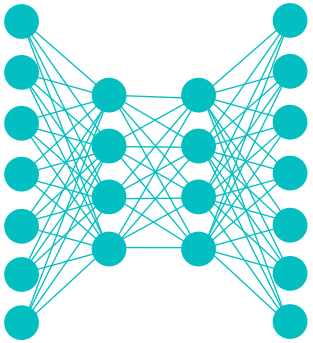
Lecture Notes for

Neural Networks and Machine Learning

Multi-task Learning

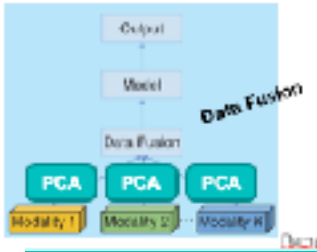


Logistics and Agenda

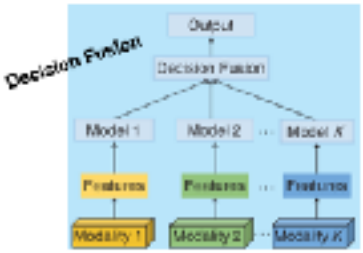
- Logistics
  - Back to regular office hours
    - No virtual option?
- Agenda
  - Multi-task networks and demo
  - Town Hall
- Next Time
  - CNN Visualization

Last Time

- **Early Fusion:** Merge sensor layers early in the process
- **Assumption:** there is some data redundancy, but modes are conditionally independent
- **Problem:** architecture parameter explosion
  - Need dimensionality reduction



- **Late Fusion:** Merge sensor layers right before flattening
- Use Decision Fusion on outputs
- **Assumption:** little redundancy or conditional independence—just an ensemble architecture
- **Problem:** just separate classifiers, limited interplay



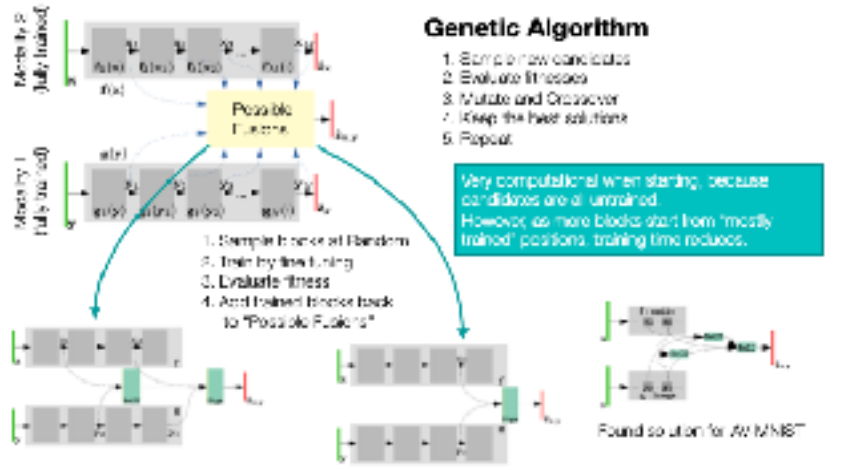
Applied Intelligence
https://doi.org/10.1007/s10489-022-04095-x

An approach for combining multimodal fusion and neural architecture search applied to knowledge tracing

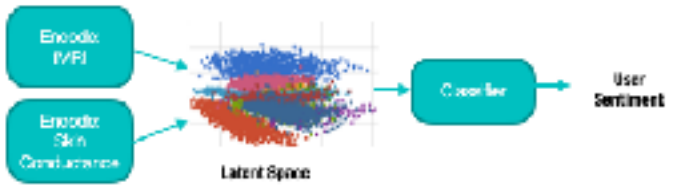
Xinyi Ding[1] · Tao Han[1] · Yili Fang[1] · Eric Larson[2]

