Logical Neural Networks

Opening The Black Box

COMP 489 Project

Daniel Braithwaite

Supervisor: Marcus Frean

Introduction + Motivation

Diffcult to intepret Artificial Neural Networks using standard activations, e.g., Sigmoid, TanH.

Why Intepretable Systems?

- Saftey Critical Systems
- Ensuring systems make Ethical decisions
- European Union General Data Protection Regulation

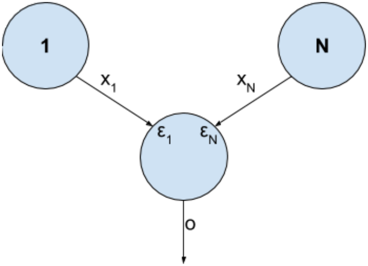
Problem Statement

Want ANNs which not only achieve high accuracy but have logic that can be defended.

Idea

- Some problems appear to have a logic decomposition
- Logical functions are easy for humans to intepret
- **Goal:** Learn these logical decompositions using backpropergation
- **Problem:** Standard Boolean Logic Gates are not continous.

Noisy Neurons



- They represent a continous paramaterisation of descrete logic gates.
- x_i is probability the i'th input is on.
- ϵ_i is the probability that input i is irrelevent. The ϵ's are the learned weights
- There exsists Noisy-AND and Noisy-OR Neurons.

Approach: Logical Neural Networks

Logical Neural Networks have layers consisting of Noisy Neurons. Can be trained with backpropagation.

Problem: Weight Initlization

- Even small networks wouldent train.
- Derived a distrubution from which to sample weights.
- Now large networks can be trained, inluding deep Logical Networks. Up-to 10 layers deep!

Experemental Approach

- Want to evaluate accuracy and peformance of Logical Neural Networks
- Will use MNIST problem.
- **Peformance:** Networks trained from 30 different initial conditions, peformance compared using confidence intervals from evaluation of network on testing set.
- **Intepretability:** Diffcult to establish. Results are obtained by visually comparing intepretations of the weights from different networks.

Experemental Results: Performance

- Logical Neural Networks have stistically equivelent peformance to Multi-Layer Perceptron Networks.

Experemental Results: Intepretability

- Logical Neural Networks are potentially more intepretable that Multi-Layer Perceptron Networks.
- Intepretability of Logical Neural Networks depends on activations used.

Experemental Results: Intepretability - No Hidden

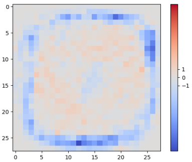


Figure: Features for a perceptron network

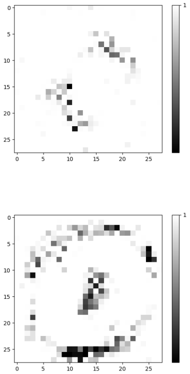


Figure: Features for a logical neural network using an AND activation

Experemental Results: Intepretability - Hidden Layers

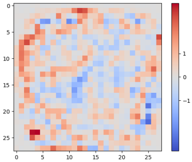


Figure: Features that positively contribute to the classification as a 1.

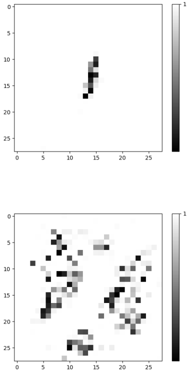


Figure: Features contributing to classification of a 1 in an AND $\rightarrow$ OR Model

Conclusion

Did we succeed in solving the problem? Well... Yes and No

- Logical Neural Networks are a promsing alternative to Multi-Layer Perceptron Networks.
- Intepretability on MNIST was "better". But again, diffcult to establish.
- Was found that intepreting Logical Neural Networks became diffcult with multiple layers.

Questions