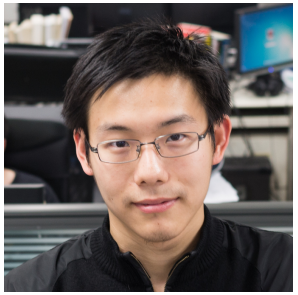
Comparing Intuitive Physics Engine and ConvNets on Physical Scene Understanding


Renqiao Zhang*


Jiajun Wu*


Chengkai Zhang


Bill Freeman


Josh Tenenbaum

CogSci 2016

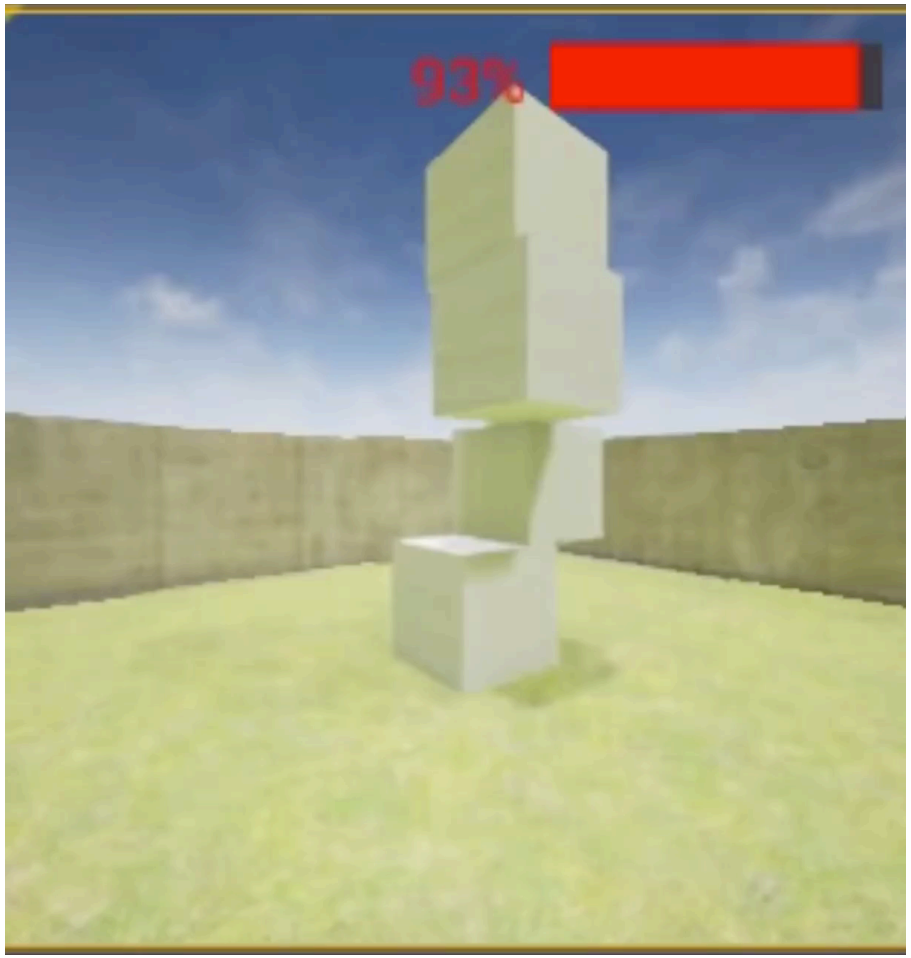
(* equal contributions)

brain+cognitive sciences

CSAIL

CENTER FOR Brains Minds+ Machines

Understanding Physical Events



Video from Facebook AI blog post

Motivation

Observations:

- Humans can understand physical events very quickly
- They can also transfer the knowledge to other tasks

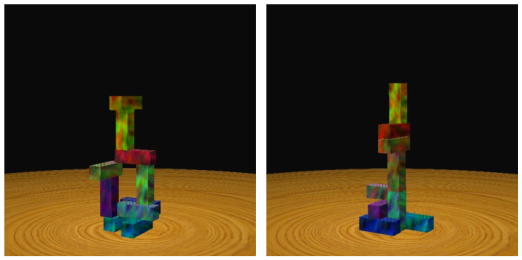
Goal: a proper computational model for this process