Accepted: 18 August 2022

Neural Architecture Search for Mode Fusion



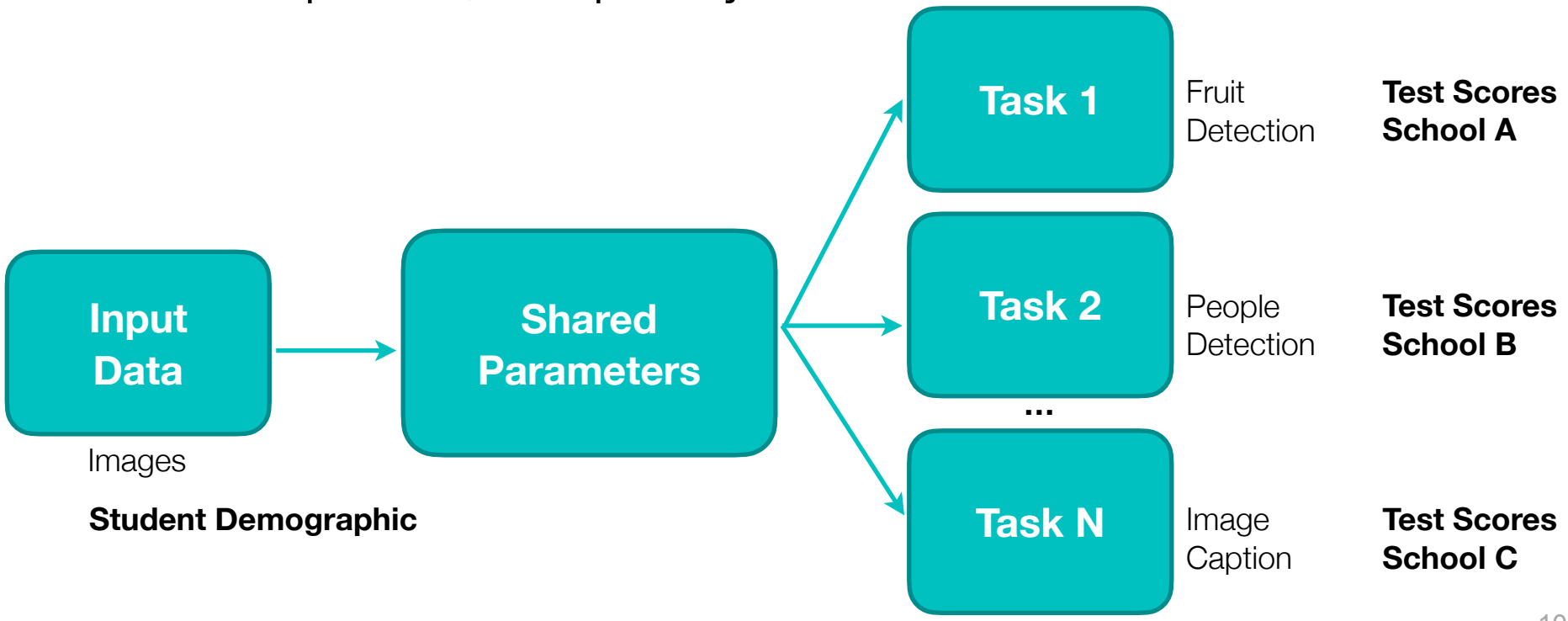
- Latent Space Transfer (universality)
  - From another domain, map to a similar latent space for the same task
  - Useful for unifying data based upon a new input mode when old mode is well understood
    - for example, biometric data
    - **I have never seen a research paper on this…**



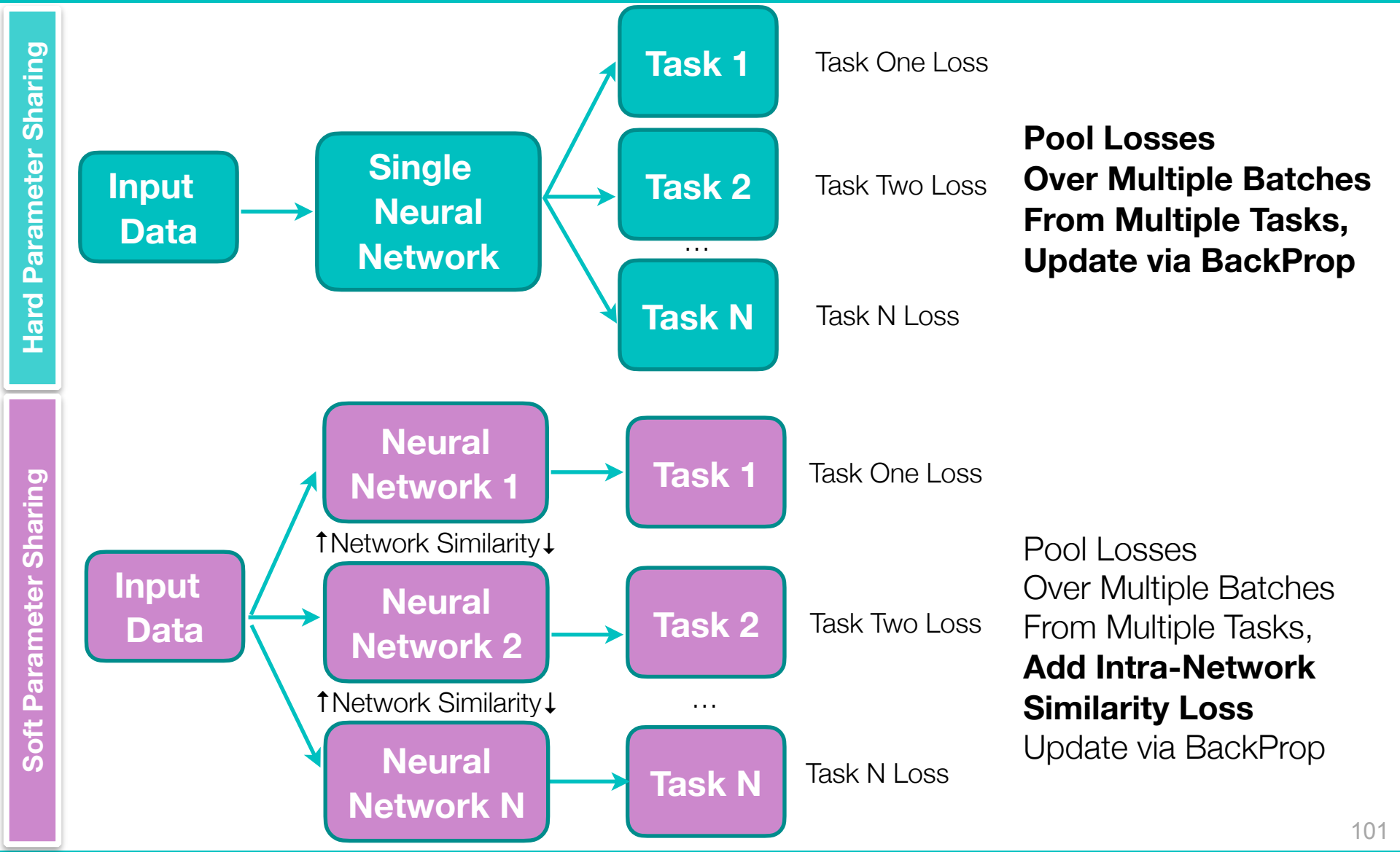
Multi-Task Models

Multi-task learning overview

- For deep networks, simple idea: share parameters in early layers
- Used shared parameters as feature extractors
- Train separate, unique layers for each task

