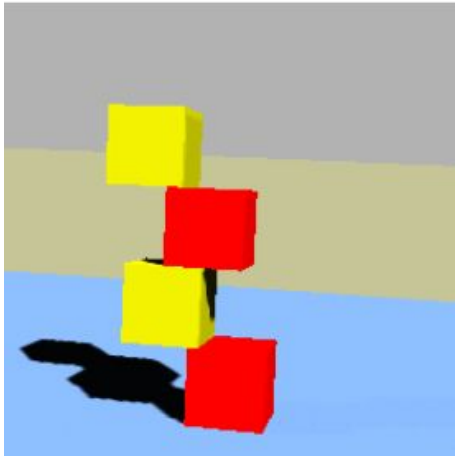
Probabilistic Simulation Models vs. CNNs in Explaining Human Scene Perception (Zhang et al., 2016)

Wednesday Bushong & Sudhanshu Srivastava

Scene Perception & Intuitive Physics



Humans have a good intuitive understanding of physics

Not surprising, since we basically need to in order to survive!

Of course, humans aren't perfect:
https://www.youtube.com/watch?v=RPBIgjaHpvA

Goal of Paper

Given what we know about humans' understanding of physics, what is a good approach to modeling the problem?

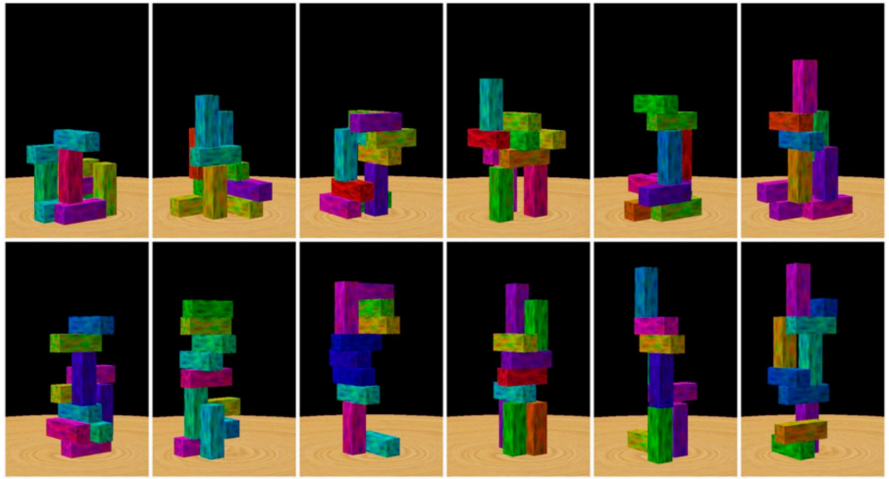
Modeling Approaches

- Intuitive Physics Engine (IPE) models: physical reasoning-based approach
  - Assumes people have rich physics models
  - Takes an initial state and simulates forward based on physics laws
- Neural networks: kill it with data!