Models:

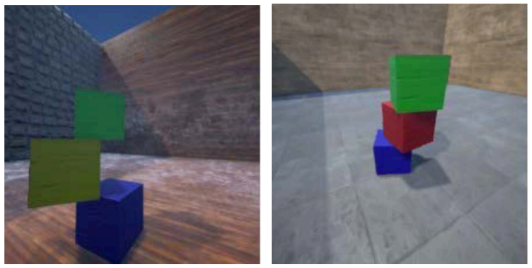
- a intuitive physics engine (IPE)
- a convolutional neural net (CNN)

Stimuli

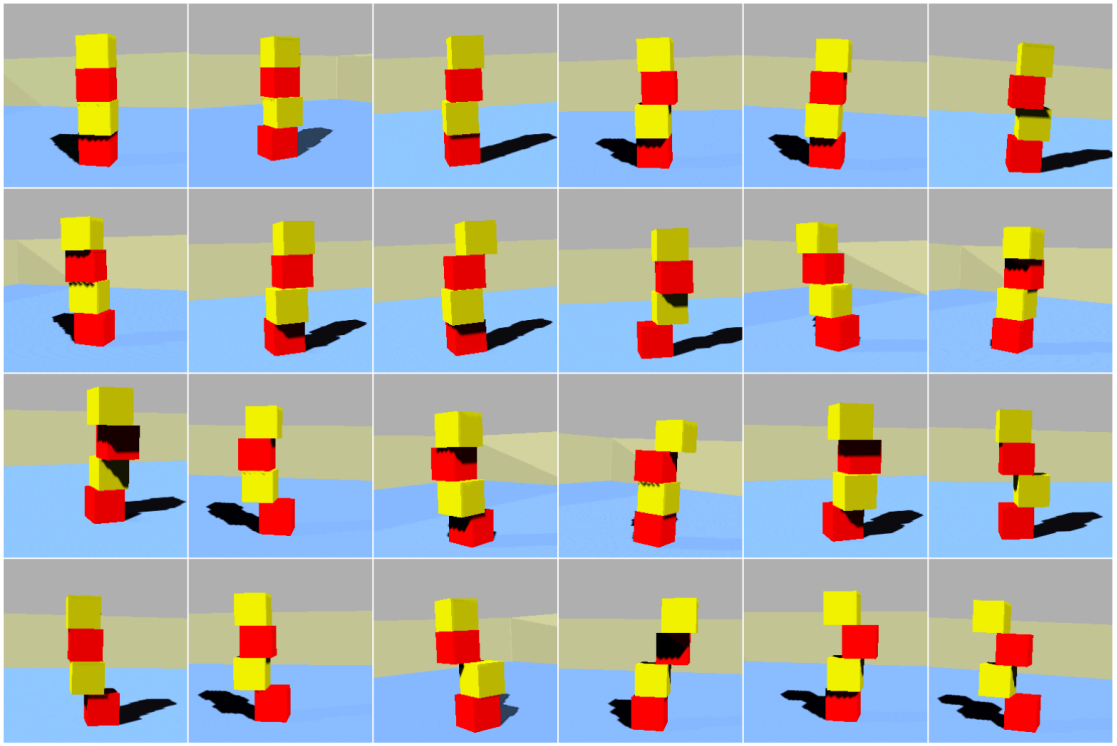
The classic tower block scenario



Stimuli in
Battaglia et al. (2013)