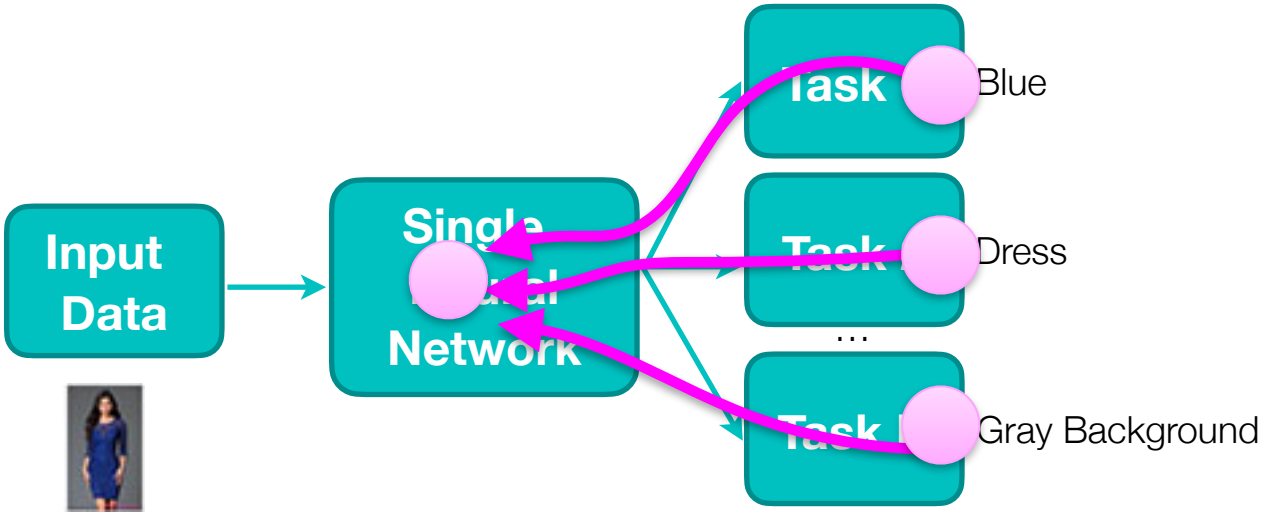


Multi-task Learning Parameter Sharing



Multi-task Optimization

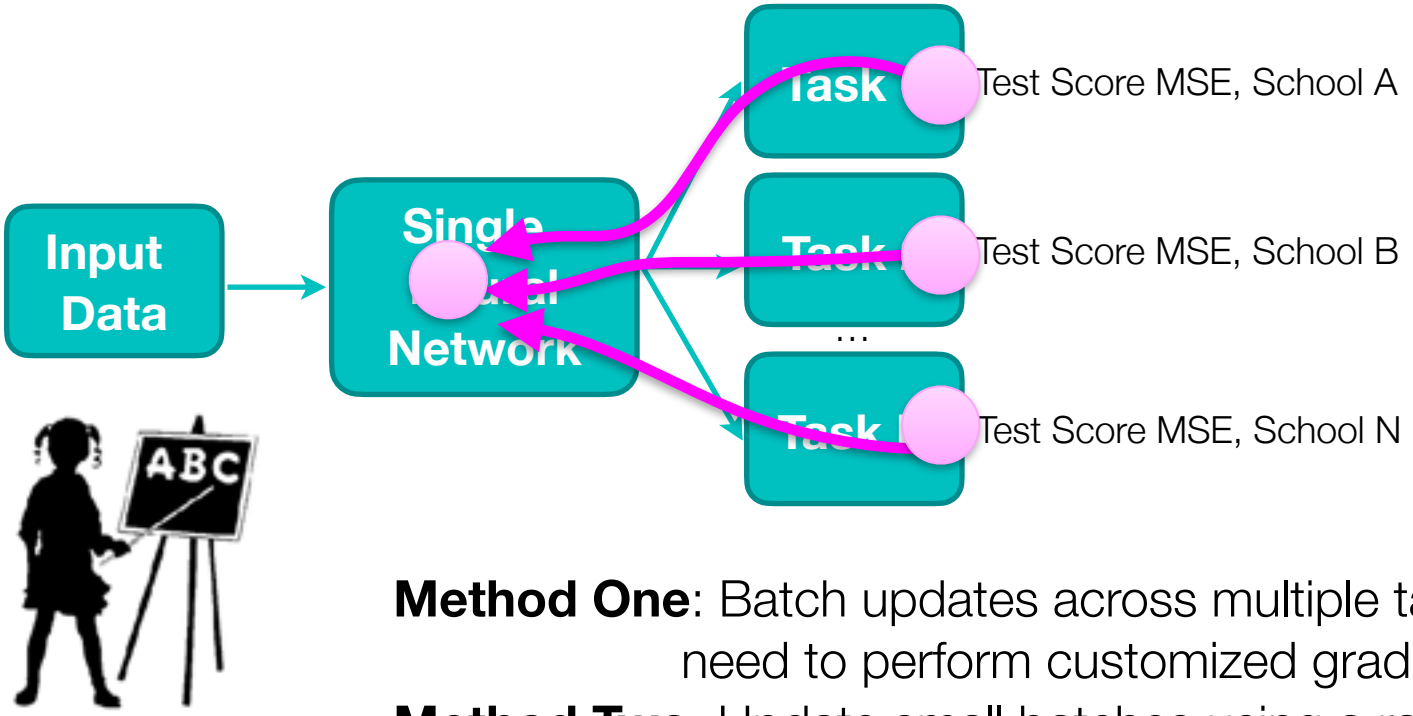
Multi-Label per Input



Measure Loss **for each label simultaneously**
Back propagate **everything at one time** for a given batch

Multi-task Optimization

Single Task Label per Input



Method One: Batch updates across multiple tasks
need to perform customized gradient calculations

Method Two: Update small batches using a random task
easier, but can cause instability in training

Multi-Task Learning in Keras with Multi-Label Data

Fashion week, colors and dresses

Optional Demo

Follow Along: https://www.pyimagesearch.com/2018/06/04/keras-multiple-outputs-and-multiple-losses/

