LSTM and the BRAIN or COGNITION

- Three different approaches:
    1. Artificial neural networks as cognitive models (e.g. of how humans might learn language or analyze visual information)
    2. Cognitive assessment of neural network models (use tools of cognitive science and linguistics to probe models' perfomance and compare to human performance)
    3. Artificial neural networks as models to understand the brain

LSTM and the BRAIN or COGNITION

1- Artificial neural networks as cognitive models (e.g. of how humans might learn language or analyze visual information)

•

LSTM and the BRAIN or COGNITION

Artificial neural networks as cognitive models (e.g. of how humans might learn language or analyze visual information)

- For Language: The first NN for language aimed to be cognitively plausible models of how language might be learned by infants (e.g., Elman's first recurrent neural networks, Rumelhart and McClelland's PDP model).

  Tasks examples: learn word-meaning associations, learn past-tense

- For Vision: perceptron aimed to model simple visual cells in the mammalian cortex. CNNs are considered as good models of how visual information is processed in the brain.

LSTM and the BRAIN or COGNITION

2- Cognitive assessment of neural network models (use tools of cognitive science and linguistics to probe models' perfomance and compare to human performance)

LSTM and the BRAIN or COGNITION

Cognitive assessment of neural network models (use tools of cognitive science and linguistics to probe models' perfomance and compare to human performance)

- Contemporary NN are focused to applications (machine translation, LLM, image classification, object detection), with looser relation to cognitive plausibility.
- However, cognitive tools have been used to assess NN performance.

Paper 1

Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies

Tal Linzen[1,2] **Emmanuel Dupoux**[1]
LSCP[1] & IJN[2], CNRS,
EHESS and ENS, PSL Research University
{tal.linzen,
emmanuel.dupoux}@ens.fr

Yoav Goldberg
Computer Science Department
Bar Ilan University
yoav.goldberg@gmail.com

Abstract

The success of long short-term memory (LSTM) neural networks in language processing is typically attributed to their ability to capture long-distance statistical regularities. Linguistic regularities are often sensitive to syntactic structure; can such dependencies be captured by LSTMs, which do not have explicit structural representations? We begin addressing this question using number agreement in English subject-verb dependencies. We

(Hochreiter and Schmidhuber, 1997) or gated recurrent units (GRU) (Cho et al., 2014), has led to significant gains in language modeling (Mikolov et al., 2010; Sundermeyer et al., 2012), parsing (Vinyals et al., 2015; Kiperwasser and Goldberg, 2016; Dyer et al., 2016), machine translation (Bahdanau et al., 2015) and other tasks.

The effectiveness of RNNs[1] is attributed to their ability to capture statistical contingencies that may span an arbitrary number of words. The word *France*, for example, is more likely to occur somewhere in

Objectives of the paper:

1. Stress test for LSTM: Success of long short-term memory (LSTM) neural networks in language processing is typically attributed to their ability to capture long-distance statistical regularities. Up to what point?

2. On the Cognitive Side: limits of RNN as a model of (child) language learning (as in Elman 1990,1991): there are rules that infants learn but LSTM don't learn. Therefore either LSTM is not the right model, or there is innate knowledge.

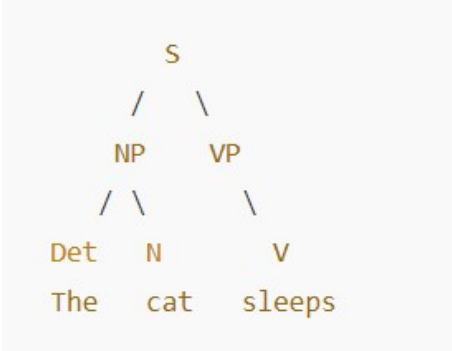
Arguments for structured representations

- Many word co-occurring statistics can be captured by treating sentences as an unstructured list of words: example: Paris-France
- Most naturally occurring agreement cases in the Wikipedia corpus are easy: they can be resolved without syntactic information, based only on the sequence of nouns preceding the verb

- Other dependencies **however** are sensitive to the syntactic structure of a sentence.
- Ex: the keys to the cabinet … on the table

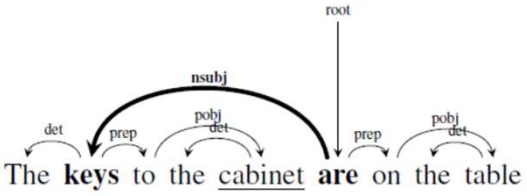
Arguments for structured syntactic representations in humans

- Given the difficulty in identifying the subject from the linear sequence of the sentence, dependencies such as subject-verb agreement serve as an argument for structured syntactic representations in humans



Example sentences- arguments for structured representations

- The **keys are** on the table.
- The **keys** to the cabinet **are** on the table.




```
                S
              /   \
            NP      VP
          / | \     / \
       Det  N   PP  V   PP
      The keys   / \   / \
               P    NP P    NP
              to   / \ on  / \
                 Det  N  Det  N
                 The cabinet The table
```

No limit to the complexity of a sentence

- The **building** on the far right that's quite old and run down **is** the Kilgore Bank Building.
- The only championship **banners** that are currently displayed within the building **are** for national or NCAA Championships.
- Yet the **ratio** of men who survive to the women and children who survive **is** not clear in this story

Attractors

Heuristic of considering only the last word or name fails

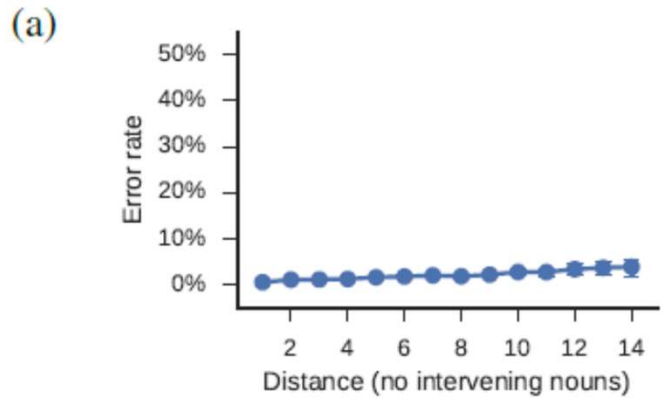
Tasks considered

- Test: Subject-verb agreement: “the kids play but the kid plays”
- Training Set=from Wikipedia

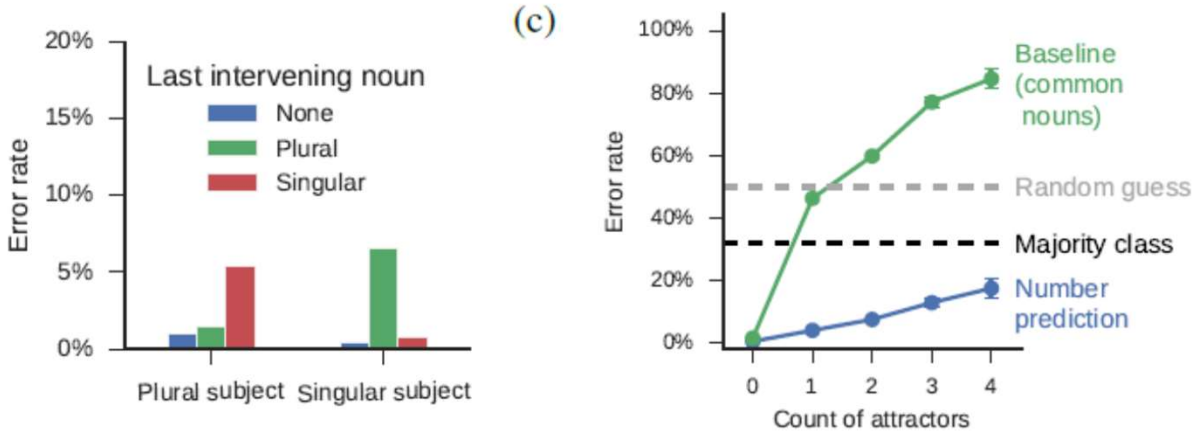
- Several training conditions:

1. Most favorable condition: training with explicit supervision directly on the task of guessing the number of a verb based on the words that preceeded it
2. Grammaticality judgment training objective (full sentences annotated as to whether or not they violate subject-verb number agreement).
3. Training of a model without any grammatical supervision, using a language modeling objective (predicting the next word)

Error rates of LSTM as a function of distance…
with explicit supervision on number prediction task



Error rates of LSTM as a function of distance… with explicit supervision on number prediction task



(Function words matter)

Results (number prediction task)

- Quantitative results indicate that most naturally occurring agreement cases in the Wikipedia corpus are easy and can be resolved without syntactic information
- This leads to high overall accuracy in all models
- The accuracy of this model (trained with supervised number prediction model) is lower on harder cases, even if it managed to recover.
- Mistakes are much more common when no overt cues to syntactic structure (function words) are available, as is the case in noun-noun compounds and reduced relative clauses -> reliant on function words. (ex “under”, “towards”, “that” …).