Overview of IPE

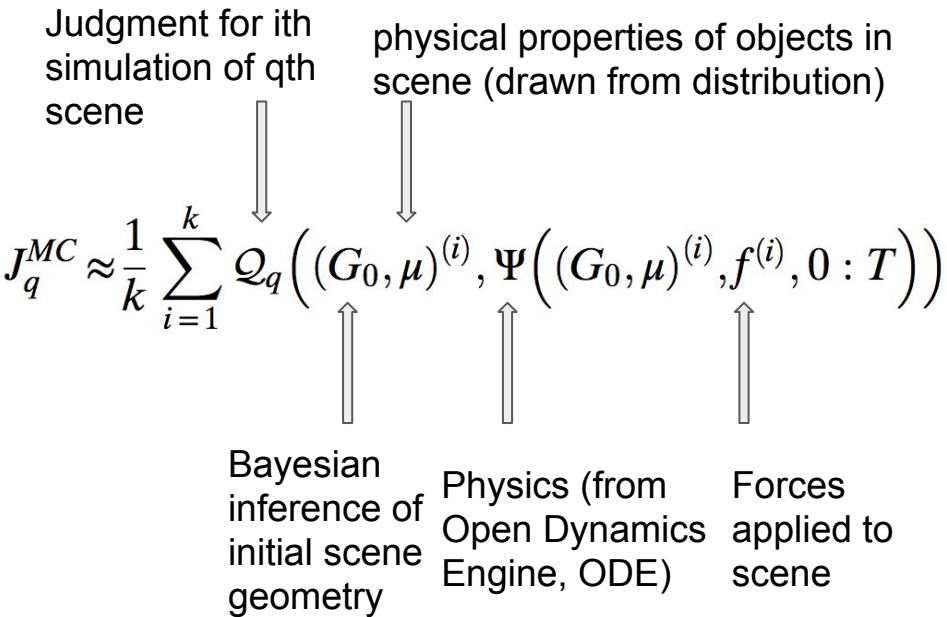
- From general physics principles, simulate forward from some starting position

- Early assessment of IPE models: Battaglia et al. (2013)
- “Does it fall?”



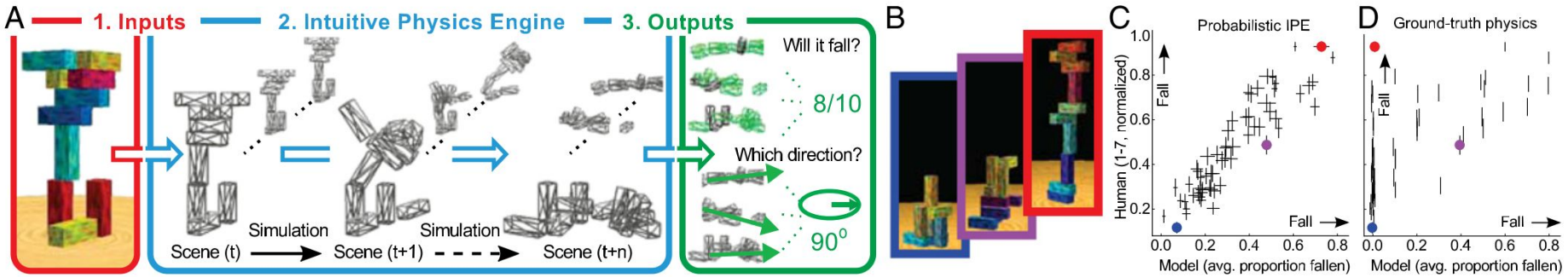
Battaglia et al. (2013) IPE model specifics

- Simulate physics forward from initial state
- Model has uncertainty about the current state, physical attributes of objects, & latent force inputs; vals for these are drawn from a distribution for each of k simulations per trial



Battaglia et al. (2013) model architecture & results

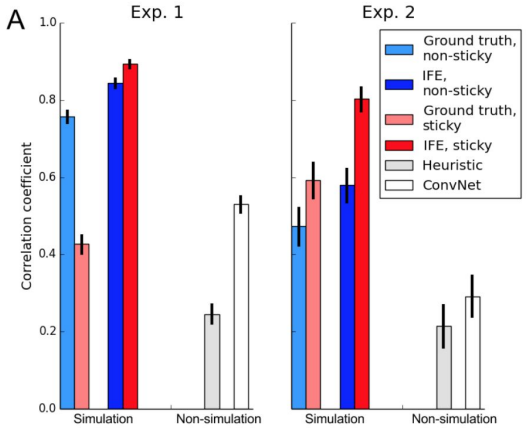
- Predict: 1) does it fall? 2) which direction does it fall?