Multi-Task Learning

School Data, Computer Surveys

Traian-Pop Traian Pop

Sponsor Follow

LukeWood Luke Wood

KerasCV Author, Full Time Keras team member & Machine Learning researcher @ Google, Part Time UCSD Ph.D student

PRO

Method One: Batch updates across multiple tasks
need to perform customized gradient calculations

Method Two: Update small batches using a random task
easier, but can cause instability in training

`Follow Along: LectureNotesMaster/03 LectureMultiTask.ipynb`

Lab Two Town Hall

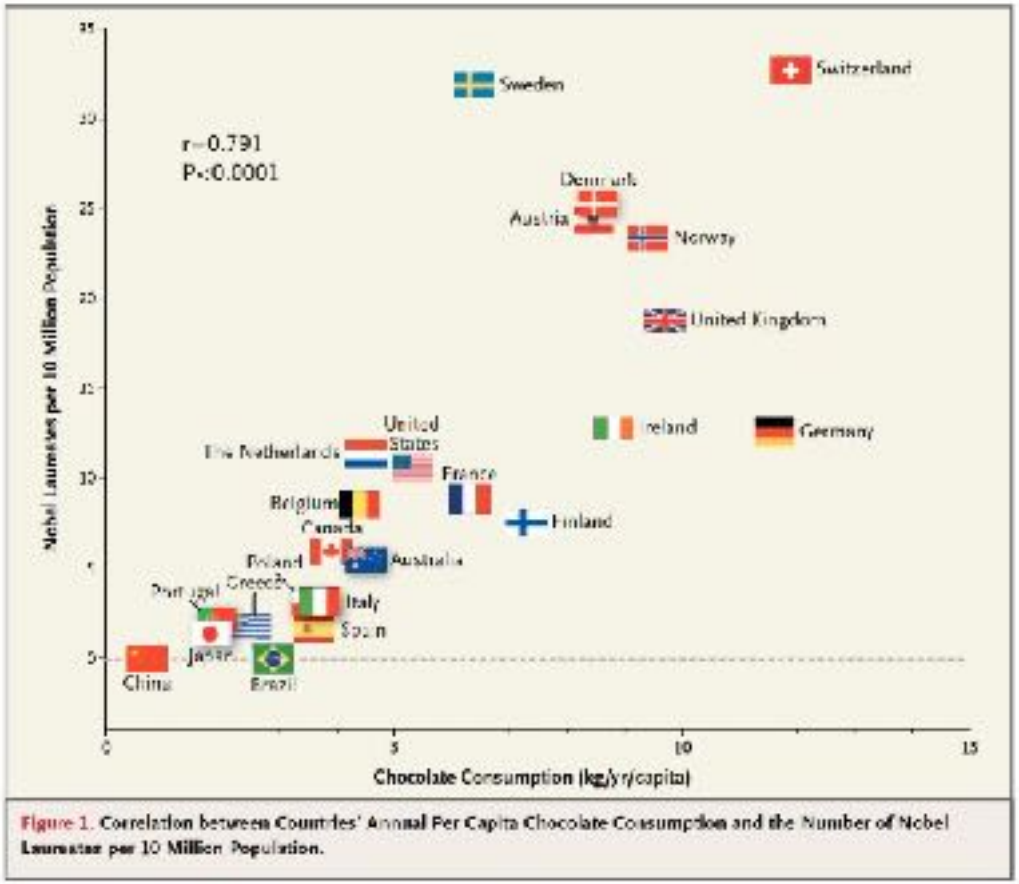


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Multi-Task Networks
Multi-Modal Networks