Related Tasks

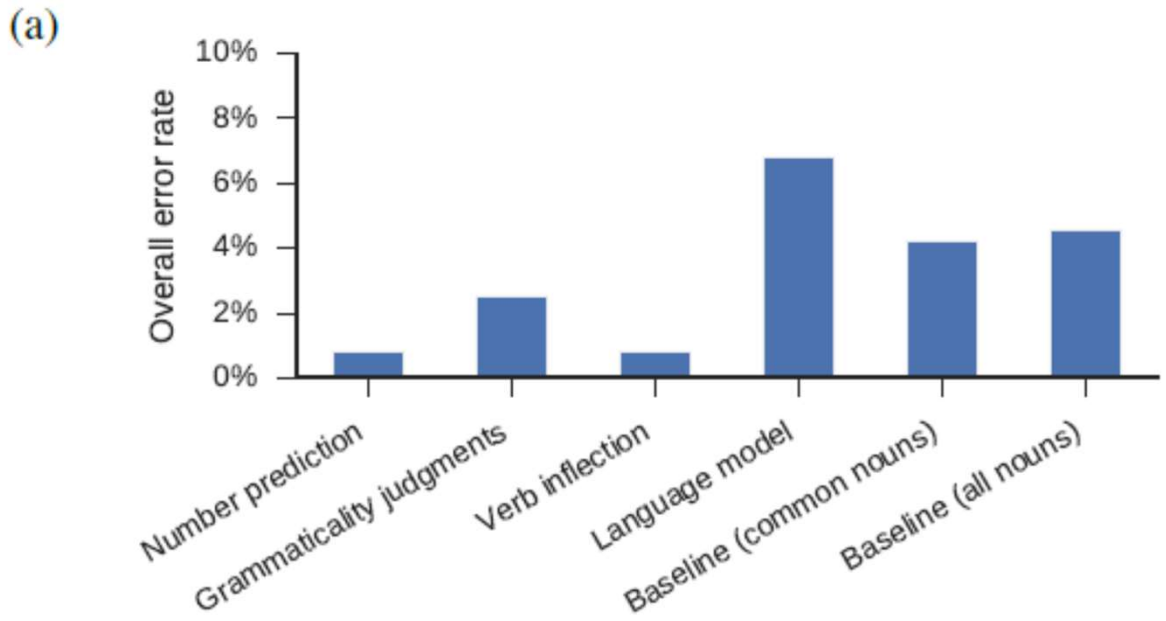
- Verb inflection (- write /writes) (access to semantics too)

Ex: work/works

- Grammaticality Judgements (weaker supervision-no syntactic clue boundaries)
- Language Model

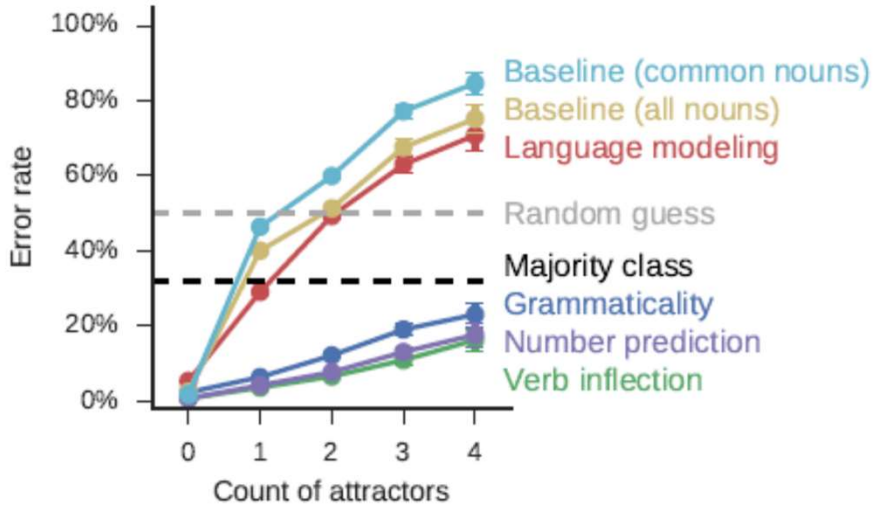
Language Models Results

Worst performer= language model



Language Models Results

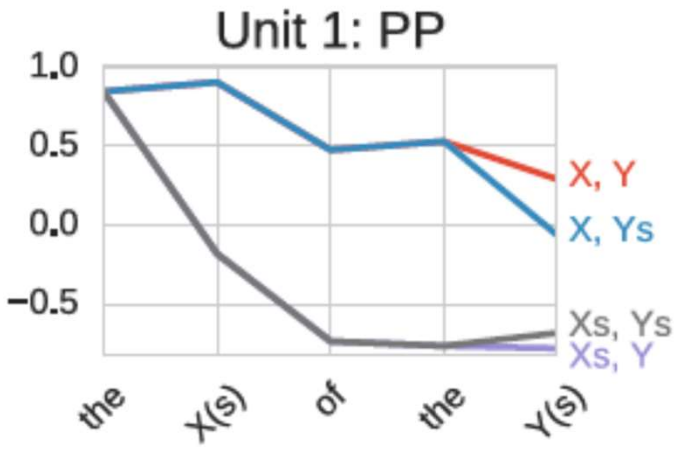
Worst performer= language model

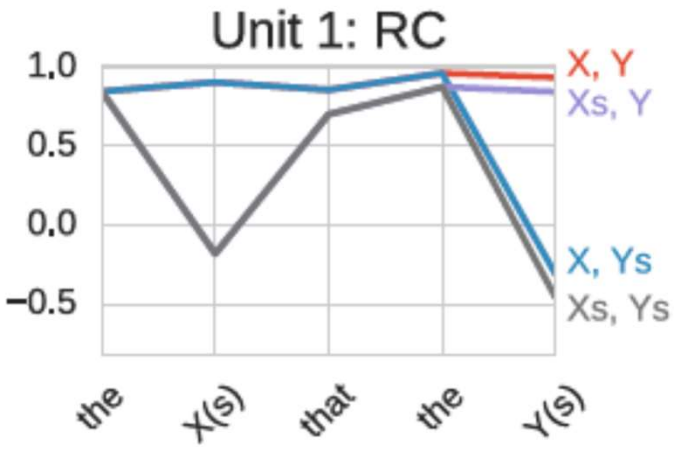


Main theoretical results:

- LSTM make errors on harder sentences (mostly if trained in the language model setting).
- Therefore, explicit supervision is necessary for learning the agreement dependency using this architecture, limiting its plausibility as a model of child language acquisition (similarly to Elman, 1990 for RNNs).

Interpreting single hidden units activations

- 

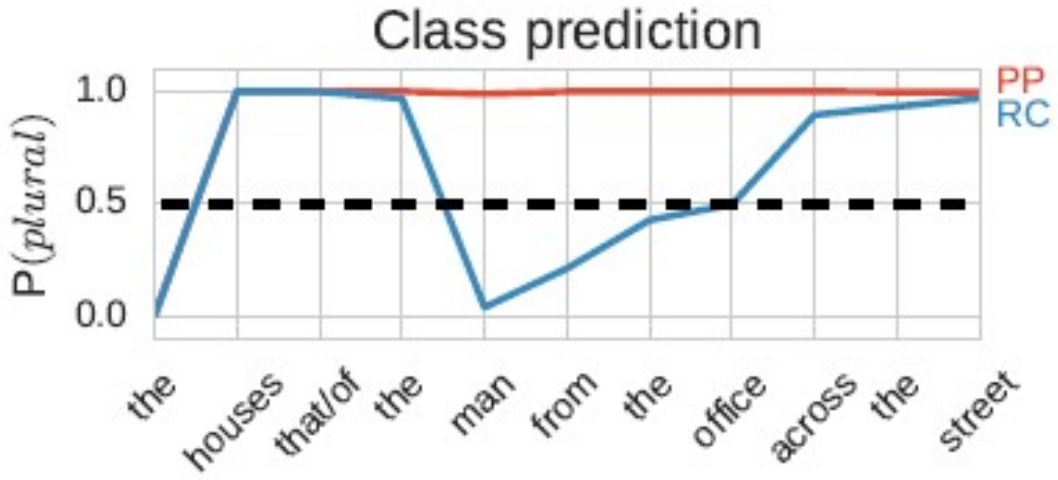


The toy(s) of the boy
2 activation patterns coherent
with first name X

The toy(s) that the boy
2 activation patterns coherent
with second name Y

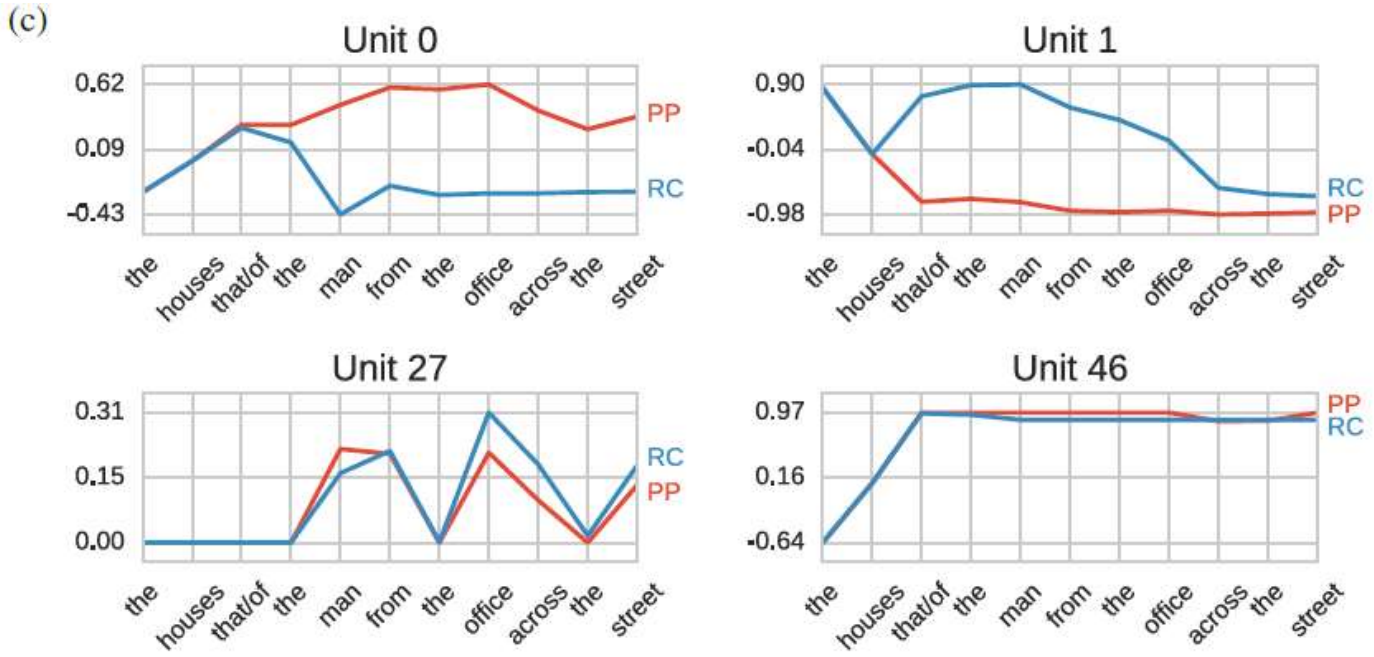
Single units activations

•



Changes to plural after «that» (not «of») but then falters.