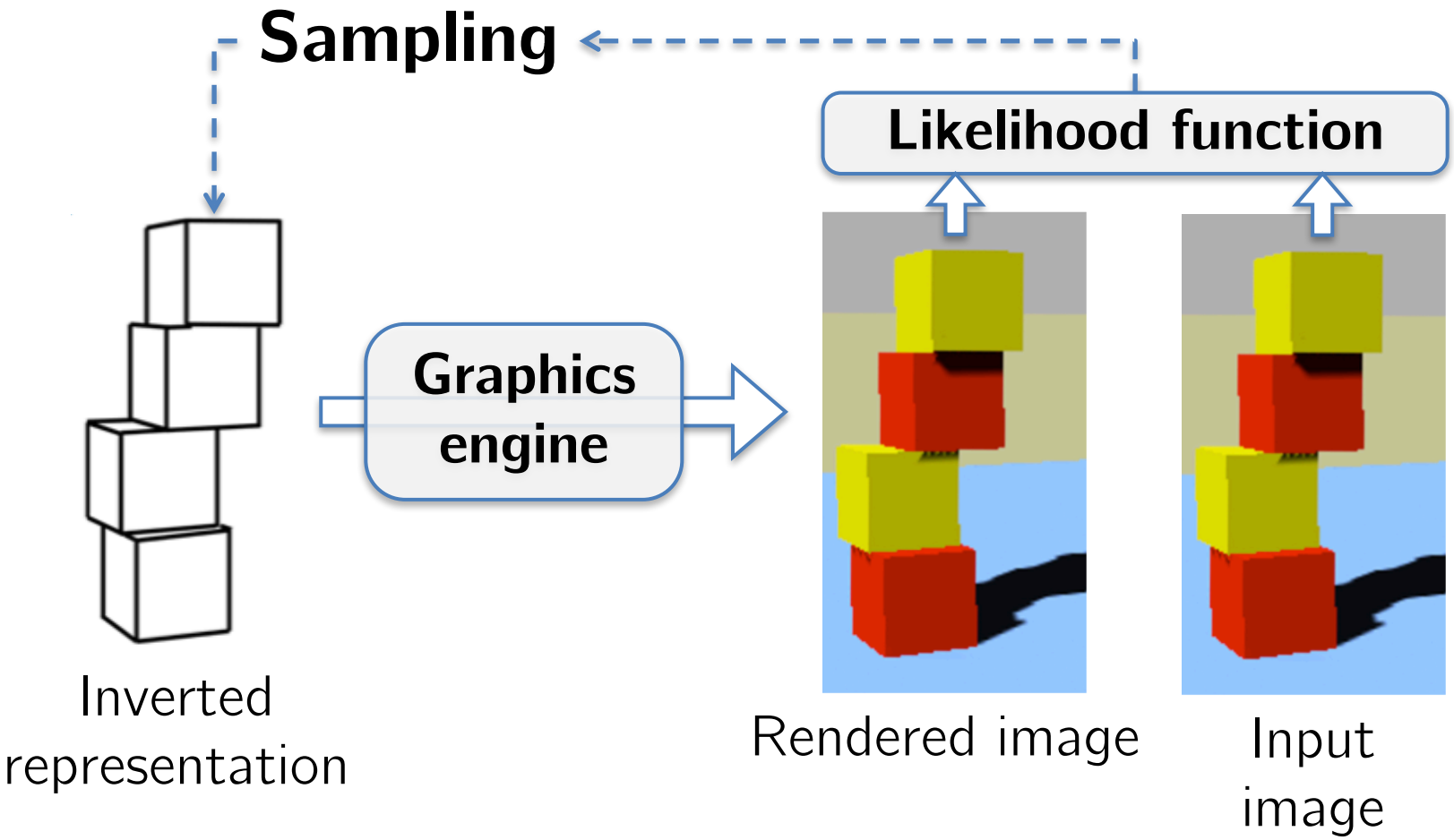
Stimuli in
Lerer et al. (2016)