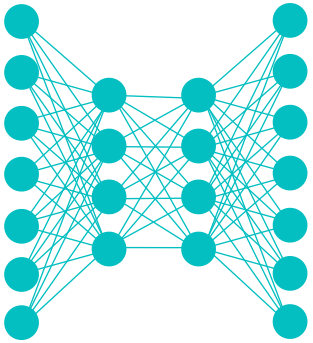
Lecture Notes for

Neural Networks and Machine Learning

Multi-Modal and Multi-Task

Next Time:
Circuits

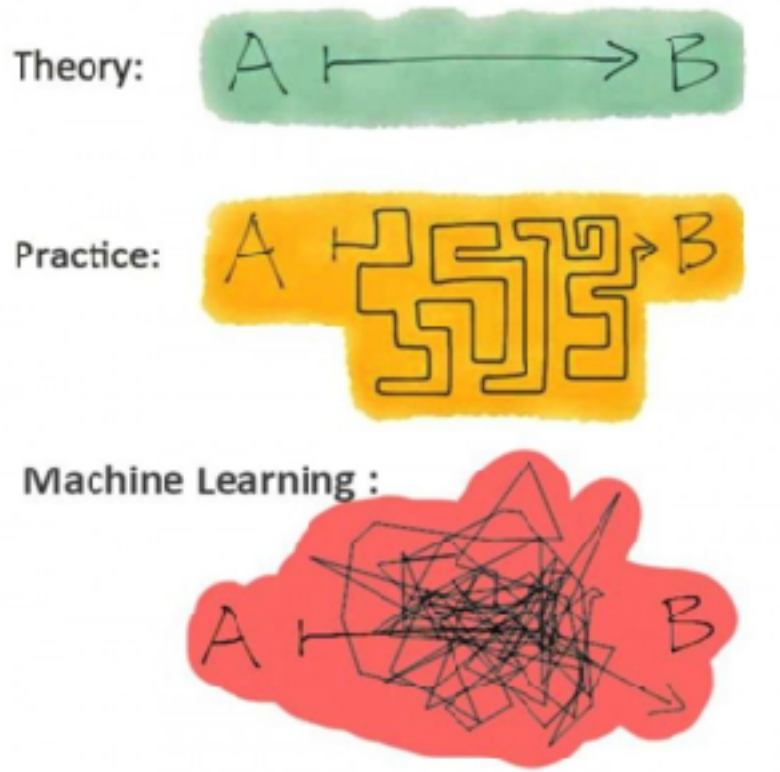
Reading: Chollet 8.1-8.5



Backup slides

Backup Slides

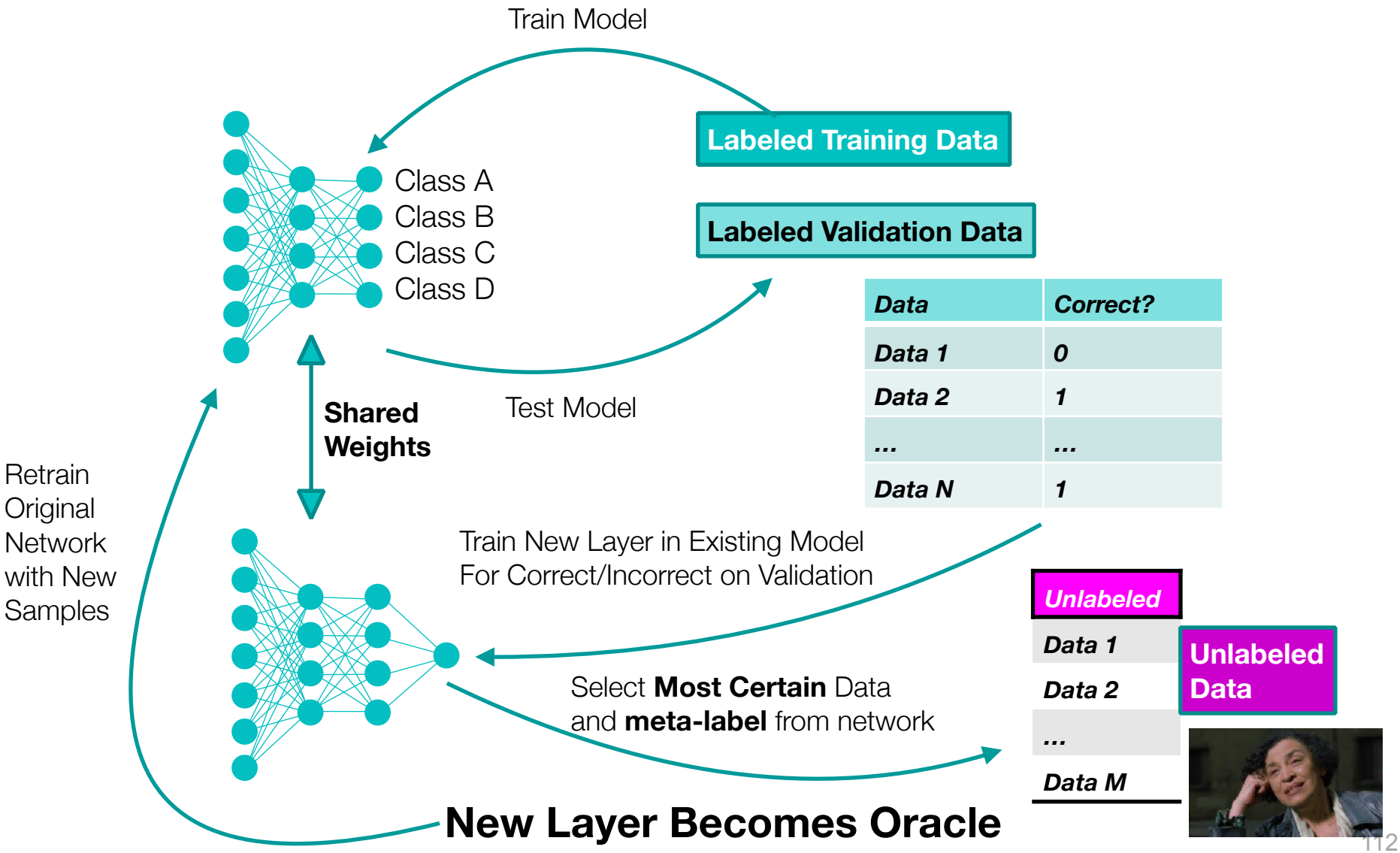
Active Transfer Learning



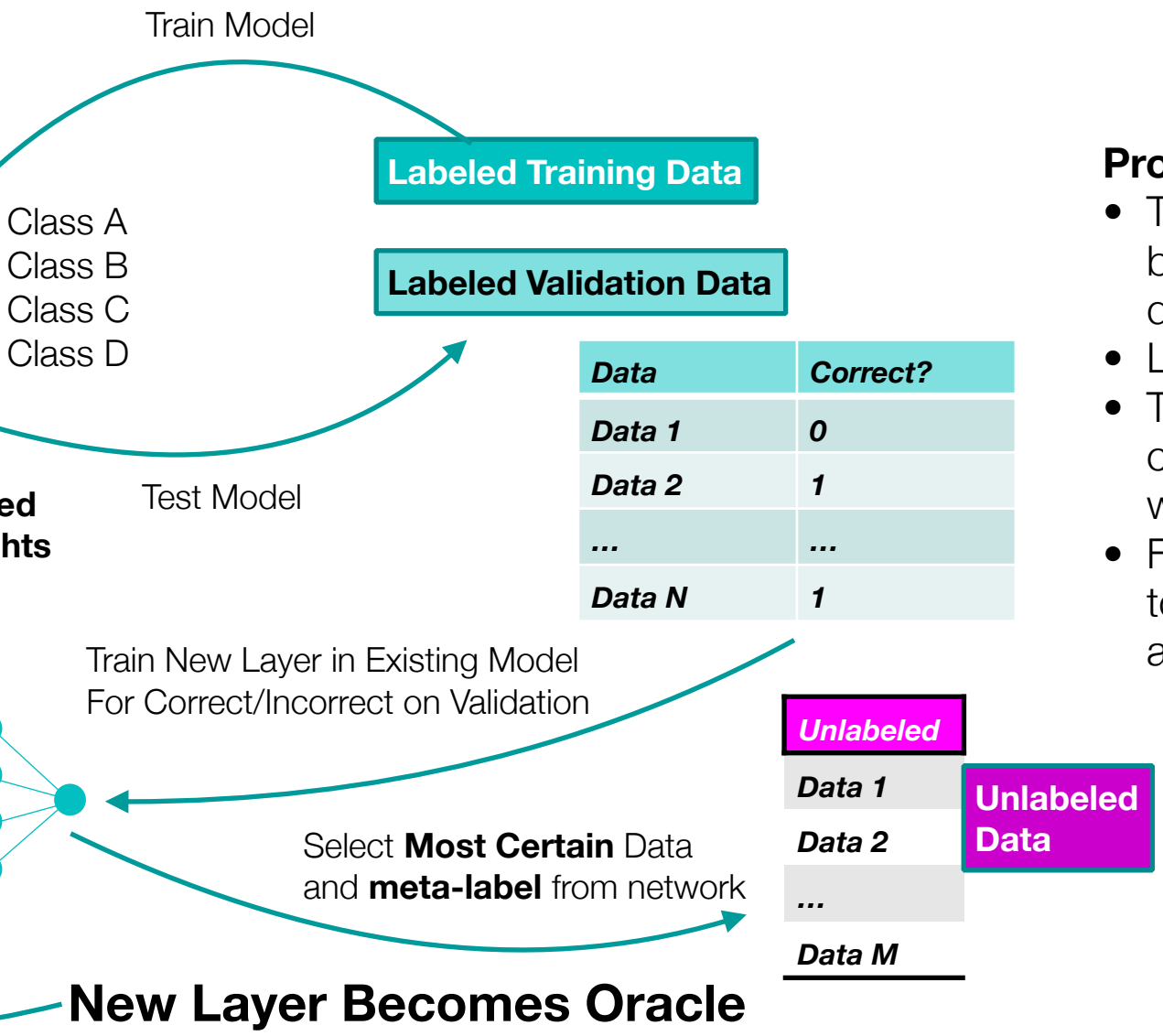
Active Learning Overview

- **Basic Idea**: Use a trained model to sample from an oracle that can magically give you a new label
  - Active Learning:
    What labels should we ask the oracle about?
- Uncertainty Sampling
  - Choose instances where the model is most uncertain or most certain
  - Various ways to measure certainty
- Diversity Sampling
  - Choose instances that are similar or different from training distribution