Single units activations



Unit 0 stably remembers that there is a relative clause, 1 notices it but forgets. Unit 27 tracks the number of most recent noun. Unit 46 stores number of main clause subject

Possible improvements

- LSTM can be encouraged to develop more sophisticated generalizations by oversampling grammatically challenging training sentences (but improvement only quantitative).
- Language modeling objective is not by itself sufficient for learning structure-sensitive dependencies. Joint training objective (with supervision as in subject-verb agreement task) can be used to supplement language models on tasks for which syntax-sensitive dependencies are important.

Other Neural Architectures

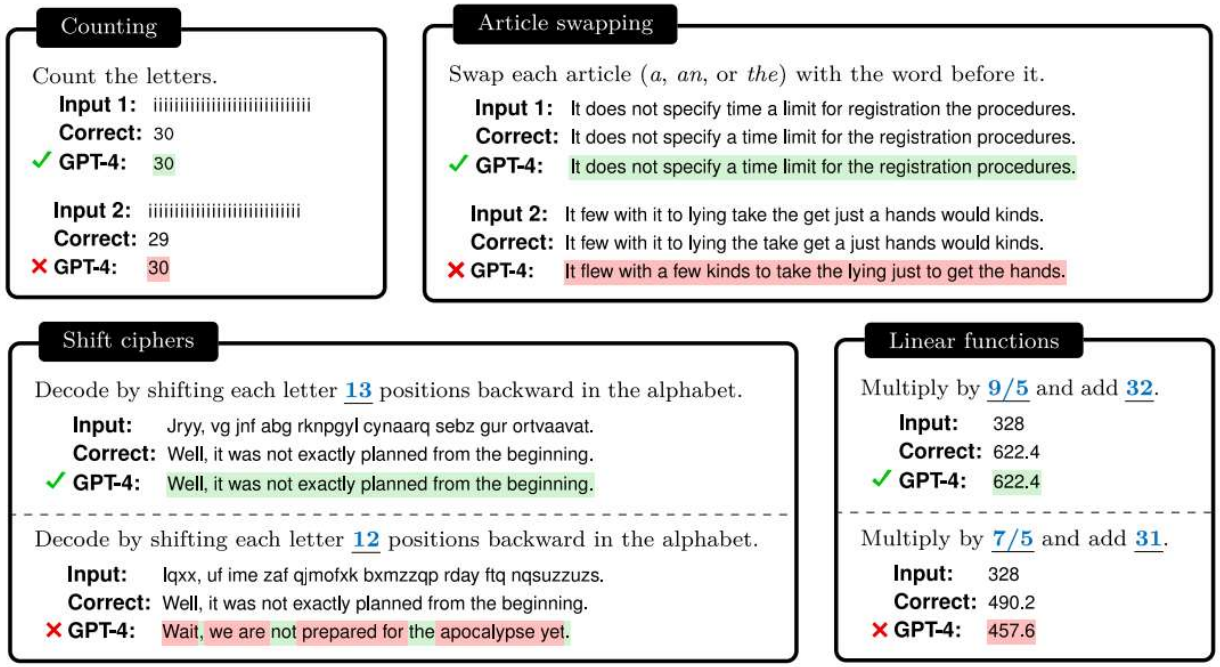
- NNs based on Transformers perform well on syntactic tasks (e.g. subject-verb agreement)
- However, they may "be right for the wrong reasons", e.g. (for BERT):

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.

T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics.

LLM learn to predict the next word, this is their strength but also weakness (sensitivity to probability and sequences):

29 lower probability than 30



(Higher probability sentence vs. lower probability)

GPT-4 struggles on some seemingly simple tasks. The GPT-4 predictions were obtained using gpt-4-0613 on the OpenAI API; other model versions (e.g., the online chat interface) may give different predictions. (note that the shift cipher with a shift of 13 is over 100 times more common in Internet text than the shift cipher with a shift of 12; and the linear function f (x) = (9/5)x + 32 is common because it is the Celsius-to-Fahrenheit conversion, while the other linear function has no special significance)

From: R. T. McCoy, S. Yao, D. Friedman, M. Hardy, and T. L. Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve. Proceedings of the National Academy of Sciences of the United States of America (PNAS),2024.

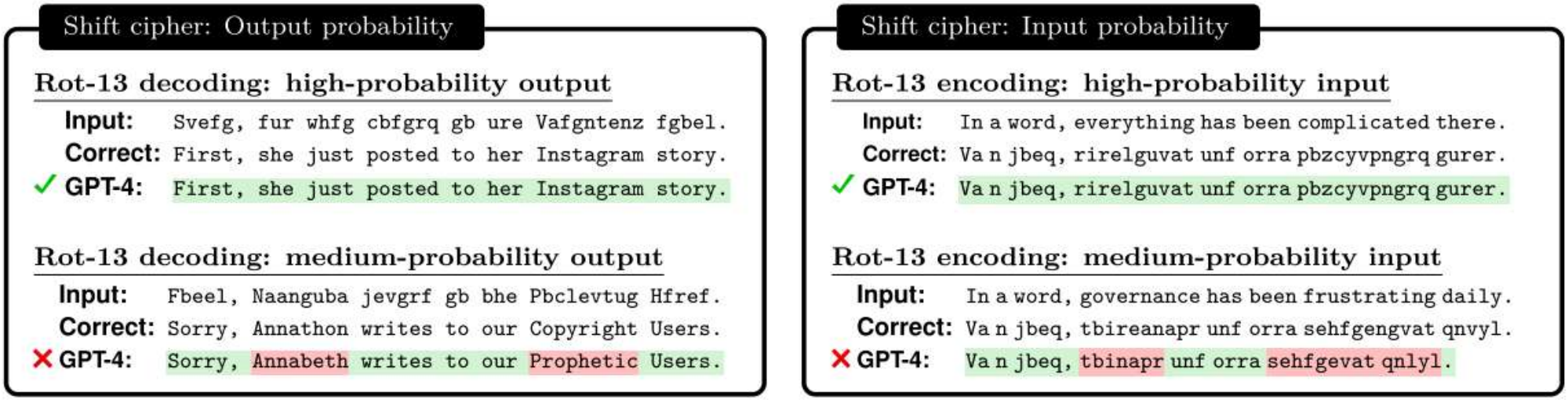
LLM learn to predict the next word, this is their strength but also weakness (sensitivity to probability and sequences):

Table 1. Effects on the performance of language models that are attributable to the fact that they are statistical next-word prediction systems

Property	Description	Example
Sensitivity to output probability	Even when the task is deterministic, LLMs achieve higher accuracy when the correct answer is high-probability text than when it is low-probability text.	When asked to reverse a sequence of words, GPT-4 gets 97% accuracy when the answer is a high-probability sentence yet 53% accuracy when the output is low probability.
Sensitivity to input probability	Even when the task is deterministic, LLMs sometimes achieve higher accuracy when the input text is high-probability than when it is low-probability, but input probability is less influential than output probability.	When asked to encode sentences in a simple cipher (rot-13), GPT-4 gets 21% accuracy when the input is a high-probability sentence yet 11% accuracy when the input is low-probability.
Sensitivity to task frequency	Even when there is no difference in the complexity of the tasks, LLMs perform better on tasks that are frequent than ones that are rare.	When asked to translate English sentences into Pig Latin, GPT-4 gets 42% accuracy when using the most common variant of Pig Latin but only 23% accuracy when using a rare variant.

From: R. T. McCoy, S. Yao, D. Friedman, M. Hardy, and T. L. Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve. Proceedings of the National Academy of Sciences of the United States of America (PNAS),2024.

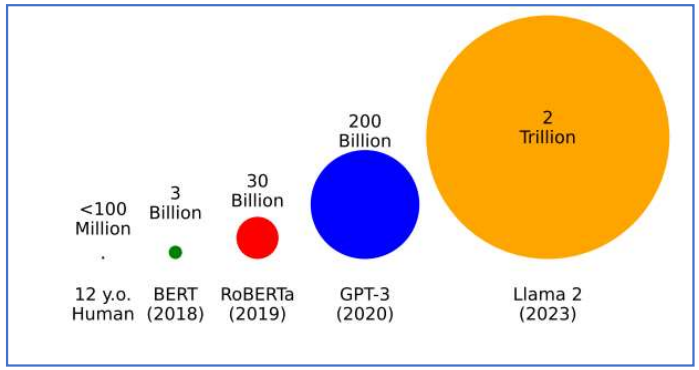
LLM learn to predict the next word, this is their strength but also weakness (sensitivity to probability and sequences):



From: R. T. McCoy, S. Yao, D. Friedman, M. Hardy, and T. L. Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve. Proceedings of the National Academy of Sciences of the United States of America (PNAS),2024.

Most of all: Argument of the Poverty of the stimulus re-adapted

- Today's best LLMs are trained on vastly more data than a child is exposed to, and some evidence suggests that a model's training dataset would need to be unrealistically large to handle some constructions in a human-like way without stronger priors



- Hence, the models will necessarily be un-humanlike (Re-edition of Chomsky's poverty of the stimulus argument).

- But: BabyLM Challenge, to train language models in low-resource data settings, where the amount of linguistic input resembles the amount received by human language learners.

- We have seen models for Language.
- Similar considerations for models for vision.

