To Fall or not to Fall

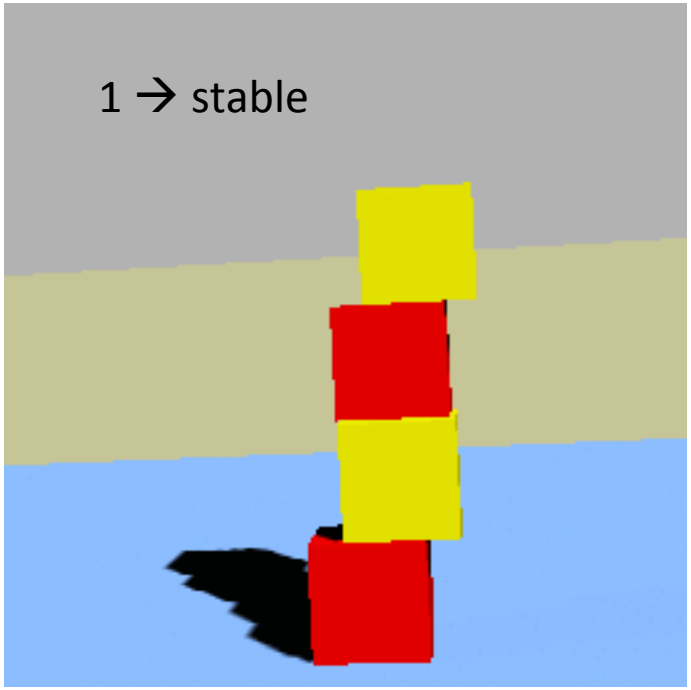
Introduction

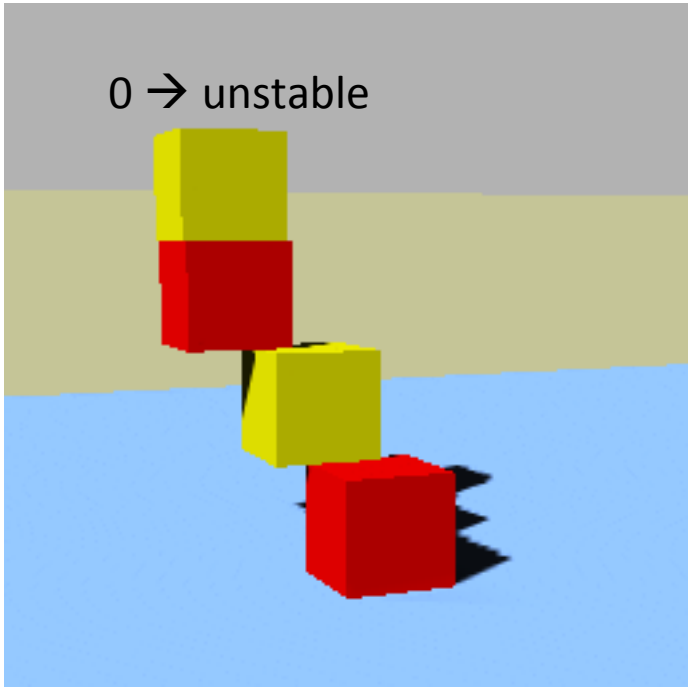
Humans demonstrate remarkable abilities to predict physical events in complex scenes. memory-based models, which make judgments based on analogies to stored experiences of previously encountered scenes and physical outcomes. Versions of the latter have recently been instantiated in convolutional neural network (CNN) architectures.

Introduction

- In this work a CNN model will be present
- concretely instantiated in algorithms that can run on raw image inputs and produce as outputs physical judgments such as whether a stack of blocks will fall.
- This approach →achieve superhuman accuracy levels
- Quantitatively predict human judgments to a similar degree.
- Still has generalization problem to novel situations in ways that people do

