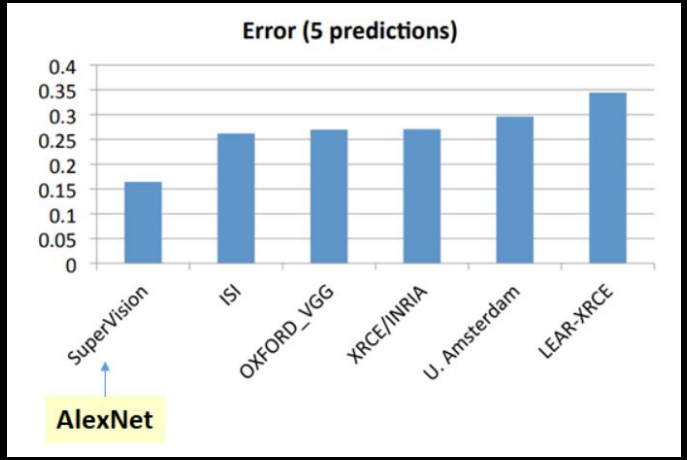
- The latest generation of neural networks has made major performance advances in object categorization.
- They can either correctly identify the object category or produce a series of plausible guesses.
- AlexNet architecture
  - (Krizhevsky, Sutskever, & Hinton, 2012)



https://mc.ai/paper-review-of-alexnet-caffenet-winner-in-ilsvrc-2012-image-classification/

AlexNet

- a massive convolutional neural network
- trained on 1.2 million images to discriminate between 1000 different object categories.
- the winner of 2012 ImageNet ILSVRC competition
  - by making approximately 40% fewer errors than the next best competitor.

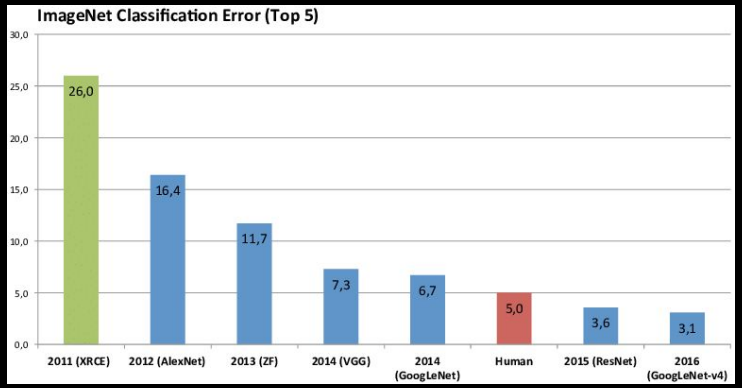


https://mc.ai/paper-review-of-alexnet-caffenet-winner-in-ilsvrc-2012-image-classification/

Next years

- In the 2013 and 2014 ImageNet competitions, virtually all of the competitors used deep convnets at least partially inspired by the AlexNet architecture.

- The best 2014 convnet (Szegedy, Liu, et al., 2014) only slightly behind human-level performance.