Paper 2

PNAS

Atoms of recognition in human and computer vision

Shimon Ullman[a,b,1,2], Liav Assif[a,1], Ethan Fetaya[a], and Daniel Harari[a,c,1]

[a]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 7610001, Israel; [b]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; and [c]McGovern Institute for Brain Research, Cambridge, MA 02139

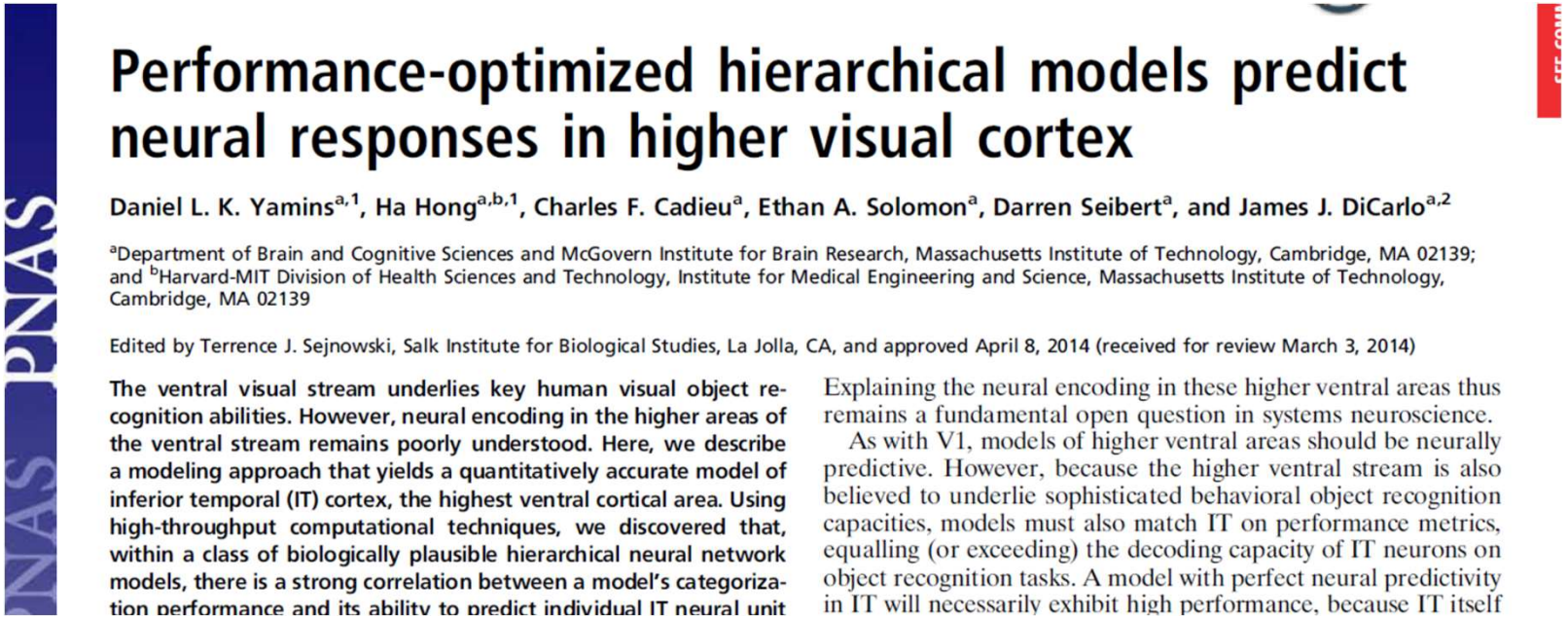
Edited by Michael E. Goldberg, Columbia University College of Physicians, New York, NY, and approved January 11, 2016 (received for review July 8, 2015)

Discovering the visual features and representations used by the brain to recognize objects is a central problem in the study of vision. Recently, neural network models of visual object recognition, including biological and deep network models, have shown remarkable progress and have begun to rival human performance in some challenging tasks. These models are trained on image examples and learn to extract features and representations and to use them for categorization. It remains unclear, however, whether the representations and learning processes discovered by current models are similar to those used by the human visual system. Here we show, by introducing and using minimal recognizable images, that the human visual system uses features and processes that are not used by current models and that are critical for recognition. We found by

descendants reaches a recognition criterion (50% recognition; results are insensitive to criterion) (*Methods* and Fig. S4). Each human subject viewed a single patch from each image with unlimited viewing time and was not tested again. Testing was conducted online using the Amazon Mechanical Turk (MTurk) (3, 4) with about 14,000 subjects viewing 3,553 different patches combined with controls for consistency and presentation size (*Methods*). The size of the patches was measured in image samples, i.e., the number of samples required to represent the image without redundancy [twice the image frequency cutoff (5)]. For presentation to subjects, all patches were scaled to 100 × 100 pixels by standard interpolation; this scaling increases the size of the presented image smoothly without adding or losing information

Deep NN and our visual system

- CNN inspired by processing of visual information in the brain
- Background theories: a major line of research tries to use CNN (and family) to predict and describe activity in areas of the brain (e.g. DiCarlo PNAS 2014)



PNAS

Performance-optimized hierarchical models predict neural responses in higher visual cortex

Daniel L. K. Yamins[a,1], Ha Hong[a,b,1], Charles F. Cadieu[a], Ethan A. Solomon[a], Darren Seibert[a], and James J. DiCarlo[a,2]

[a]Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Harvard-MIT Division of Health Sciences and Technology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139

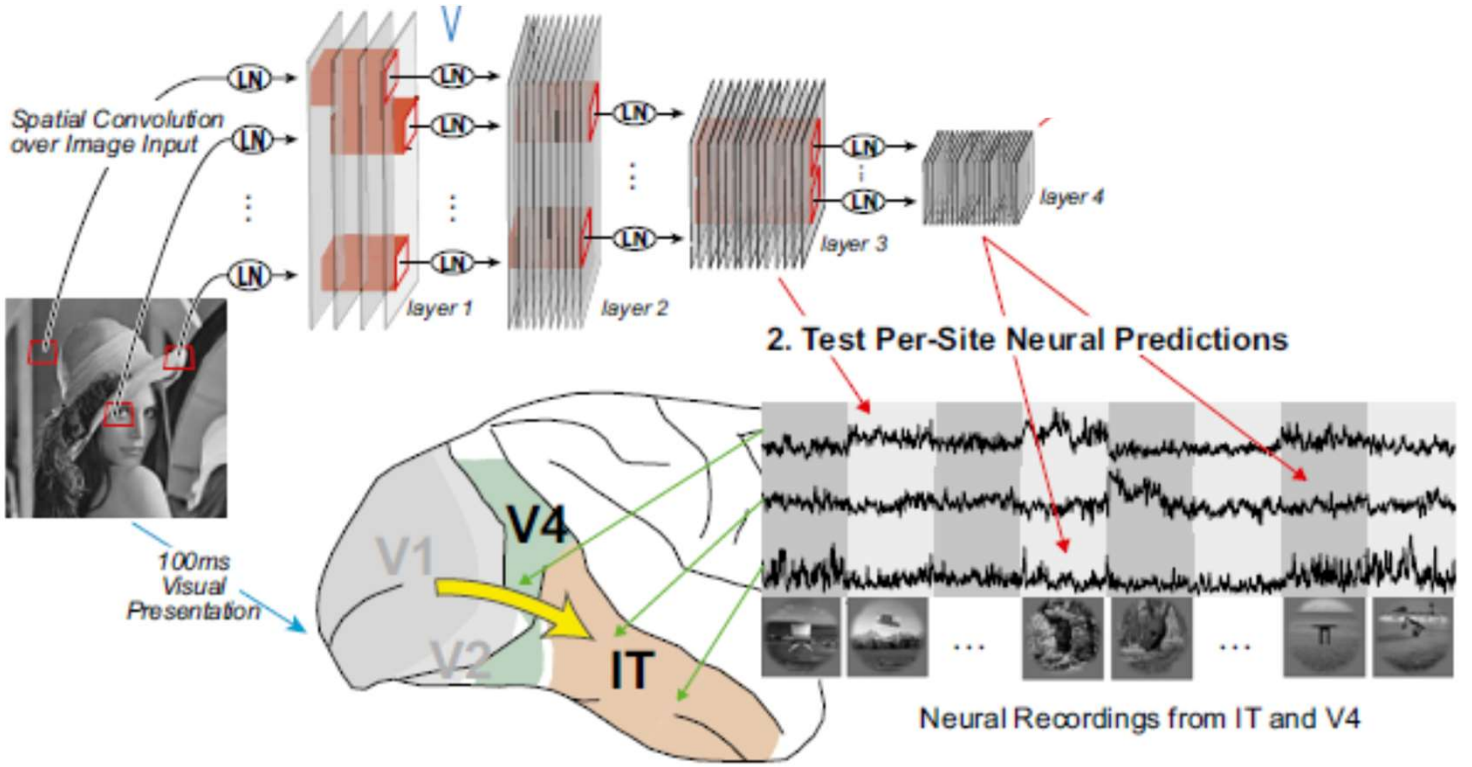
Edited by Terrence J. Sejnowski, Salk Institute for Biological Studies, La Jolla, CA, and approved April 8, 2014 (received for review March 3, 2014)

The ventral visual stream underlies key human visual object recognition abilities. However, neural encoding in the higher areas of the ventral stream remains poorly understood. Here, we describe a modeling approach that yields a quantitatively accurate model of inferior temporal (IT) cortex, the highest ventral cortical area. Using high-throughput computational techniques, we discovered that, within a class of biologically plausible hierarchical neural network models, there is a strong correlation between a model's categorization performance and its ability to predict individual IT neural unit

Explaining the neural encoding in these higher ventral areas thus remains a fundamental open question in systems neuroscience.