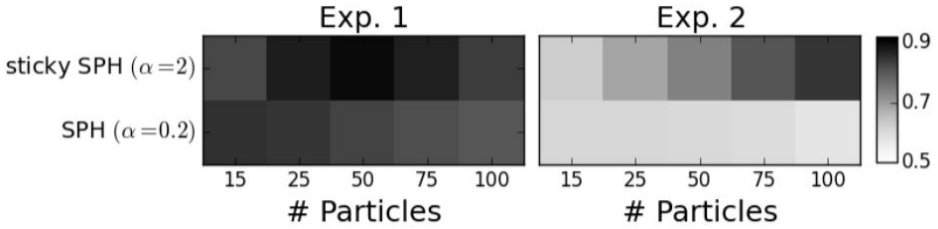


- Pretty good performance compared to humans

Other example domains of IPEs

- Our very own Chris Bates has done similar work in liquid dynamics (Bates et al., 2015)
- Demo of task: http://web.mit.edu/cjbates/www/liquidfun/liquidfun/Box2D/lfjs/cogfluid_interactive.html



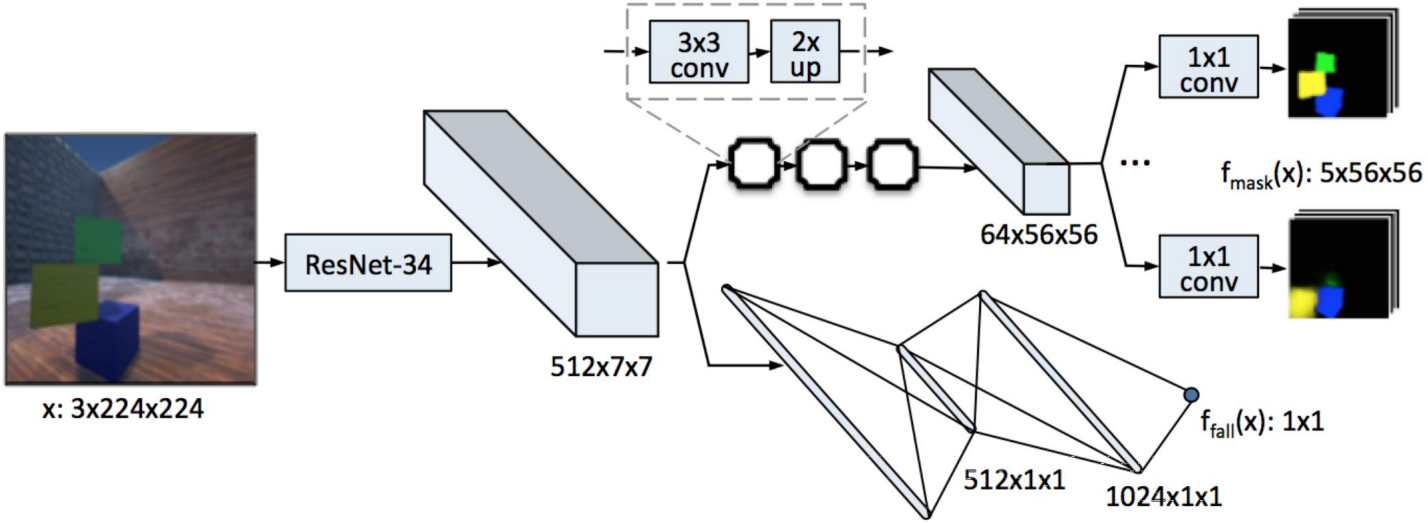


Why might IPEs be a good model for humans?

- Generative physics knowledge allows for very good generalization to new scenarios
- Previous work shows that IPEs are subject to similar physics “illusions” as humans (Battaglia et al., 2013)
- Generative/Bayesian models in general capture human cognition in many domains (e.g., language: Frank & Goodman, 2012; multisensory integration: Kording et al., 2007; and many others)