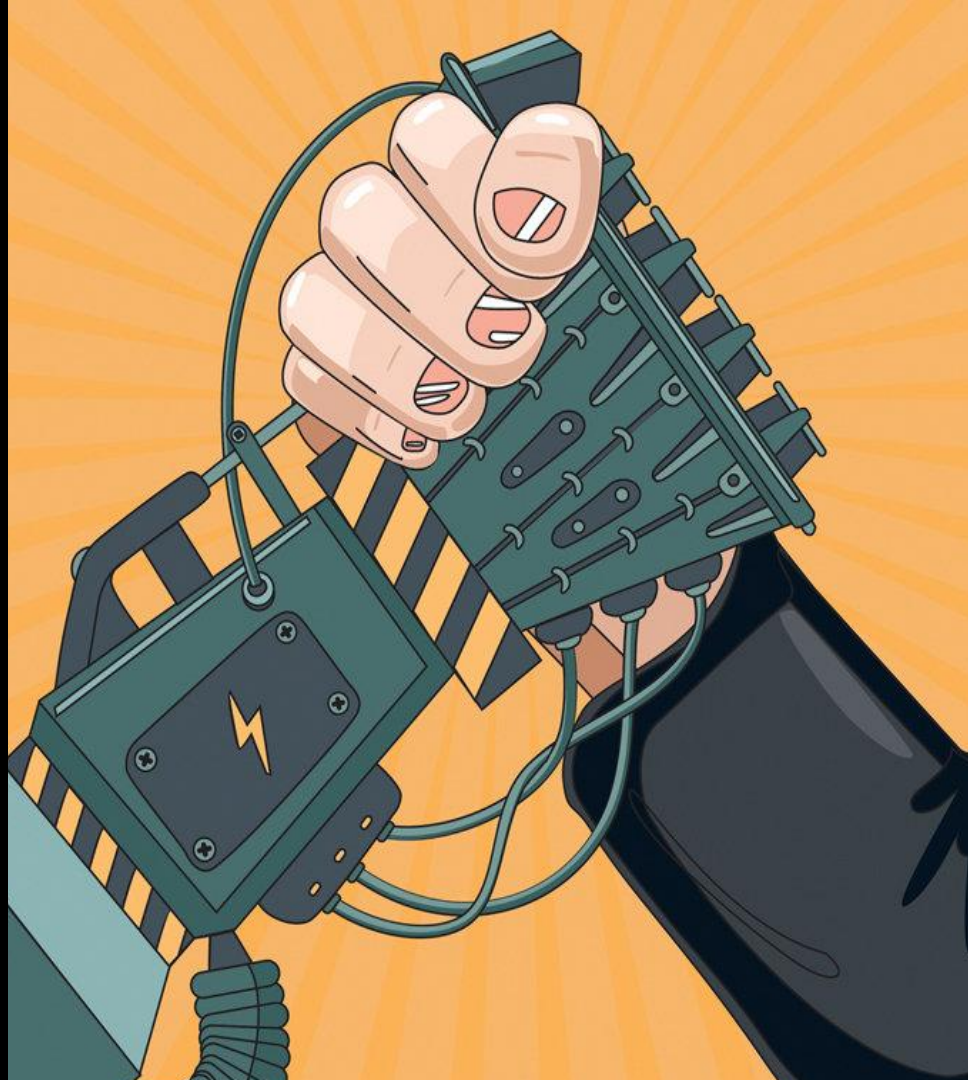
https://www.edge-ai-vision.com/2018/07/deep-learning-in-five-and-a-half-minutes/

Relation to cognitive science

- These advances should interest the cognitive science community, especially since categorization is a foundational problem and some leading models are neural networks.
  - (e.g., Kruschke, 1992; Love, Medin, & Gureckis, 2004)

- **However,** there has been little work evaluating the newest generation of neural networks as potential cognitive models or as large-scale tests of existing psychological theories.

This paper offers a first step towards this goal by using convnets to predict human typicality ratings from raw naturalistic images.



Typicality ratings?

- They reflect the graded structure of concepts.

- “... the more typical an exemplar is of a category, the more quickly it is verified to be a member of that category” (Gruenenfelder, 1984).

Which one is a more typical “dog”?

hairless Chihuahua [chi-wah-wah] or Golden Retriever

People rate a Golden Retriever as a more typical “dog” than a hairless Chihuahua.

Which one is a more typical “fish”?

Goldfish or Shark

People rate a goldfish as a more typical “fish” than a shark.

for any task that requires relating an item to its category, typicality will influence performance, whether it is

the speed of categorization,
ease of production,
ease of learning,
usefulness for inductive inference,
or word order in language

(Murphy, 2002)

However,

- There are reasons to suspect that convnets may not see the same typicality structure in images that people do, despite approaching human-level classification performance.

TWO REASONS

Firstly,

- The model parameters are trained strictly to optimize its ability to predict category labels, as opposed to predicting missing features or building a generative model of the data.

- It may be hard to learn prototypes with this objective:
  - laboratory studies with human learners show that it **(i.e. to predict category labels)** discourages people from abstracting category prototypes when compared to feature prediction tasks.
    - (Yamauchi & Markman, 1998; Yamauchi, Love, & Markman, 2002)

Secondly,

- Recent work has shown it is easy to construct adversarial images that fool convnets but are easily recognized by people.
  - (Szegedy, Zaremba, et al., 2014)

- By modifying the image slightly, the model can be induced to mistake any image for any other category with an arbitrary degree of confidence.

- **Nonetheless,**
  - these types of deformations must be rare occurrences in real images since the classifier generalizes well to unseen images.

If convnets predict human typicality, there would be implications for current psychological theories.

