Week 10: Black boxes and interpretability

Deep dream warm-up

https://www.youtube.com/watch?v=DgPaCWJL7XI&t=6s

https://www.youtube.com/watch?v=dJ1VorN9Cl0

Reading reactions

Bloop-Gloop game

Bloop-Gloop game

1. **Rules:**
    a. Three shapes, sequence length 3
    b. 27 total guesses, once you've exhausted those the rules are mathematically equivalent
2. **Defense:** Write a rule on paper. Fold. Set aside.
    a. Rule for *Gloop*. *Bloop* is everything else.
3. **Offense:** Goal: figure out the rule in as few examples as possible
4. **Scoring:**
    a. Defense: Number of points = Number of examples the offense took before they guessed it
    b. Offense: Number of points = Total - (number of examples)
    c. Every time the offense "guesses" the total number of examples you are allowed is subtracted by 1
5. **Full game:** Both sides (b/c defense is easier)

Modifications

- Add colors?
- Sequence length 5? 10?
- Others?

Discussion questions

- Is Bloop-Gloop easy or hard?
- How hard would this be this to do with a neural network with 1 million parameters and 1 million inputs?
- How do we audit black box models? Who should? Why?
- Would you trust a neural network to fly an airplane or drive a car?

Neural network demo

https://cs.stanford.edu/people/karpathy/convnetjs/demo/cifar10.html