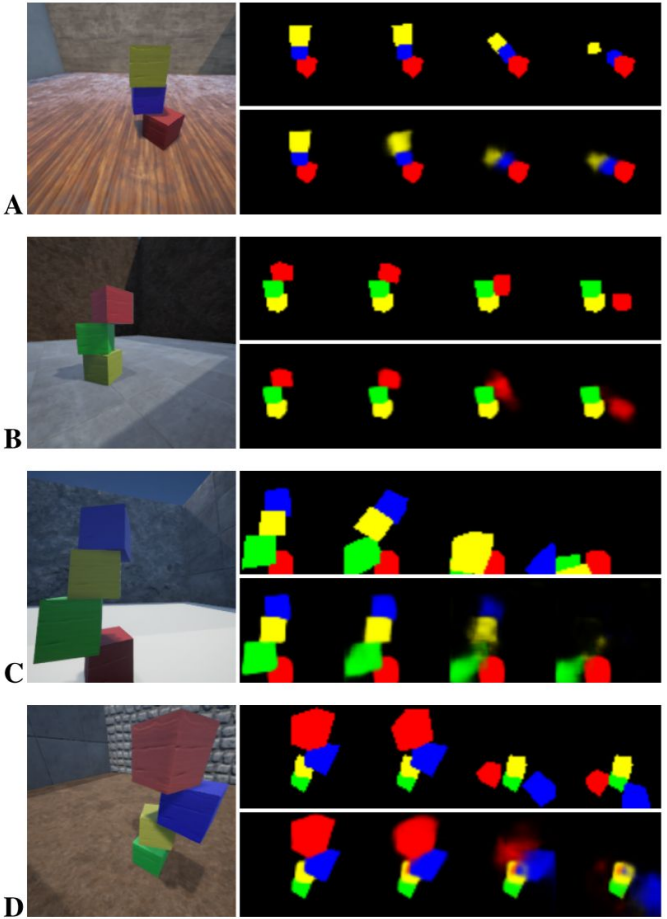
Previous CNN Work: Lerer et al. (2016)

- Predict: 1) does it fall? 2) where do the blocks go?



Lerer et al. (2016): Generalization

- Test model on #s of blocks it hadn't been exposed to before
- Generalizes decently to stacks of 3 & 4 blocks when those examples are held out
- Crucially, doesn't generalize as well as humans



Discussion:

What are the benefits and drawbacks of IPEs vs. CNNs?

Are these different for applications vs. cognitive models?

Current Study

- Assess whether IPEs or CNNs are a better model of human behavior

- 4 behavioral experiments using Battaglia-esque block-falling task
- Compare performance with an IPE and several CNN frameworks