Our Stimuli

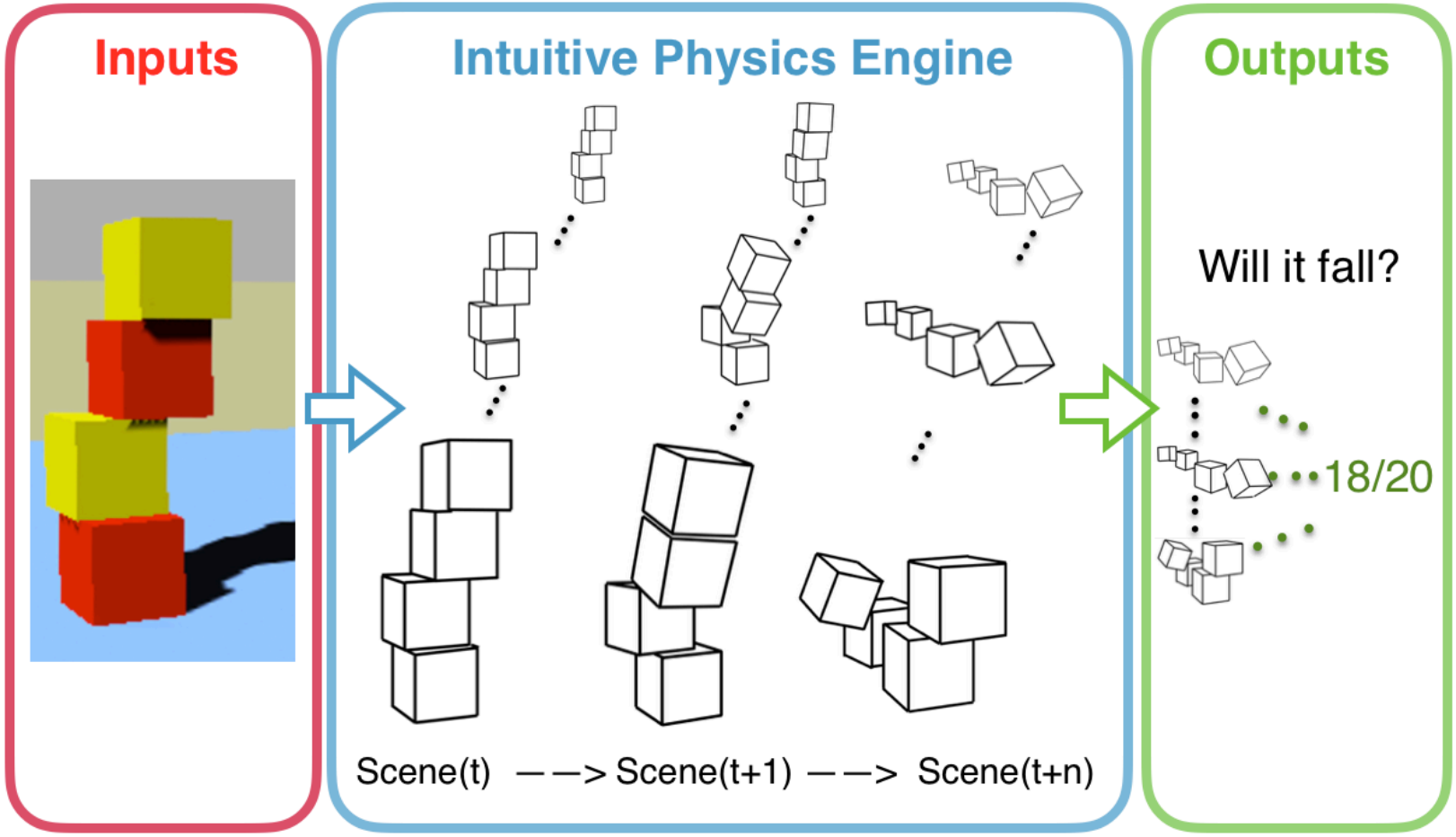
- Battaglia, Hamrick, Tenenbaum. PNAS, 2013.
- Lerer, Gross, Fergus. ICML, 2016.