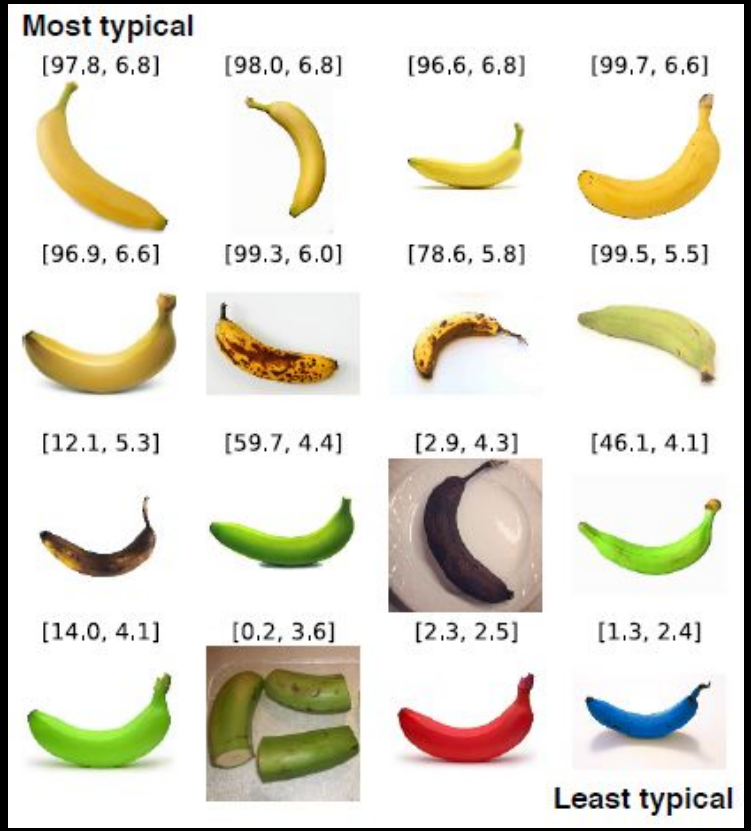
e.g. by testing different models on the same massive dataset, we are able to explore classic questions of **whether aspects of conceptual structure are bottom-up reflections of the world or top-down impositions by the mind.**

Methods

People are asked to rate a collection of images for category typicality, and three convnet architectures and a baseline system are tested on their ability to predict these ratings.

Stimuli

- Typicality ratings were collected for eight categories from the ImageNet challenge:
  - banana, bathtub, coffee mug, envelope, pillow, soap dispenser, table lamp, and teapot
- They selected a set of 16 new images from each class that do not appear in the ImageNet training set.
  - chosen via Google searches to span a maximum range of variation while focusing on a single, large, unoccluded object from a standard view.


Most typical
[97.8, 6.8]
[98.0, 6.8]
[96.6, 6.8]
[99.7, 6.6]
[96.9, 6.6]
[99.3, 6.0]
[78.6, 5.8]
[99.5, 5.5]
[12.1, 5.3]
[59.7, 4.4]
[2.9, 4.3]
[46.1, 4.1]
[14.0, 4.1]
[0.2, 3.6]
[2.3, 2.5]
[1.3, 2.4]
Least typical


[60.6, 6.6]
[58.5, 6.6]
[57.3, 6.6]
[66.5, 6.2]
[72.0, 6.1]
[80.7, 6.0]
[9.5, 5.9]
[35.4, 5.7]
[67.6, 5.6]
[63.0, 5.2]
[9.8, 3.2]
[16.4, 3.1]
[1.0, 3.0]
[1.5, 2.9]
[1.0, 2.8]
[9.1, 2.4]


[51.6, 6.7]
[80.9, 6.7]
[85.0, 6.5]
[10.5, 6.5]
[34.8, 6.4]
[84.3, 6.3]
[54.3, 6.2]
[38.3, 4.9]
[4.2, 4.8]
[22.8, 4.7]
[35.8, 4.6]
[61.2, 4.5]
[4.8, 4.3]
[1.3, 3.6]
[3.3, 3.5]
[18.0, 3.1]


[20.7, 6.8]
[1.6, 6.7]
[20.3, 6.7]
[6.5, 6.7]
[96.3, 6.5]
[94.5, 6.3]
[99.8, 6.2]
[93.6, 5.8]
[99.7, 5.8]
[0.6, 5.5]
[0.2, 4.6]
[0.5, 4.4]
[0.3, 4.2]
[0.0, 4.2]
[0.0, 4.1]
[0.2, 3.4]

Behavioral experiment

- Human typicality ratings were collected:
  - 30 participants in the USA
- Each participant rated every image from all 16 categories.
  - “How well does this picture fit your idea or image of the category?”
    - from “1” to “7”
- Participants viewed a grid of all of these images before beginning each category.