As with V1, models of higher ventral areas should be neurally predictive. However, because the higher ventral stream is also believed to underlie sophisticated behavioral object recognition capacities, models must also match IT on performance metrics, equalling (or exceeding) the decoding capacity of IT neurons on object recognition tasks. A model with perfect neural predictivity in IT will necessarily exhibit high performance, because IT itself

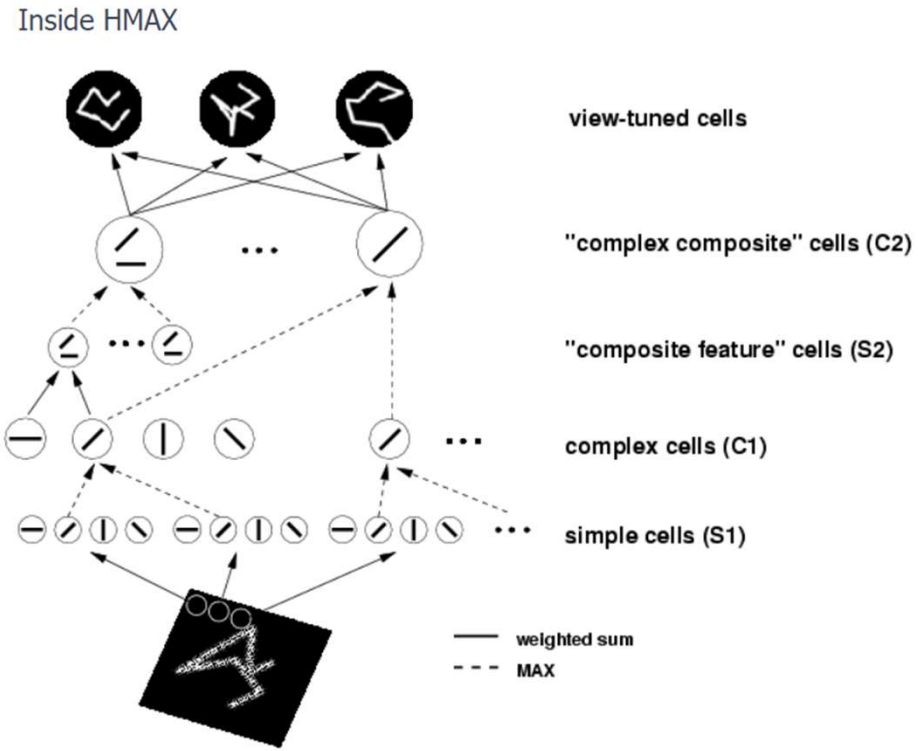
Deep NN and our visual system



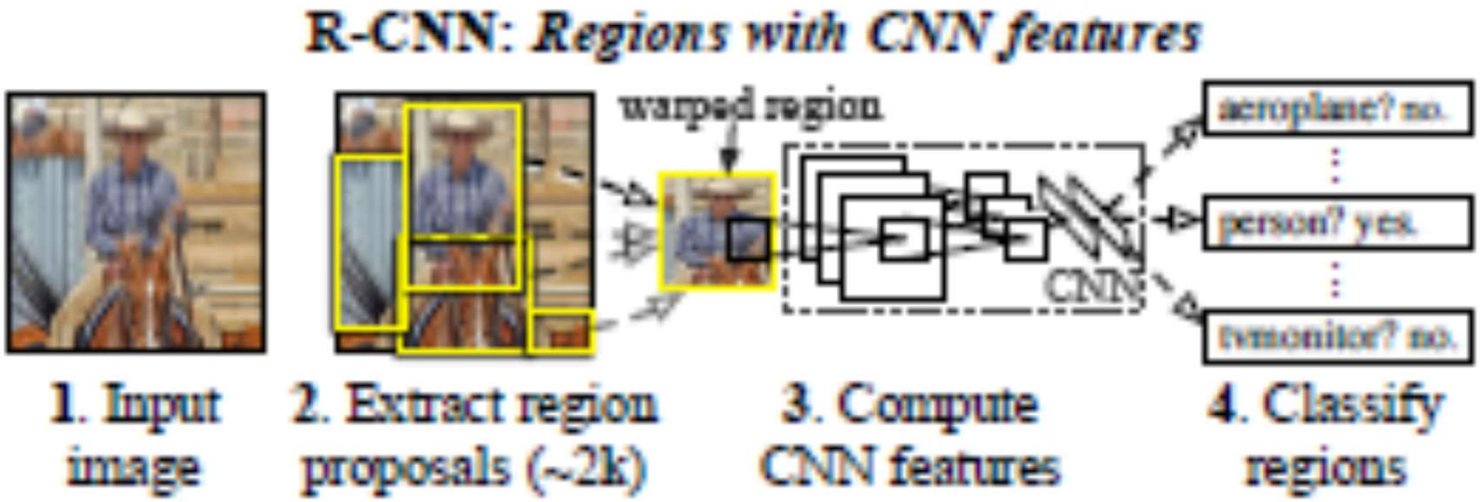
Deep NN and our visual system: Ullman et al. Atoms of Recognition

- Central problem in the study of human vision: discovering the visual features and representations used by the brain to recognize objects.
- Neural network models rival human performance on image classification.
- Representations and learning processes discovered by current models are similar to those used by the human visual system?

Deep NN models considered: HMAX



Deep NN models considered. R-CNN

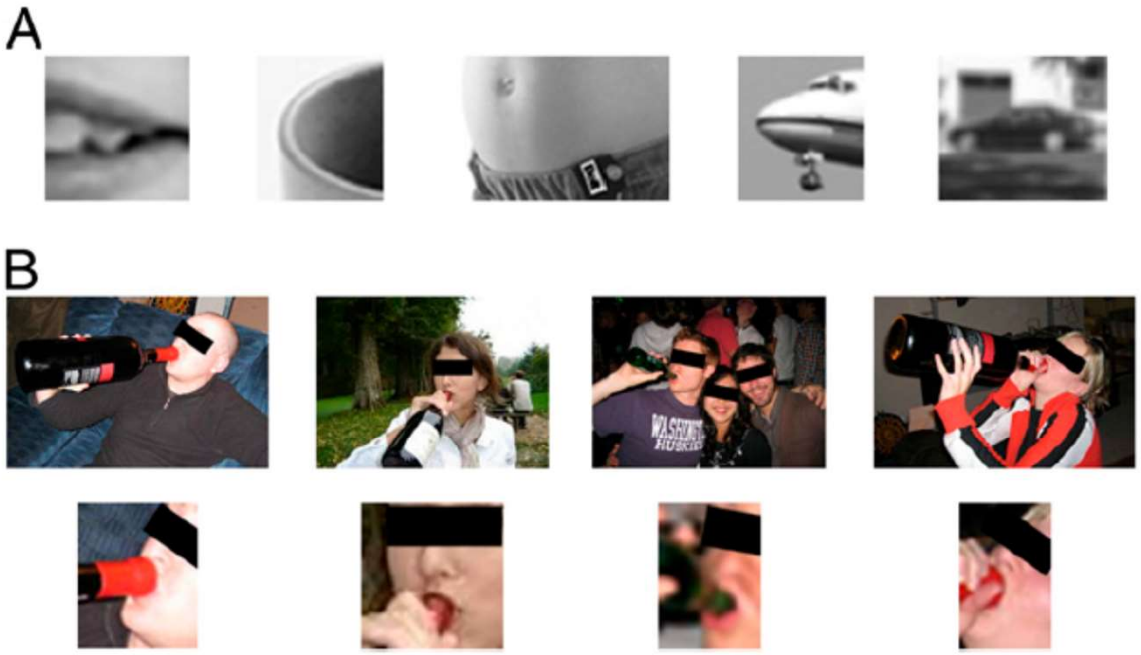


Deep NN and our visual system

A

MIRC: Minimal recognizable configurations

Deep NN and our visual system



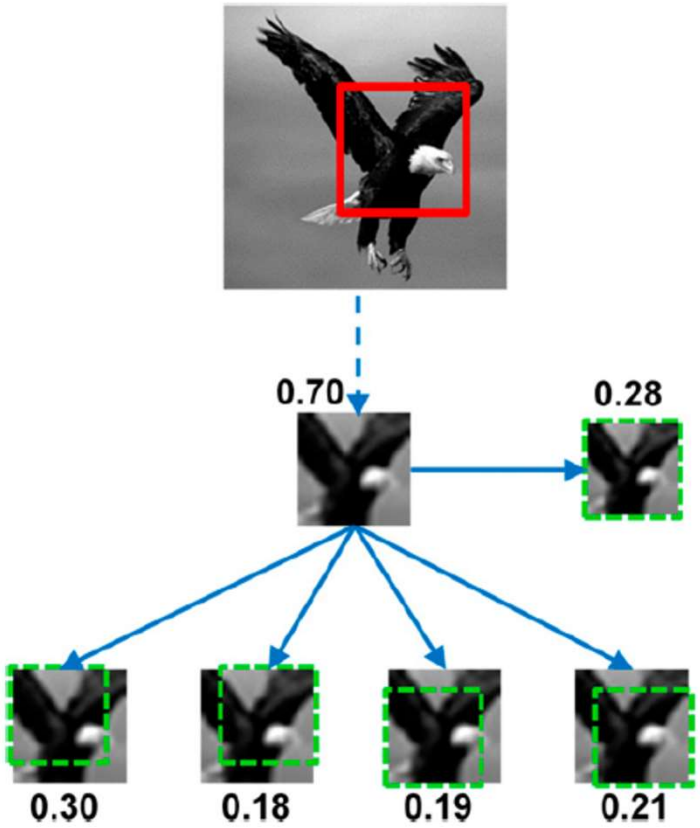
MIRC: Minimal recognizable configurations