IPE: Inverse Graphics



Battaglia, Hamrick, Tenenbaum. PNAS, 2013

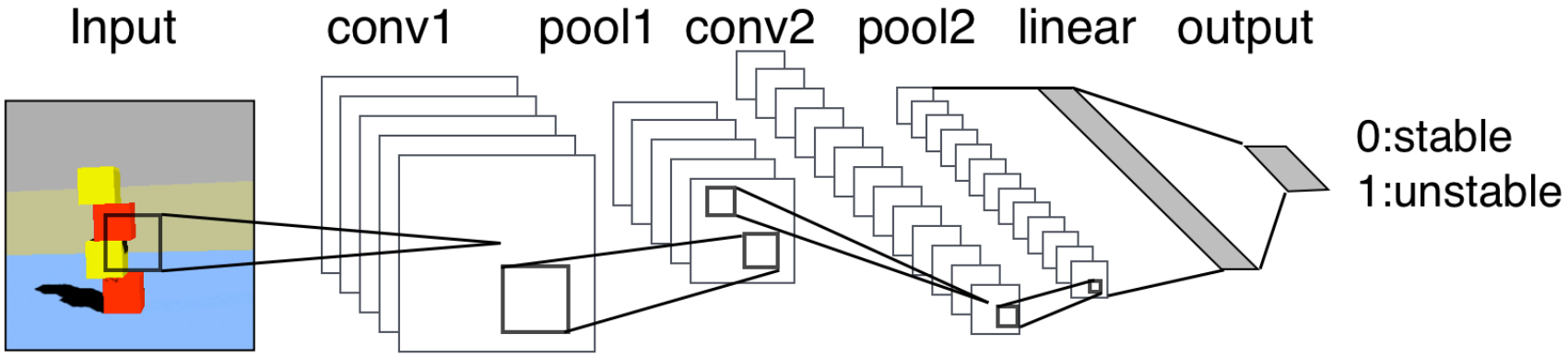
IPE: Simulation



Battaglia, Hamrick, Tenenbaum. PNAS, 2013

Convolutional Networks

- LeNet: smaller, 2 convs and 2 linear layers
- AlexNet: widely used, 5 convs and 3 linear layers
- AlexNet pretrained on ImageNet



Observation: Asymmetry

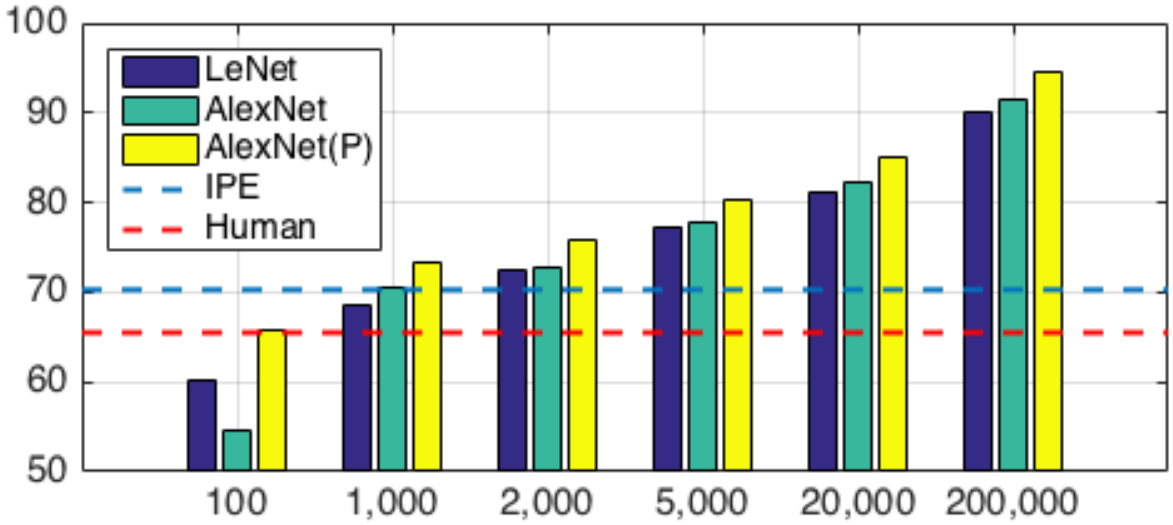
Method	Stable	Unstable	All
Human	38.0	92.9	65.5
IPE	40.7	99.0	70.3
LeNet (200K)	91.3	89.0	90.1
AlexNet (200K)	91.5	92.3	91.9
AlexNet (Pretrained, 200K)	94.5	94.7	94.6

Accuracies (%) of humans, IPE, LeNet, and AlexNet

Observations:

- Networks have better overall performance.
- There is an asymmetry in human accuracies.
- The IPE can model the asymmetry, but not the CNNs.