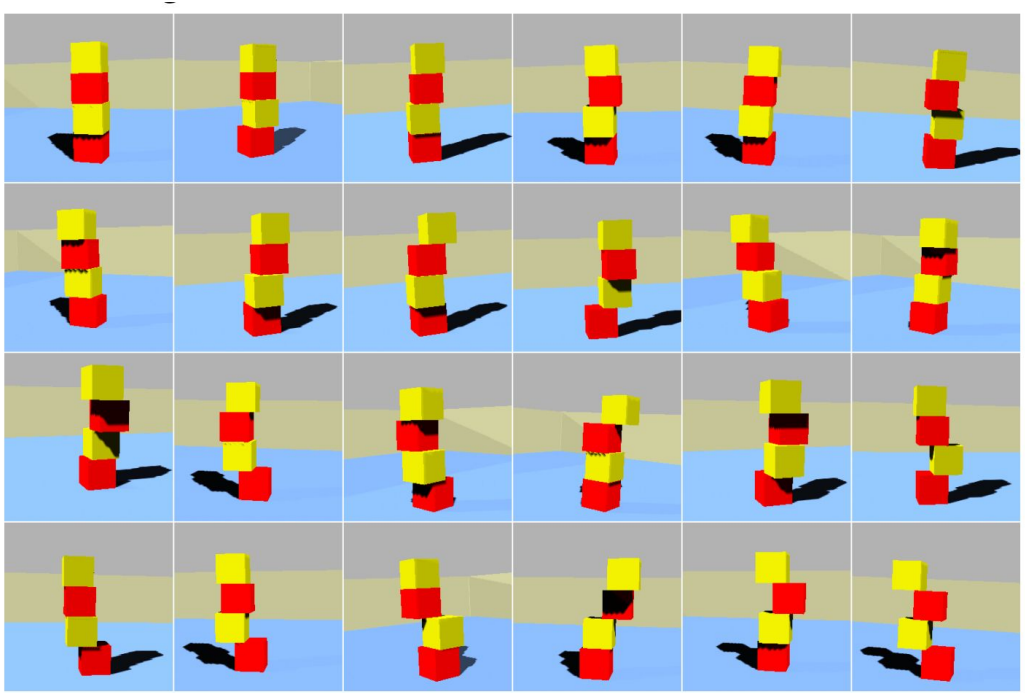
Behavioral Experiment

80 Mechanical Turk subjects

14 trials each (4 “easy”, 10 randomly chosen from stimulus set)

Stacks of 4 blocks

Task: will it fall?



Computational Models

- IPE: Same model as Battaglia et al. (2013)
  - k (# of simulations) = 20
  - other parameters determined by fit to human data in Experiment 1
- CNNs:
  - LeNet, fine-tuned for task
  - AlexNet: one version pre-trained on ImageNet & fine-tuned for task, one with no pretraining

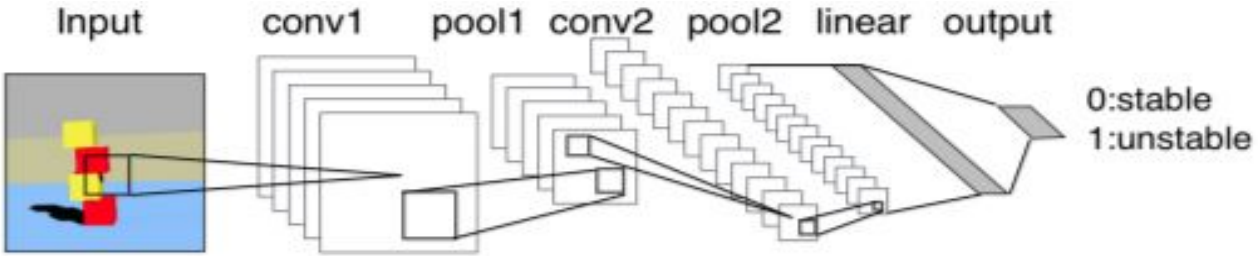


Figure 3: The structure of LeNet

Analyses

Experiment 1: IPEs vs. humans

Experiment 2: How much training data is needed for CNNs to look like IPEs & humans?

Experiment 3: How do the models do on stacks of blocks that look unstable to humans but are actually stable?

Experiment 4: Knowledge transfer (generalization) to untrained stimuli (stacks of 3 or 5 blocks)

Results: IPEs vs Humans

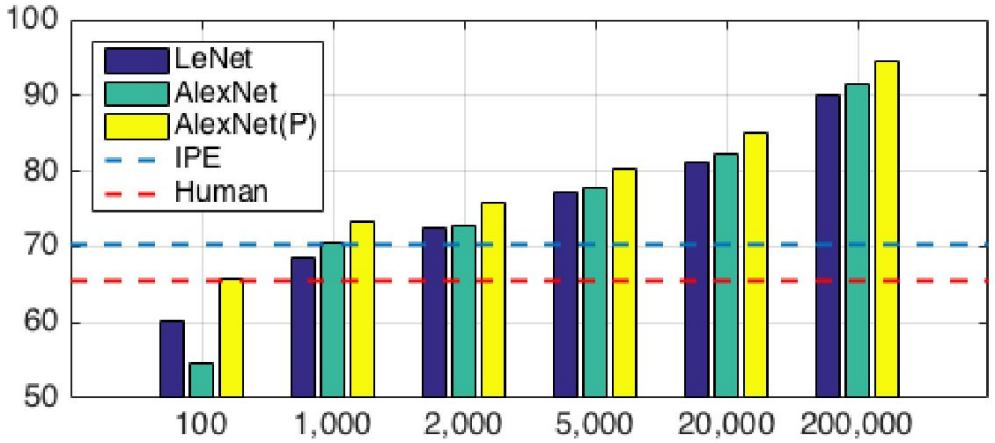
- IPE model had highest accuracy with no noise added (surprise surprise)
- Highest accuracy + correlation with humans was found for sigma = 0.1, phi = 40
- These vals are used for remainder of experiments