Convolutional networks

- They tested three different convnet architectures:
  - OverFeat (Sermanet et al., 2014a)
  - AlexNet (Krizhevsky et al., 2012)
  - GoogLeNet (Szegedy, Liu, et al., 2014).

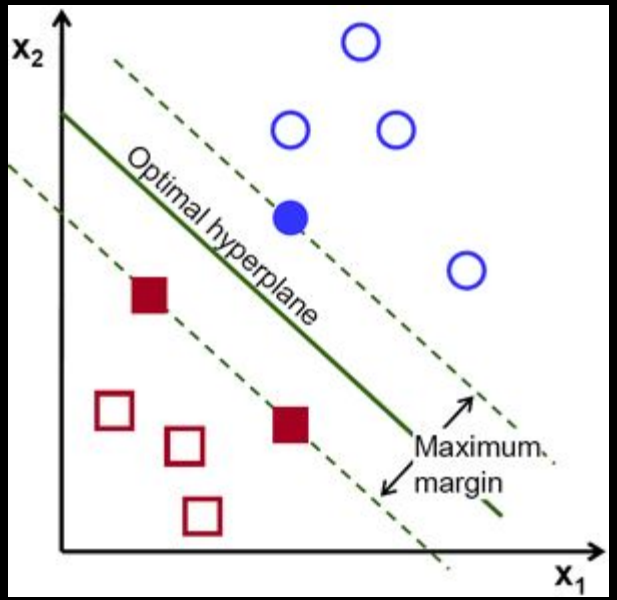
- While both OverFeat and GoogLeNet are derivatives of AlexNet, GoogLeNet is deeper and uses more sophisticated multi-resolution modules.

- ImageNet contests
  - OverFeat produced an top-five error rate of 14.2% in the 2013 contest.
    - for over 85 percent of test images, the correct label appeared in the top five guesses.
  - AlexNet achieved an error rate of 16.4% in 2012.
  - GoogLeNet achieved 6.7% in 2014.

They assume that **typicality ratings** are related to the strength of a model's classification response to the category of interest.

Baseline SIFT model

- They also test a non-convnet baseline.
  - using code from the ImageNet 2010 challenge (Russakovsky et al., 2014)
- Eight one-versus-all linear SVMs were trained.
  - – one for each category in the rating task
- SVM confidence was used to predict typicality.



Results

Results

- The mean typicality rating for each image was computed by averaging across participants.

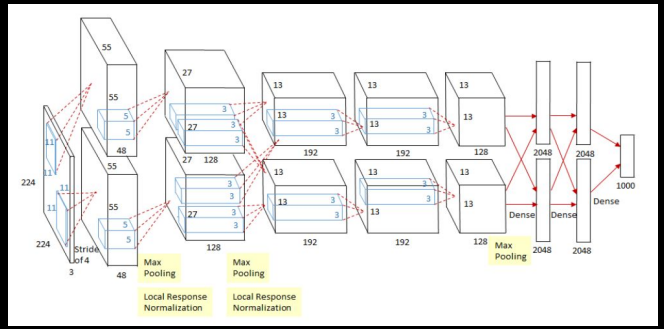
- Since human ratings were not expected to scale linearly with model ratings, Spearman's rank correlation (ρ) was used to assess fit between human and model ratings.

- Also, the reliability of the human typicality ratings was assessed with a split-half correlation.
    - Across 25 random splits, the average reliability across all eight categories was $\rho = 0.92$

Results

- The convnets predicted human ratings about equally well regardless of whether raw or contrast typicality was used.
    - **raw typicality:**
        - raw category score from the last fully-connected layer (4096 units)
        - a measure of similarity between the prototype and top-level hidden unit activations.

$$y_j = \sum_{i=1}^{4096} w_{ij} x_i$$

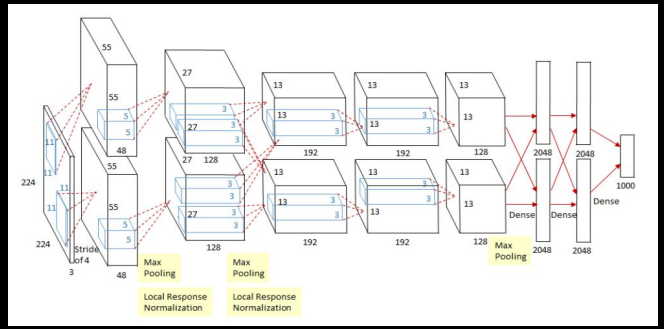


https://mc.ai/paper-review-of-alexnet-caffenet-winner-in-ilsvrc-2012-image-classification/