Uncertainty Sampling with a Neural Network



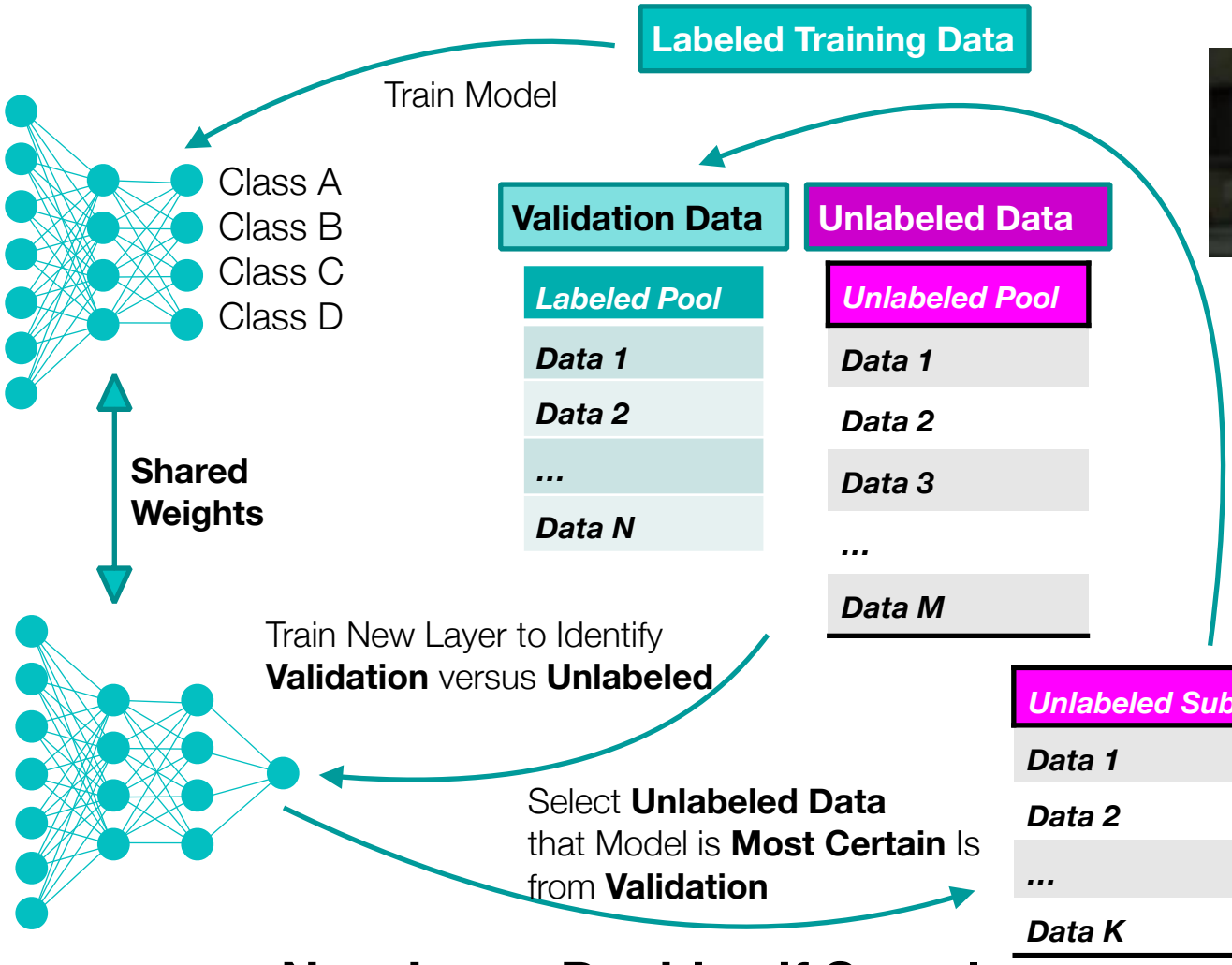
Uncertainty Sampling with a Neural Network



Problems:

- Training pool is represented by classes the model already does well predicting
- Limited diversity of Samples
- Training pool can become contaminated easily from a few wrong predictions
- For Oracle: we might be asking to get labels that the model is already good at classifying

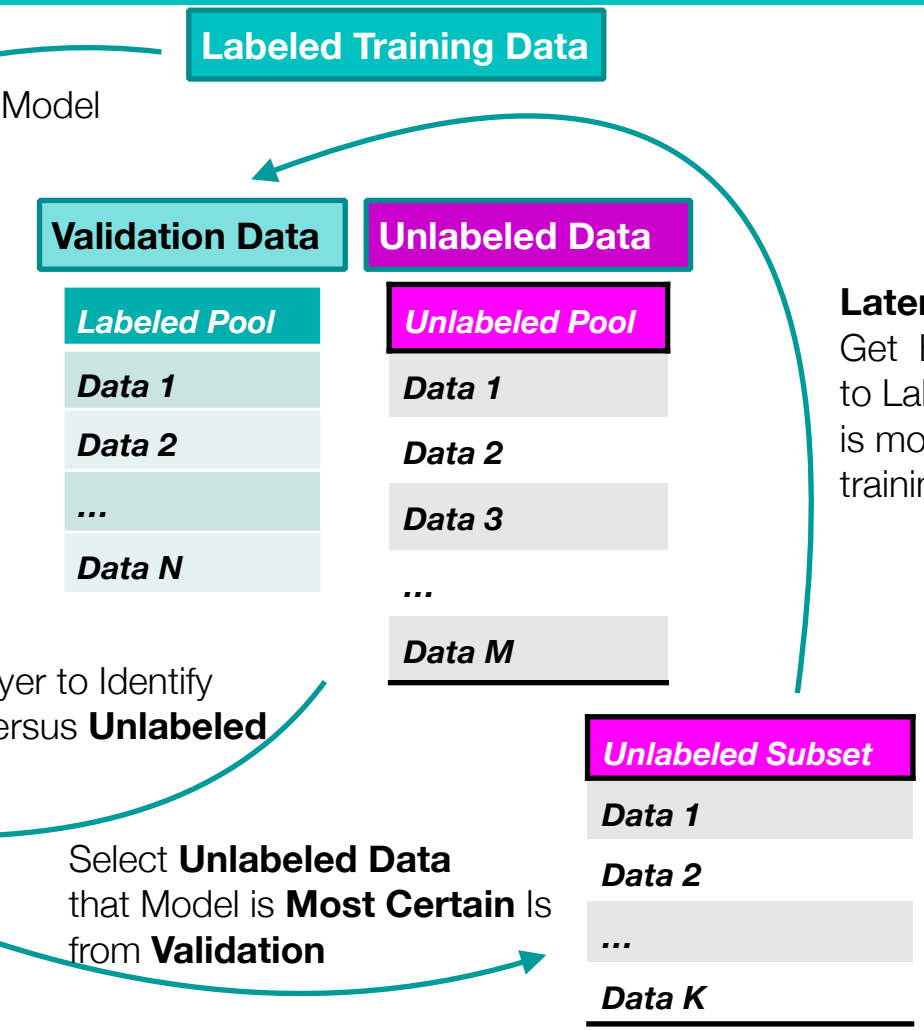
Diversity Sampling with a Neural Network