		ϕ				
		0	35	40	45	50
σ	0	94.2	87.2	79.5	71.3	63.8
	0.05	91.3	83.4	76.1	69.1	61.8
	0.1	83.2	75.7	70.3	62.6	56.4
	0.15	72.2	66.8	59.4	54.2	51.2
	0.2	58.5	53.8	52.1	51.0	50.9
Corr		≥ 0.45	≥ 0.54	≥ 0.56	≥ 0.58	≥ 0.60

Results: Limited Data

CNNs needed ~1000-2000 training examples to match IPE & human performance

Not-pretrained AlexNet suffers most from limited data

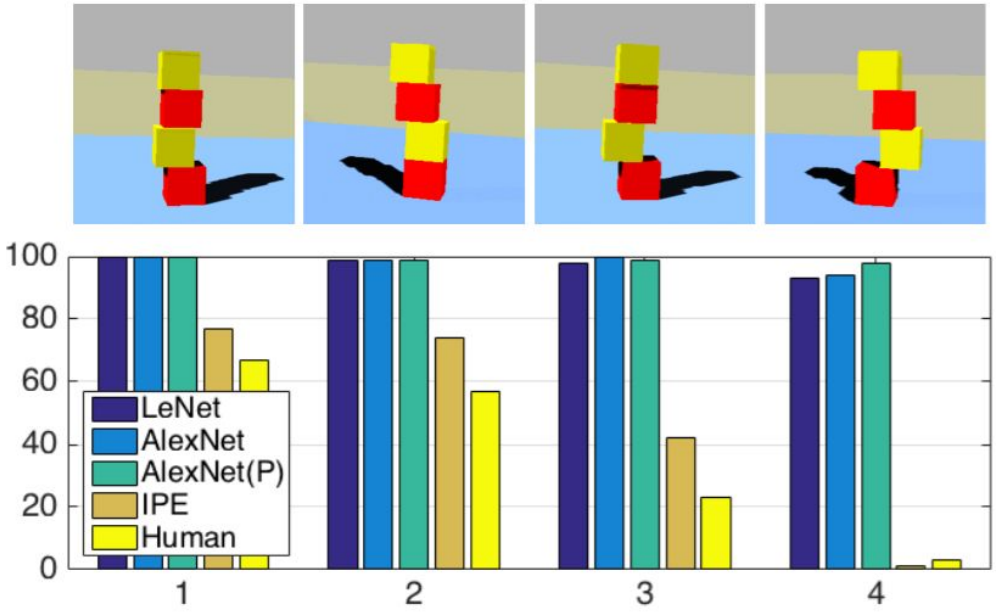


Method	Stable	Unstable	All
Human	38.0	92.9	65.5
IPE	40.7	99.0	70.3
LeNet (200K)	91.3	89.0	90.1
AlexNet (200K)	91.5	92.3	91.9
AlexNet (Pretrained, 200K)	94.5	94.7	94.6
LeNet (1,000)	68.0	69.3	68.7
AlexNet (1,000)	71.8	70.1	70.9
AlexNet (Pretrained, 1,000)	72.5	74.2	73.4

