■ FSI シンポジウム

" 東京大学 FSI シンポジウム「未来社会のための AI」" が，学外者も参加 OK だったので聴講に.

スターを観に行きたかっただけ，というミーハーな気持ちだったことは，正直否めない.

http://engineeringchallenges.org/challenges.aspx

- restore and improve uban infrastructure
  - combining vision with robotics - http://arxiv.org/abs/1603.02199, http://arxiv.org/abs/1806.10293, http://arxiv.org/abs/1704.06888, http://sermanet.github.io/imitate
- advance health informatics - expert care, anyware
- explainability
  - aliency map using integrated gradients, localization, attention basd models
- many advances depnd on being abel to understand tet
  - transformer model, "attention is all you need" - http://arxiv.org/abs/1706.03762
  - bidirectional encoder "BERT" - http://arxiv.org/abs/1810.04805
    - 1: pre-train a model on this "fill in the blanks" task using large-amounts of self-supervised text
    - 2: fine-tune this model on indivisual language
- engineer th tool of scientific discovery
  - https://www.blog.google/technology/ai/using-tensorflow-keep-farmers-happy-and-cows-healthy/
  - Deep learning for image-based cassava disease detection - www.ncbi.nlm.nih.gov/pmc/articles/PMC6553696
- AutoML: automated machine learning - https://cloud.google.com/automl
  - current: solution = ML expertise + data + computation
  - can we turn this into: solution = data + computation
  - neural architecture search with reinforcement learning - http://arxiv.org/abs/1611.01578
  - efficientnet: rethinking model scaling for deep convolutional neural networks - http://arxiv.org/abs/1905.11946
  - http://arxiv.org/abs/1904.07392
  - http://arxiv.org/abs/1901.11117
- More computational power needed
  - reduce precision ok
  - handful of specific operations
  - "What if 100M of our users started talking to their phones for three minutes per day?"
  - running speech models on CPUs, we'd need to double the number of computers in Google data-centers
  - TPUv1 - http://arxiv.org/abs/1704.04760
  - bfloat16 - http://arxiv.org/abs/1603.04467 - originally introduce by Google in our Tensor Flow white paper
    - multipier area and energy are proportional to the square of mantissa bits
  - TPUv2, 16GB of HBM, 600GB/s mem BW
    - TPUv2 pod 11.5 petaflows, 4TB HBM, 2-D toroidal mesh
  - TPUv3, 430 teraflops, 128 GB HBM
    - TPUv3 pod 100 peta flowps 32TB HBM
  - cloud.google.com/edge-tpu/
  - coral.withgoogle.com/
- What's wrong with how we do ML
  - current practice for solving a task with ML. data + ML experts -> solution
  - current practice for solving a task with AutoML. data -> solution
  - still start with little to no knowledge
    - transfer learning adn multi-task learning usually help, but are often done very modestly
- A visoin fo where we could go

- Bigger models, but sparsely activated
- Per-Example Routing - http://arxiv.org/abs/1701.06538
- What do we want:
  - large model, but sparsely activated
  - single model to solve any tasks(100s to 1Ms)
  - dynamically learn & grow pathways through large model
- https://ai.google/principles
  - 1. Be socially beneficial.
  - 2. Avoid creating or reinforcing unfair bias.
  - 3. Be built and tested for safety.
  - 4. Be accountable to people.
  - 5. Incorporate privacy design principles.
  - 6. Uphold high standards of scientific excellence.
  - 7. Be made available for uses that accord with these principles.
- ~75 google research papers form 2018/19 related ML bias, privacy and/or safety
  - http://ai.google/research/pubs/
- https://ai.googleblog.com/2019/01/looking-back-at-googles-research.html