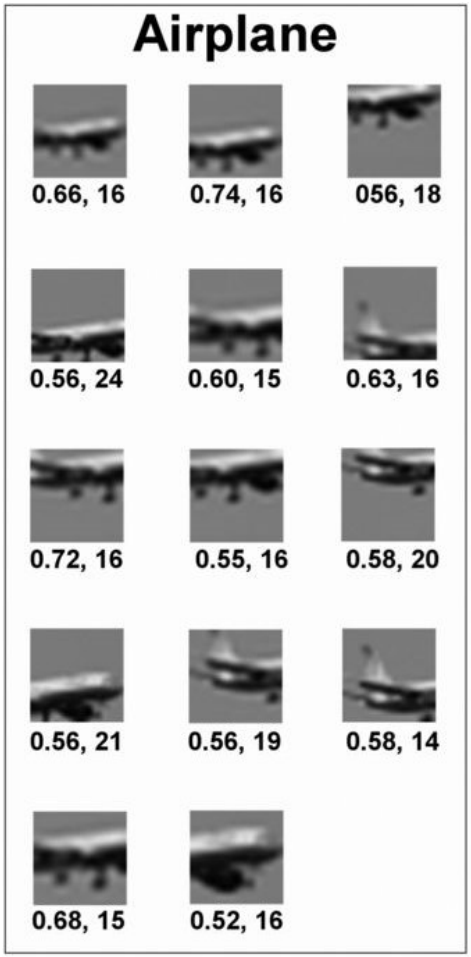
Deep NN and our visual system: Ullman et al. Atoms of Recognition

- By using minimal recognizable images: the human visual system uses features and processes that are not used by current models and that are critical for recognition.
- The role of the features is revealed uniquely at the minimal level, where the contribution of each feature is essential.
- A full understanding of the learning and use of such features will extend our understanding of visual recognition and its cortical mechanisms and will enhance the capacity of computational models to learn from visual experience and to deal with recognition and detailed image interpretation

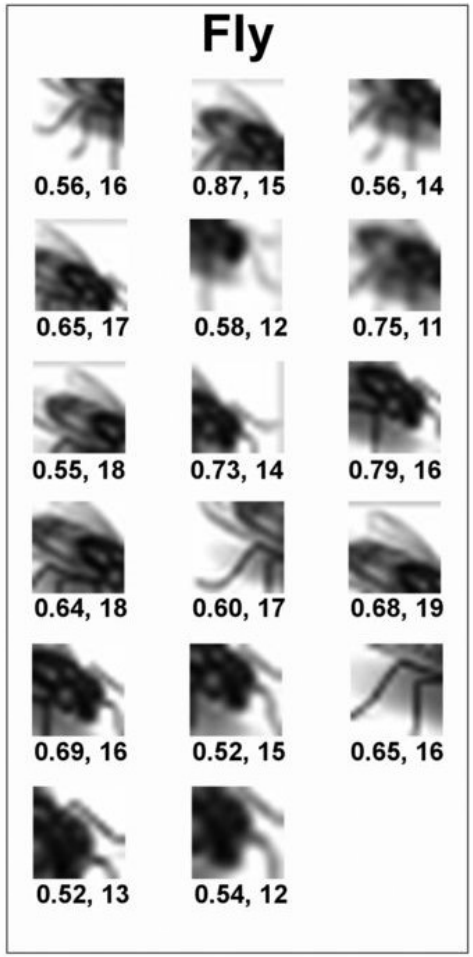
Deep NN and our visual system: MIRC and sub-MIRCS

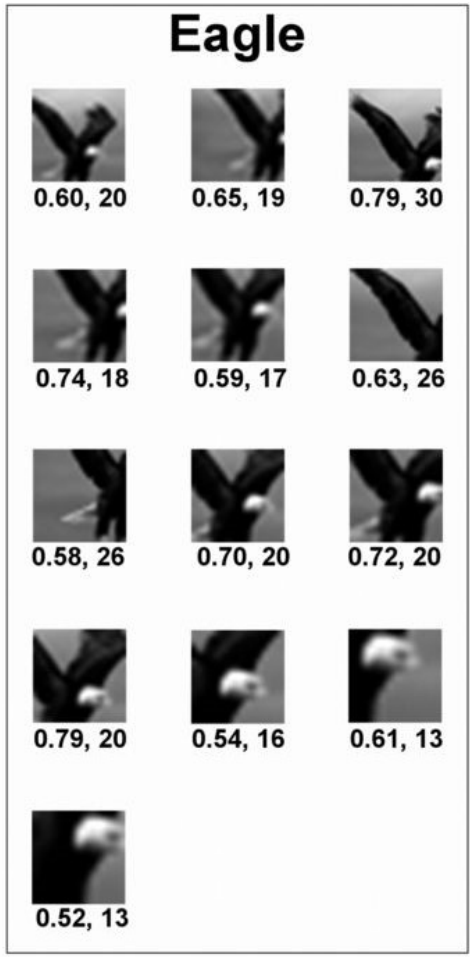
•



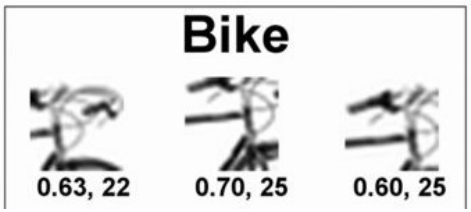

Airplane
0.66, 16
0.74, 16
056, 18
0.56, 24
0.60, 15
0.63, 16
0.72, 16
0.55, 16
0.58, 20
0.56, 21
0.56, 19
0.58, 14
0.68, 15
0.52, 16

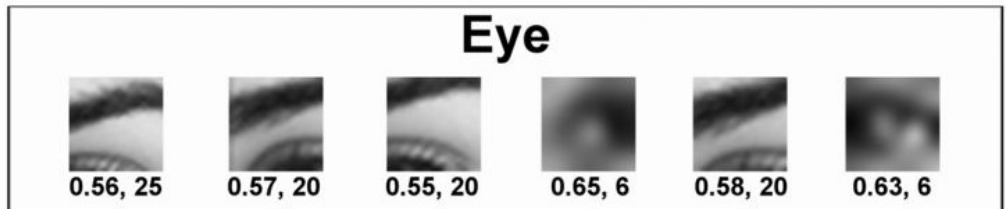

Ship
0.56, 24
0.79, 17
0.84, 18
0.66, 18
0.54, 25
0.67, 26
0.53, 20
0.56, 21
0.63, 21
0.57, 13
0.52, 12
0.52, 13
0.56, 13
0.57, 13


Fly
0.56, 16
0.87, 15
0.56, 14
0.65, 17
0.58, 12
0.75, 11
0.55, 18
0.73, 14
0.79, 16
0.64, 18
0.60, 17
0.68, 19
0.69, 16
0.52, 15
0.65, 16
0.52, 13
0.54, 12

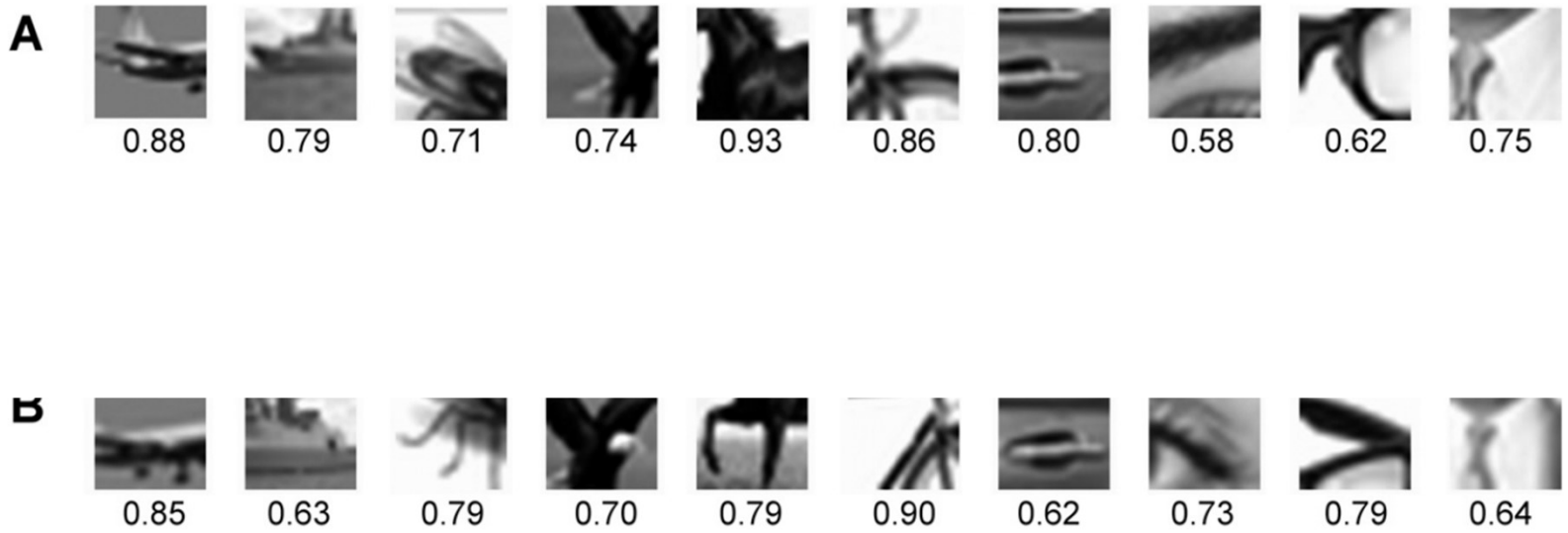

Eagle
0.60, 20
0.65, 19
0.79, 30
0.74, 18
0.59, 17
0.63, 26
0.58, 26
0.70, 20
0.72, 20
0.79, 20
0.54, 16
0.61, 13
0.52, 13