Results: Deceptively Stable Block Piles

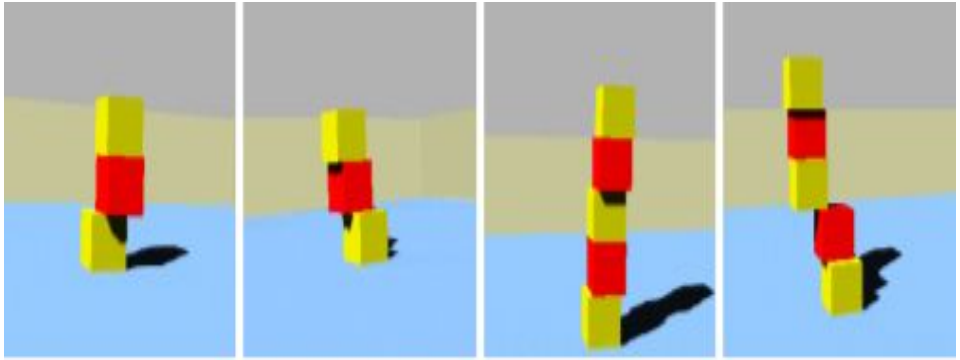
Human & IPE performance decreases as block piles *appear* to be more unstable

CNN performance hardly changes at all



Results: Knowledge Transfer

- Humans subjects and IPE were consistent when evaluated on 3 or 5 blocks, but CNN's performance decreased drastically.
- Change from 4 to 5 is easier for CNN as compared to 4 to 3.



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

Table 3: Results on the task of transfer learning

Analysing CNN’s performance

To determine stable/unstable-

1)Figure out there are blocks(a CNN can do this),

2)Find their positions(a CNN can do this),

3)Check if the centre of mass of each block and each group of blocks has something below it to support(a CNN can do this with enough data. [“A powerful enough Neural Net can approximate any function.”- Henry])

^boils down to finding the mean of the x positions of the blocks above a point and checking if it lies between x_min and x_max of the block just below.

Analysing CNN’s performance(contd.)

With the last slide in mind, it makes sense that:

1)Lack of training data causes poor performance.

2)A CNN does better on Edge cases- it knows the x values and has probably figured out to find the mean.

3)It does drastically worse on knowledge transfer. It is probably still trying to calculate the average of 4 blocks while there are 3 or 5. In case of 3, there is no fourth block, so it tosses a coin and answers. [taking a wild guess] In case of 5, it can find out the average, which coincidentally works sometimes. E.g., if the 4 block average predicts unstable, the 5 block setup will definitely be unstable.

Conclusions

- IPEs and CNNs achieve human-like performance in general
- However, CNNs are not affected by boundary cases (blocks that look unstable but are stable), while IPEs & humans are
- CNNs also do not generalize well to different numbers of blocks in a stack

- Take-home point seems to be: if you want something very accurate for the stimulus type it was trained for, go with a CNN; if you want great generalization to new situations but human-like failure in edge cases, go with an IPE

Limitations of Study

- Are IPEs and CNNs even really that comparable?
- IPEs as explanations of human behavior assume that humans already have physics knowledge, whereas CNNs are building from the ground up
- That is, IPEs *completely* ignore the learning problem

- Discuss!!!

How can we improve these models?

- For practical applications, we want CNNs to have their awesome accuracy on edge cases, but also to display generalization like IPEs
- As cognitive models, we want some kind of learning story from IPEs: where does physics knowledge come from? How can we explain developmental data?

To Discuss:

Humans do bad on edge cases which are slightly stable, but look unstable. This rarely happens for cases that are slightly unstable, but look stable. What could the reason be?

2 cents(very likely overkill): Classifying unsafe as safe would mean our ancestors would hang on branches that are likely to break, but classifying safe as unsafe would mean we are just overcautious. Similar to the case of Pareidolia, where we see faces in clouds.

Do our brains have an IPE?

Fischer et al. (2016) found that “physical scene understanding engages a systematic set of brain regions replicated across three studies”

The 3 tasks: 1)Predict Falling blocks,2)Predict trajectory,3)Movie watching

The regions involved have been previously attributed to motor action planning, tool use, and general problem-solving.

Tool use and intuitive physical inference have previously been shown to be correlated via clinical trials. Apraxia patients find it difficult to use familiar tools, and also do poorly on inferring the use of a novel tool based on its physical structure.