Size of Training Set



Accuracies (%) of networks with different sizes of training data

Observations:

- Networks achieves human-like accuracies with 1K images.
- AlexNet performs better, but only with enough data.
- Pretrained networks suffer less from the lack of data.

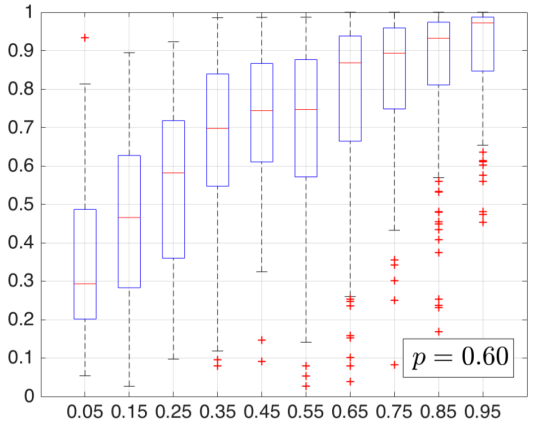
Accuracy with Limited Training Data

Method	Stable	Unstable	All
Human	38.0	92.9	65.5
IPE	40.7	99.0	70.3
LeNet (200K)	91.3	89.0	90.1
AlexNet (200K)	91.5	92.3	91.9
AlexNet (Pretrained, 200K)	94.5	94.7	94.6
LeNet (1,000)	68.0	69.3	68.7
AlexNet (1,000)	71.8	70.1	70.9
AlexNet (Pretrained, 1,000)	72.5	74.2	73.4

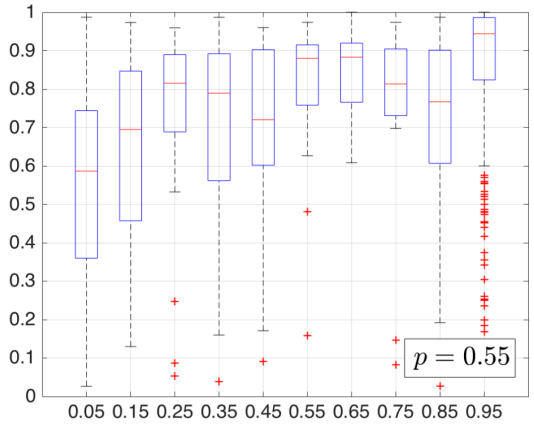
Accuracies (%) of humans, IPE, LeNet, and AlexNet

Still, only the IPE can model the asymmetry, not the CNNs.

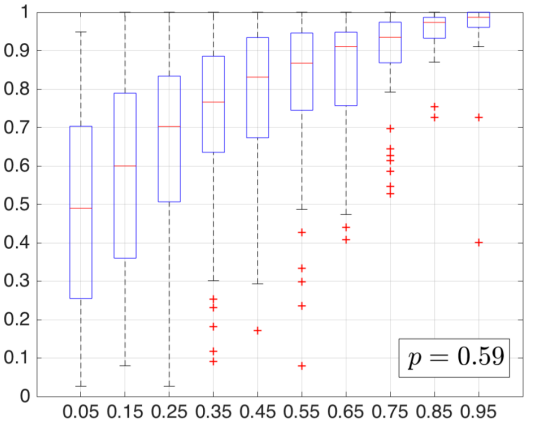
Correlation with Human Responses



(a) Human (Y-axis) vs IPE (X-axis)



(b) Human (Y) vs LeNet 200K (X)



(c) Human (Y) vs LeNet $1,000$ (X)