Horse
0.52, 29
0.69, 9
0.65, 15

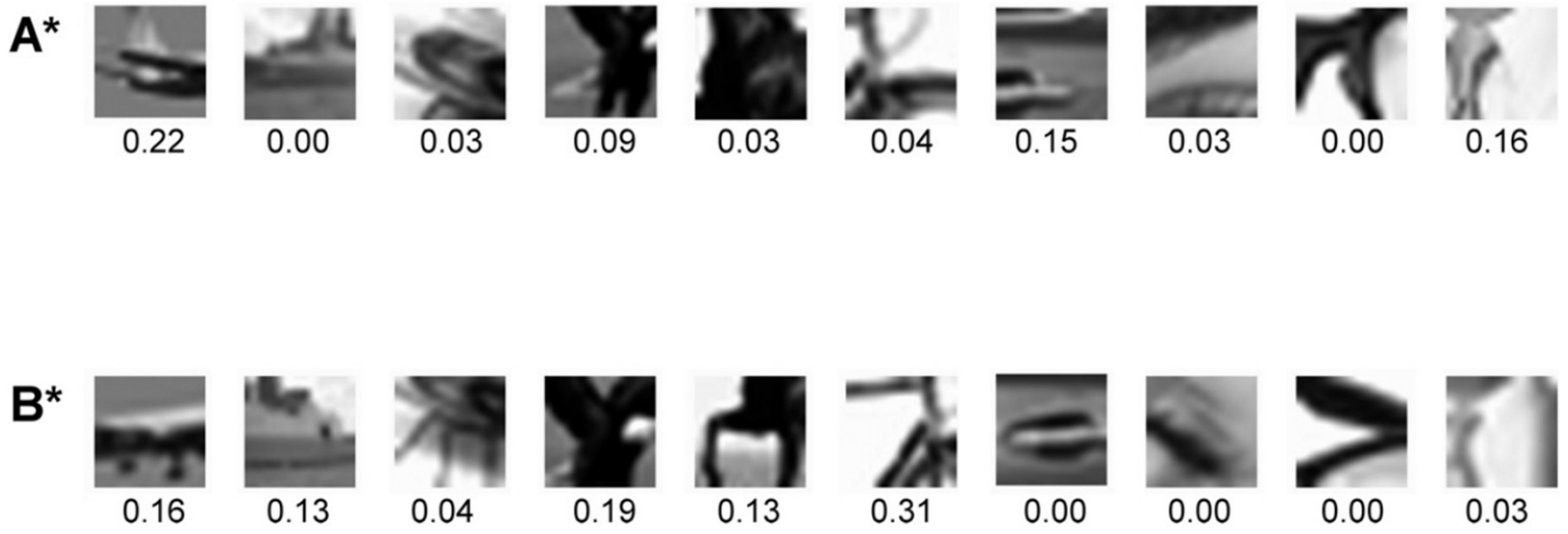

Bike
0.63, 22
0.70, 25
0.60, 25


Eye
0.56, 25
0.57, 20
0.55, 20
0.65, 6
0.58, 20
0.63, 6

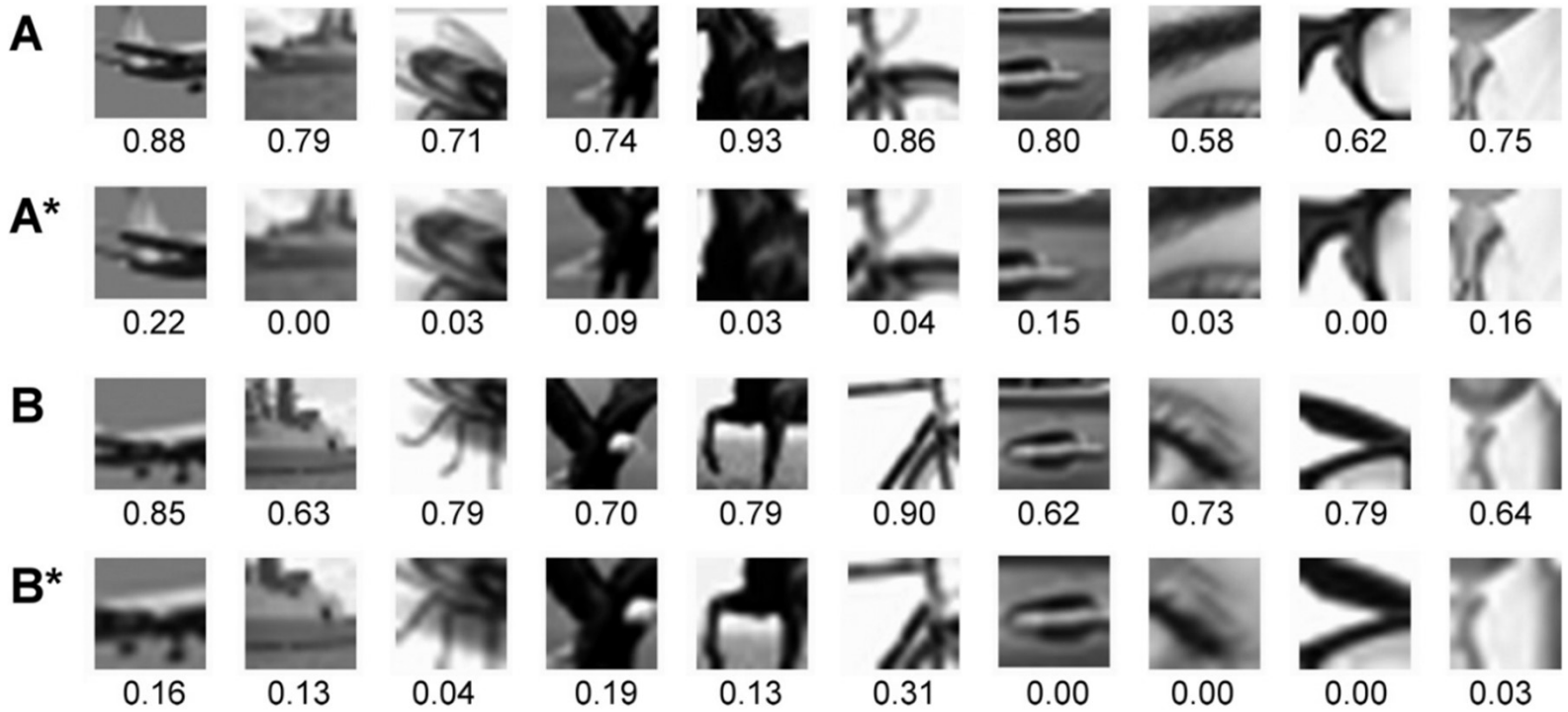
MIRC


A
0.88
0.79
0.71
0.74
0.93
0.86
0.80
0.58
0.62
0.75
B
0.85
0.63
0.79
0.70
0.79
0.90
0.62
0.73
0.79
0.64

SUBMIRC


A*
0.22
0.00
0.03
0.09
0.03
0.04
0.15
0.03
0.00
0.16
B*
0.16
0.13
0.04
0.19
0.13
0.31
0.00
0.00
0.00
0.03

MIRC (A, B) vs. SUBMIRC (A*, B*)



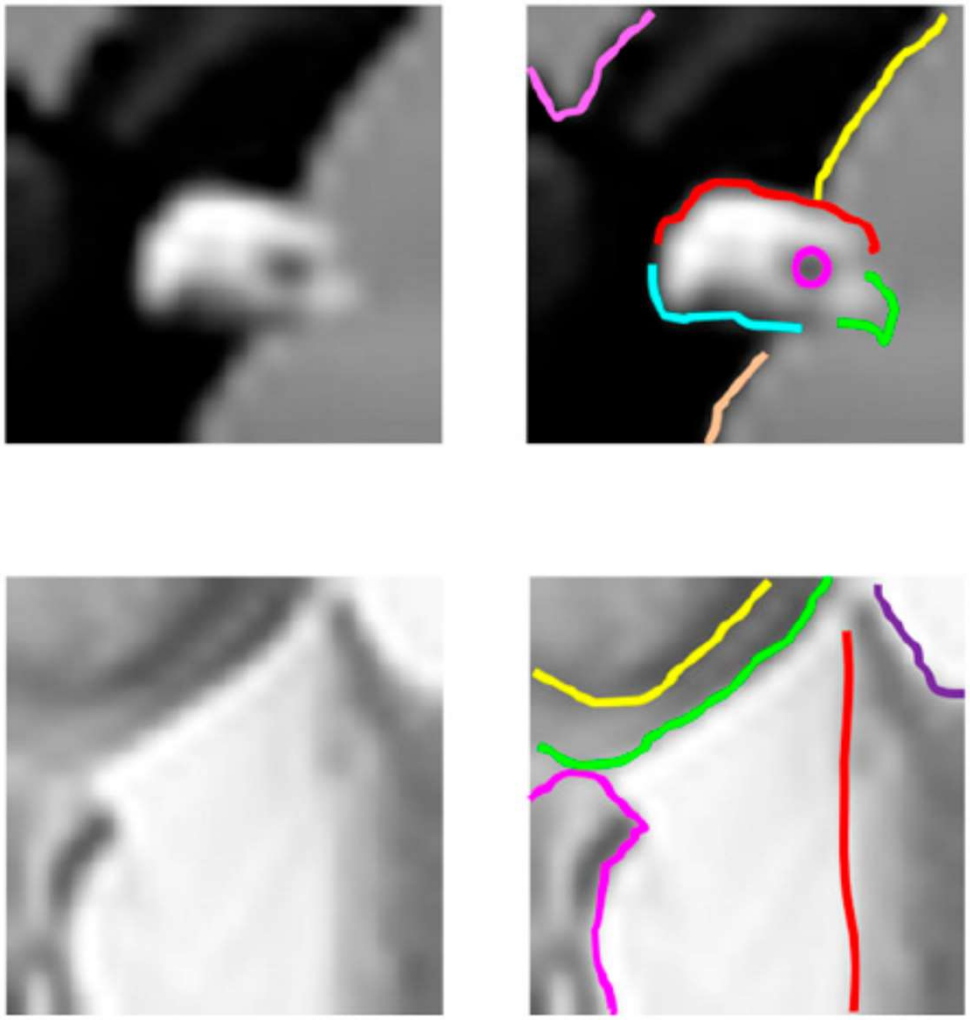
Results:

- In humans better recognition of MIRC than in network models
- In humans sharp decrease between MIRC and SUB-MIRC recognition rate; not in neural network models
- Network models:
  - They are not good at recognizing minimal images;
  - and no decrease between MIRC and SUB-MIRC.