Results

- The convnets predicted human ratings about equally well.
    - **contrast typicality:**
        - normalized classification score from the softmax layer
            - produces a probability distribution over the j = 1:1000 classes
        - after computing the raw score, examples that score highly for other categories are penalized.

$$z_j = \frac{e^{y_j}}{\sum_{j=1}^{1000} e^{y_j}}$$

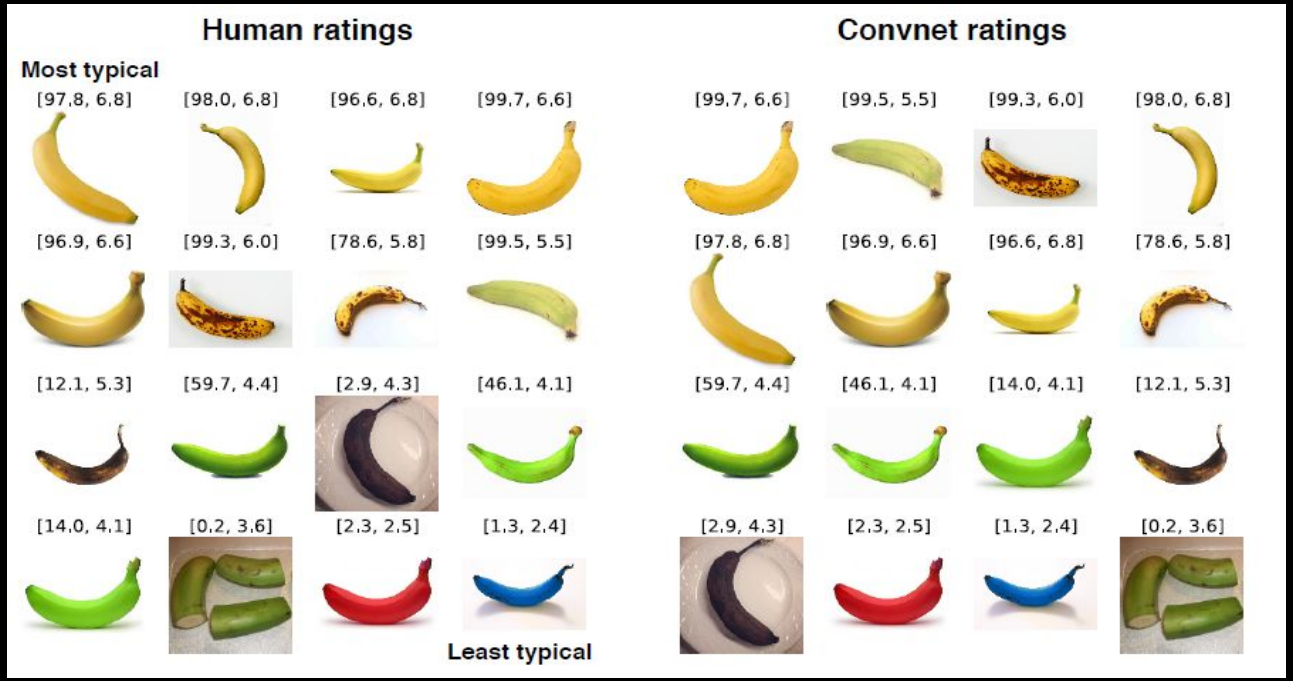


https://mc.ai/paper-review-of-alexnet-caffenet-winner-in-ilsvrc-2012-image-classification/

Table 1: Rank correlations for human and machine typicality.

Category	OverFeat	AlexNet	GoogLe	Combo	SIFT
Banana	0.82	0.8	0.73	0.84	0.4
Bathtub	0.68	0.74	0.48	0.78	0.39
Coffee mug	0.62	0.84	0.84	0.85	0.63
Envelope	0.79	0.62	0.75	0.78	0.38
Pillow	0.67	0.55	0.69	0.59	0.11
Soap Disp.	0.74	0.79	0.82	0.75	0.09
Table lamp	0.69	0.8	0.7	0.83	0.48
Teapot	0.38	0.21	0.07	0.28	-0.23
Average	**0.67**	**0.67**	**0.63**	**0.71**	**0.28**

The full set of results for contrast typicality ratings is shown.
A combination model that averages the predictions of the three convnets showed a slightly higher correlation of r = 0:71.
The convnets predicted human ratings about equally well.