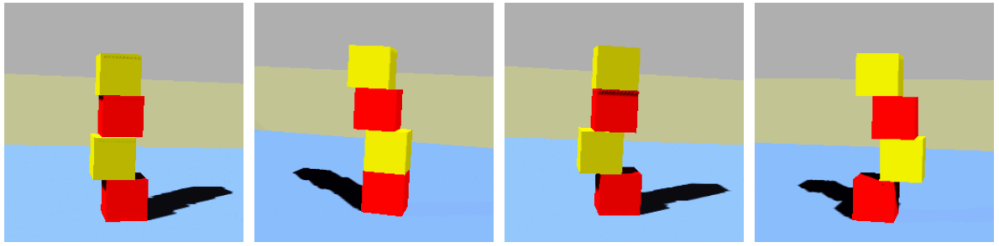
Correlations between human and model responses (p for Pearson's coefficients)

Both models correlate with human responses reasonably well.

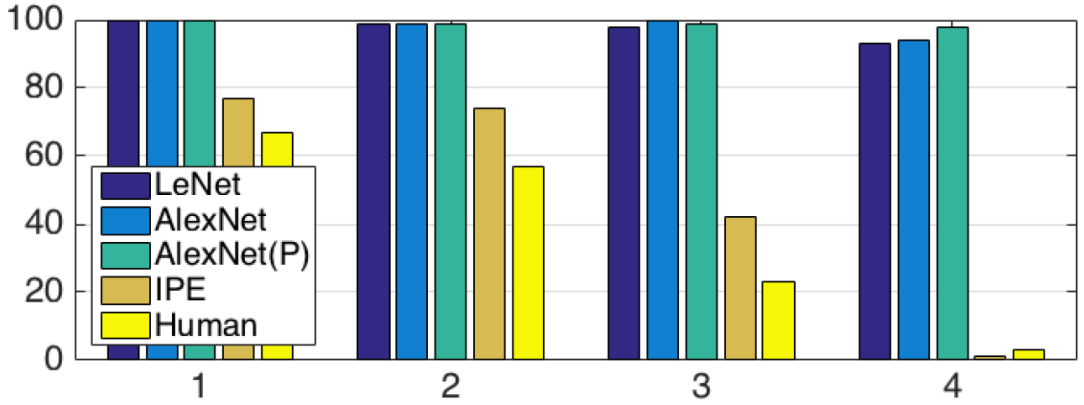
Visual Instability

Stable images may have different visual instability.
How would models perform in these cases?

Images



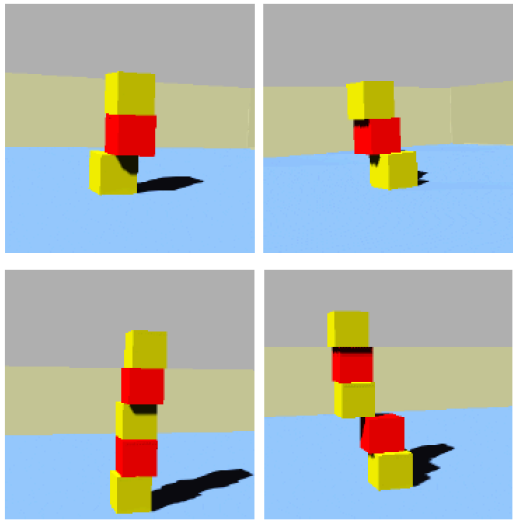
Results



Observation: Both human and the IPE perform worse as visual instability increases, but not CNNs.

Generalization

Can models generalize to images with a different number of blocks?



Data

Model	Training	Test Set			
		3	4	5	Avg
LeNet (200K)	4	50.5	88.5	64.0	67.7
AlexNet (200K)	4	52.5	89.5	65.5	69.2
AlexNet (P, 200K)	4	51.0	95.0	78.5	74.8
LeNet (1,000)	4	57.0	64.0	66.0	62.3
AlexNet (1,000)	4	54.0	62.0	64.5	60.2
AlexNet (P, 1,000)	4	55.0	71.0	72.0	66.0
IPE (0.1, 10x)	N/A	72.0	64.0	56.0	64.0
Human	N/A	76.5	68.5	59.0	68.0

Results

Observation: Both human and the IPE perform worse as the number of blocks increases, but not CNNs.