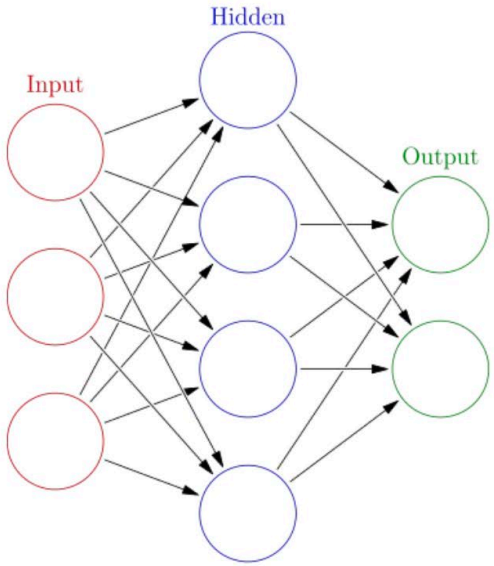
Device Placement over LLM

Project Idea by Andy Zhou

Paper Recap: Device placement with RL

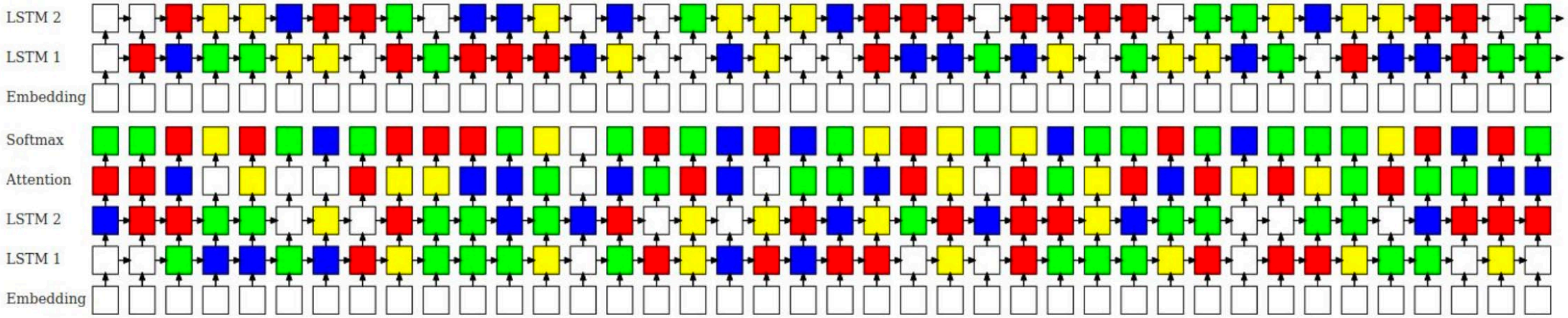


Neural Network

Execution Devices

images borrowed from Mark’s presentation and the original paper

Paper Recap: Device placement with RL



images borrowed from Mark's presentation and the original paper

Significance

- Not requiring expert design or specific rules
- We can inject less prior (bias) to the system

My Goal

- Use the algorithm on modern LLMs (e.g. LLaMA 7b)
- Would it be faster than doing all compute with CPU?
  - Would it bring benefit to a **home computer?**

- Secret sauce and ablations?

Extensions?

- Model: Tensorflow -> PyTorch (C++)? / Rust?
- Similar for the RL part
- Potential improvements to the system:
  - Better RL algorithm? - *PPO*
  - Better model for prediction? - *GNN*
  - Better overall design? – *Placeto*
- BO for parameters