Typicality ratings from people and OverFeat are shown.
The values above each image [x1 ; x2] show the convnet contrast typicality rating and the mean participant rating.

Results

- To gain further insight into how the convnets predict typicality, they analyzed the structure present at each layer of processing.

- They calculated the correlation between human and convnet typicality ratings as a function of network depth.

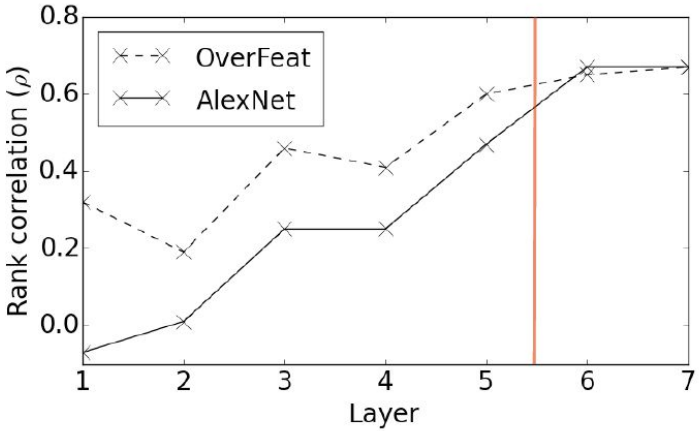


Figure 3: Correlation between human and convnet typicality ratings as a function of network depth. The red line indicates a transition from convolutional (1-5) to standard layers (6-7).

Performance steadily improves with depth.

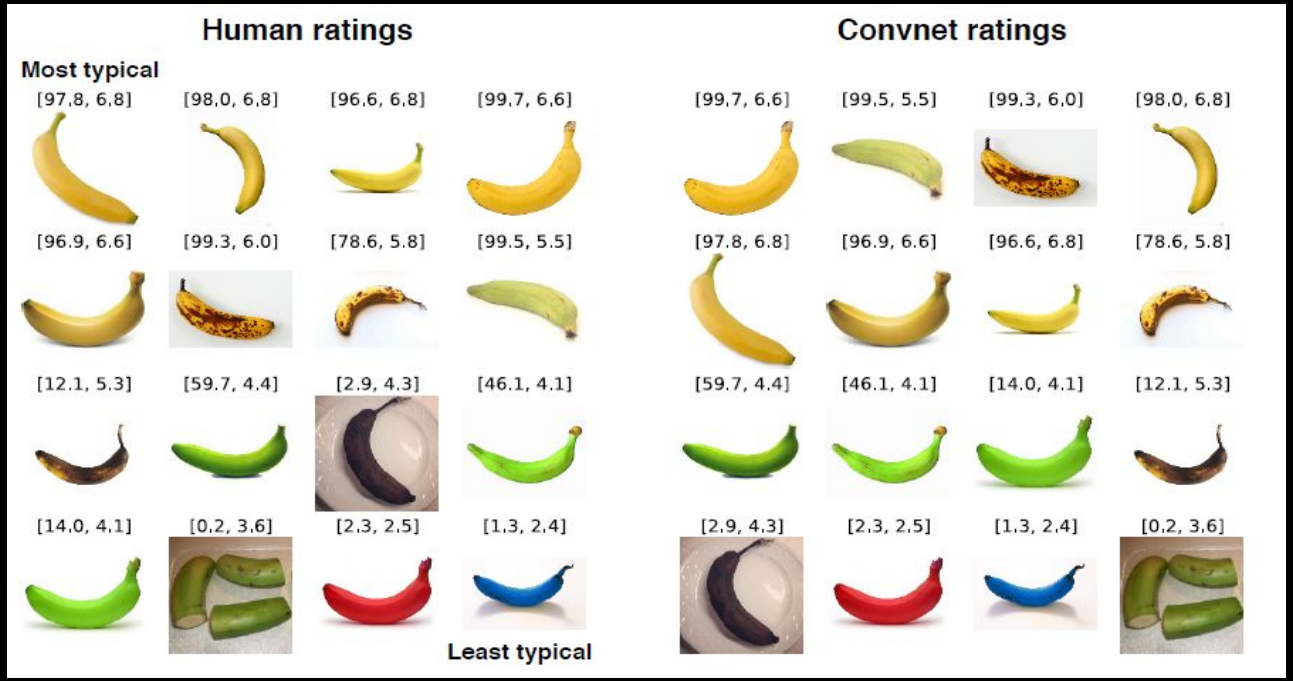
Discussion

Discussion

- The results suggest that deep convnets learn graded categories that can predict human typicality ratings, at least for some types of everyday categories.