Dataset

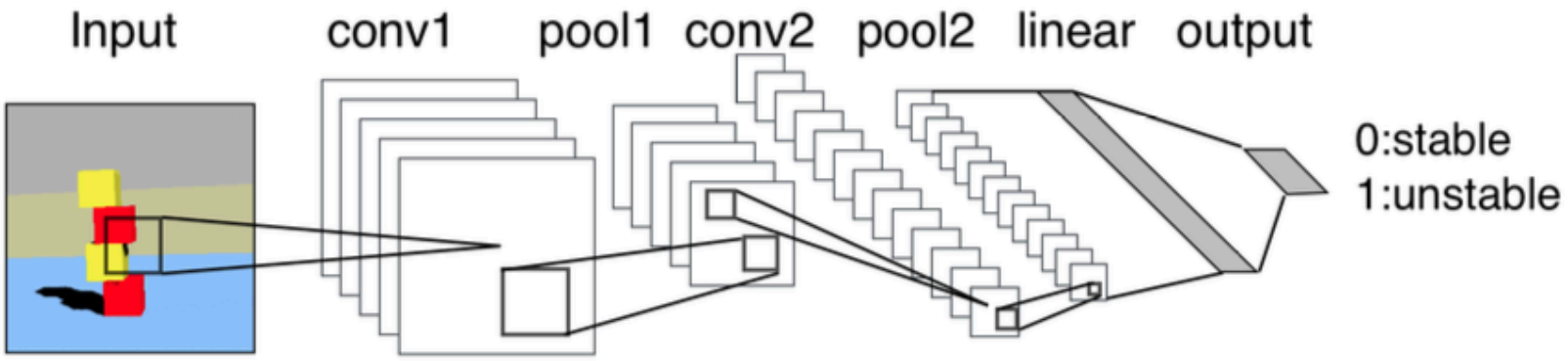




Dataset

- Actual Size of dataset = 200K
- Required size for train = 100, 1000
- For each scale, we sample five training sets independently, train one network on each set, and compute the average of their performance
- Cross Validation: 10 fold crossvalidation. 90%-10%
- Whether blocks are stable, *i.e.*, groundtruth labels, can be derived from the coordinates of blocks.
- A block will fall if and only if the center of mass of all blocks above it, including itself, does not fall on top of the block under it.
- For rendering, we generate images of resolution 256* 256

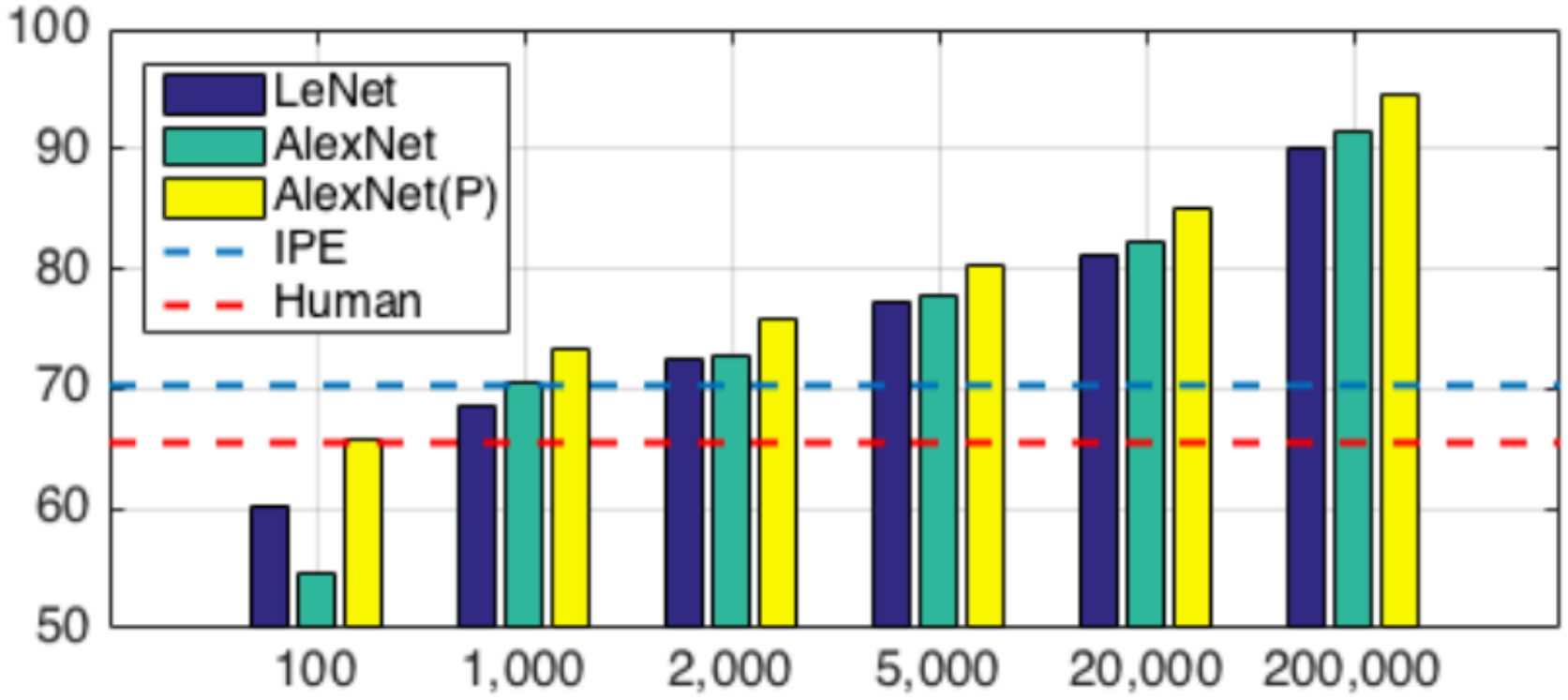
CNN setups


Input
conv1
pool1
conv2
pool2
linear
output
0:stable
1:unstable

Setups

- Convolutional Neural Network(Deep Learner)
  - two convolutional layers, each followed by a pooling layer and an activation layer . There are then two fully connected linear layers at the end.
  - learning rate to 0.01
- Random Forest:
  - Number of trees = 50
  - Depth of trees = 2
- SVM

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



Current Predicted Results

	CNN	RF	SVM
100	%60.3	%51.1	%58.0
1000	%61.01	%53	%60.8