Internal interpretation

- An additional limitation of current modeling compared with human vision is the ability to perform a detailed internal interpretation of MIRC images.
- Humans can consistently recognize multiple components internal to the MIRC .
- Such internal interpretation is beyond the capacities of current neural network models, and it can contribute to accurate recognition, because a false detection could be rejected if it does not have the expected internal interpretation.

Example



Feed forward or also top down?

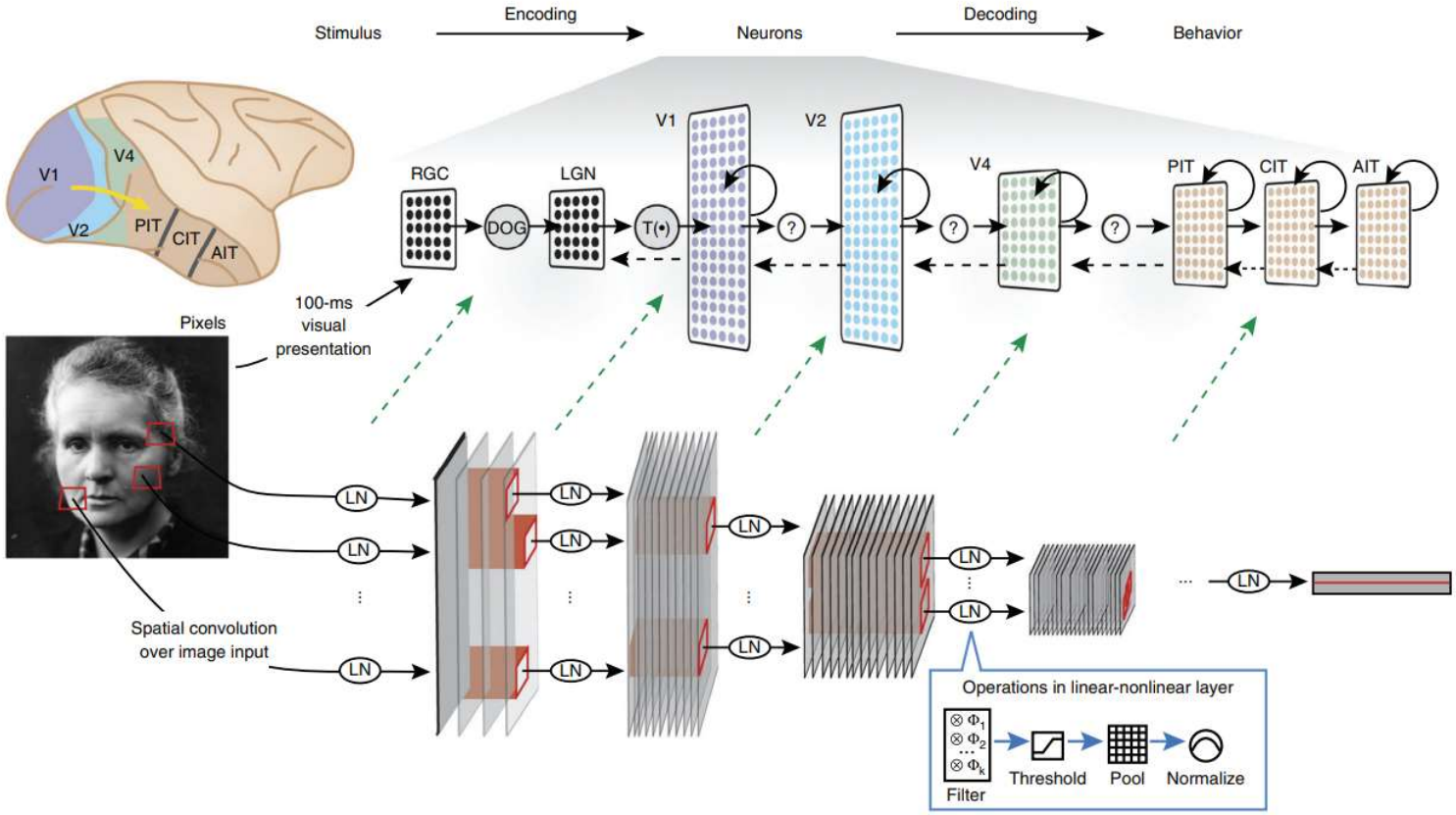
- Features are used in the visual system as a part of the cortical feed-forward process or by a top-down process, which currently is missing from the purely feedforward computational model?
- Top down processes are likely to be involved: detailed interpretation appears to require features and interrelations that are class-specific, i.e., their presence depends on a specific class and location.
- Two main stages:
  - Initial activation of class candidates, which is incomplete and with limited accuracy.
  - The activated representations then trigger the application of class specific interpretation and validation processes, which recover richer and more accurate interpretation of the visible scene

LSTM and the BRAIN or COGNITION

3- Artificial neural networks as models to understand the brain

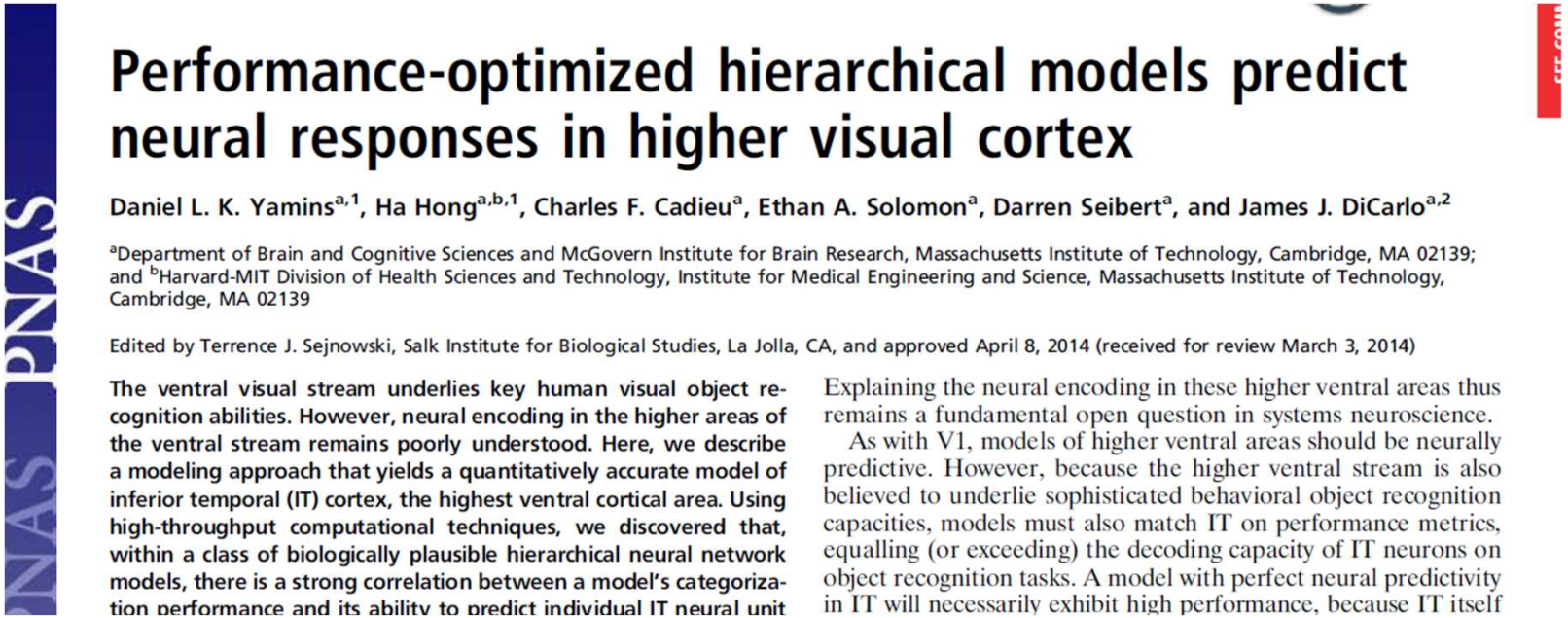
Deep NN and our visual system

- CNN inspired by processing of visual information in the brain



Deep NN and our visual system

- Several researchers to use CNN (and family) to predict and describe activity in areas of the brain



PNAS

Performance-optimized hierarchical models predict neural responses in higher visual cortex

Daniel L. K. Yamins[a,1], Ha Hong[a,b,1], Charles F. Cadieu[a], Ethan A. Solomon[a], Darren Seibert[a], and James J. DiCarlo[a,2]

[a]Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Harvard-MIT Division of Health Sciences and Technology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Terrence J. Sejnowski, Salk Institute for Biological Studies, La Jolla, CA, and approved April 8, 2014 (received for review March 3, 2014)

The ventral visual stream underlies key human visual object recognition abilities. However, neural encoding in the higher areas of the ventral stream remains poorly understood. Here, we describe a modeling approach that yields a quantitatively accurate model of inferior temporal (IT) cortex, the highest ventral cortical area. Using high-throughput computational techniques, we discovered that, within a class of biologically plausible hierarchical neural network models, there is a strong correlation between a model's categorization performance and its ability to predict individual IT neural unit

Explaining the neural encoding in these higher ventral areas thus remains a fundamental open question in systems neuroscience.

As with V1, models of higher ventral areas should be neurally predictive. However, because the higher ventral stream is also believed to underlie sophisticated behavioral object recognition capacities, models must also match IT on performance metrics, equalling (or exceeding) the decoding capacity of IT neurons on object recognition tasks. A model with perfect neural predictivity in IT will necessarily exhibit high performance, because IT itself

Cross Fertilization between Neuroscience discoveries and Neural Networks

- Learning how the brain processes visual information can also improve CNNs
- Example:

Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations

Joel Dapello*,[1,2,3], Tiago Marques*,[1,2,4]
Martin Schrimpf[1,2,4], Franziska Geiger[2,5,6,7], David D. Cox[8,3], James J. DiCarlo[1,2,4]

- CNN models with a neural hidden layer that better matches primate primary visual cortex (V1) are also more robust to adversarial attack