- The low correlations from the SIFT baseline suggest that human typicality ratings are not just a property of any classifier trained on a large dataset with reasonable features.

Category	OverFeat	AlexNet	GoogLe	Combo	SIFT
Banana	0.82	0.8	0.73	0.84	0.4
Bathtub	0.68	0.74	0.48	0.78	0.39
Coffee mug	0.62	0.84	0.84	0.85	0.63
Envelope	0.79	0.62	0.75	0.78	0.38
Pillow	0.67	0.55	0.69	0.59	0.11
Soap Disp.	0.74	0.79	0.82	0.75	0.09
Table lamp	0.69	0.8	0.7	0.83	0.48
Teapot	0.38	0.21	0.07	0.28	-0.23
Average	**0.67**	**0.67**	**0.63**	**0.71**	**0.28**



Typicality ratings from people and OverFeat are shown.
The values above each image [x1 ; x2] show the convnet contrast typicality rating and the mean participant rating.

For bananas, **people** may have ranked the images based on their similarity to an “ideal”; in this case, a yellow spot-free banana.

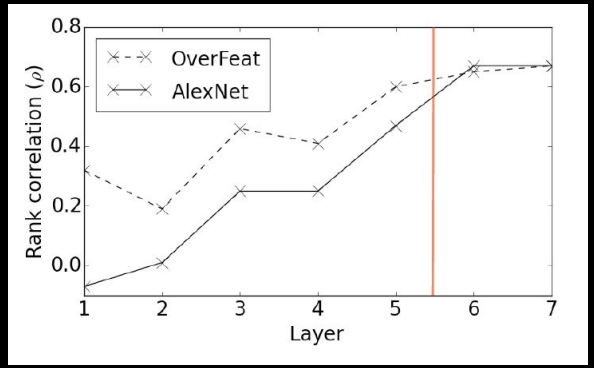
In contrast, **OverFeat** rated a greenish plantain and a spotted banana about as highly as more ideal bananas.

This suggests that typicality may be more a **top-down** imposition from the mind rather than a **bottom-up** property of visual experience with bananas (most bananas are not perfect).

Discussion

- The depth at which typicality emerges suggests the difficulty of extracting this structure from the raw data.

- Again, this confirms that typicality does not automatically emerge from a large dataset with simple feature extraction.
  - The data must be viewed through the right lens before the structure is apparent.



- However, the relationship between classifier performance and typicality effects remains unclear, making it difficult to isolate any unique contributions of the architectures beyond their abilities as classifiers.

LIMITATIONS

Conclusion

Conclusions

- This paper evaluated the ability of convnets to predict human typicality ratings for images of category examples.
  - convnets trained on 1000-way classification were able to predict human typicality ratings.
- Different operationalizations of typicality provided equally good fits,
  - suggesting there was no particular benefit for an explicitly contrastive measure of typicality

Conclusions

- The role of the training data versus the model in capturing typicality,
  - simple features did not provide good prototypes for prediction even with many examples per class.

- Convnets were less sensitive to category ideals than people,
  - suggesting that feature extraction on a large dataset may not be fully sufficient for ideals to arise.

- Given their results, it may be promising to use these methods to study more fine-grained structure within categories.

Conclusions

- Whether or not convnets can match these aspects of behavior, they are still far too limited compared to the human ability to learn and use new concepts.
    - While the convnet was trained on an average of 1200 images per class, people need far less data in order to learn a new category (Lake, 2014).
- In addition, human concepts support the flexible use of the same knowledge across many tasks – classification, inference, generation, and explanation – a remarkable quality that current machine learning approaches do not capture.

While the current best algorithms are limited compared to people, further exercises in understanding their synthetic psychology may serve to both advance machine learning and psychological theory.

Thank you for listening!

orhan.soyuhos@studenti.unitn.it