Later:
Get Human to Label data that is most dissimilar to training

New Layer Decides if Samples are Added to Validation Data

Diversity Sampling with a Neural Network

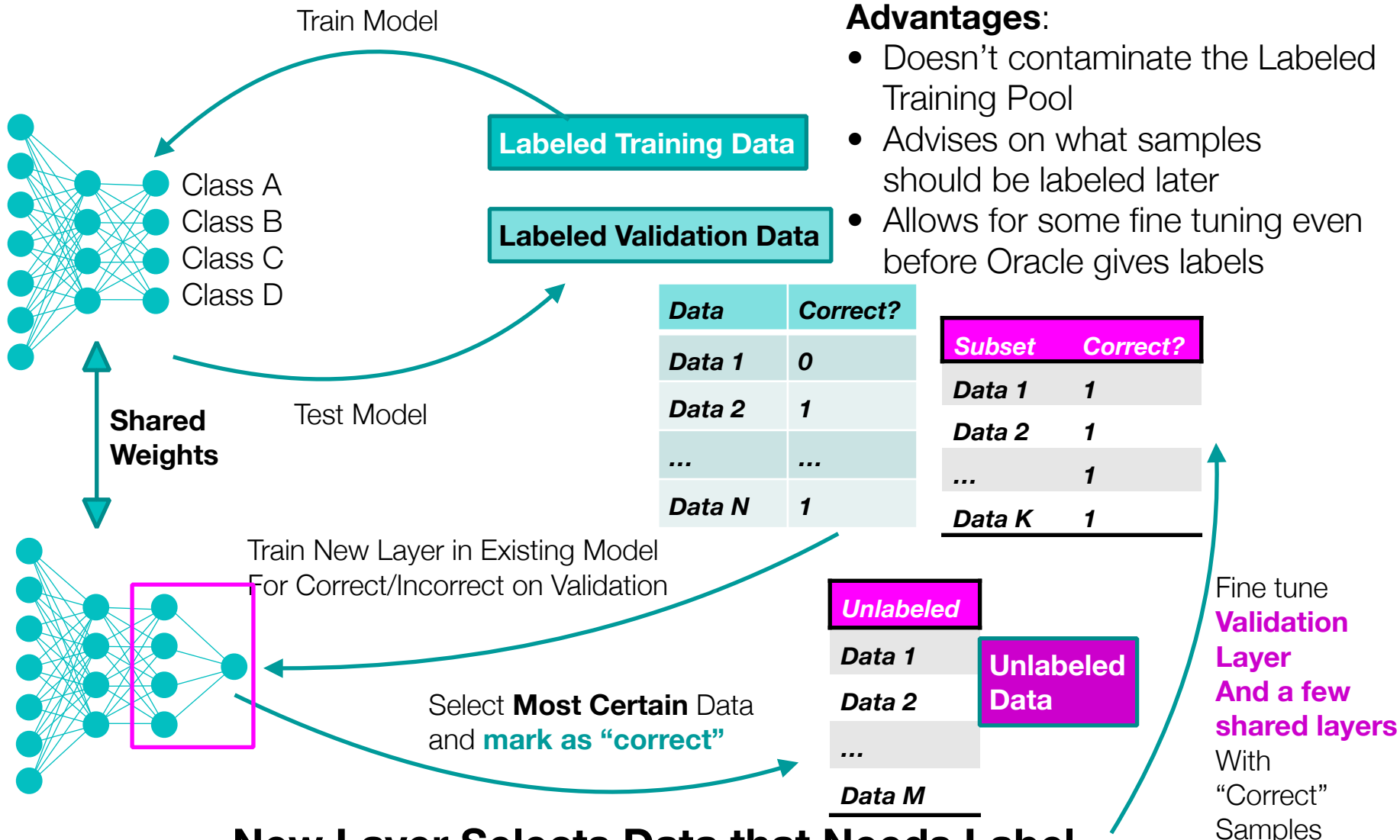


Later:
Get Human
to Label data that
is most dissimilar to
training

Discussion:

- Training pool is not contaminated
- Expands validation data in well mannered way, not adding too "far away" samples
- Validation versus Unlabeled might not be the best comparison, because it ignores confusions in the training data
- For Oracle: we can get labels to inputs that the model is likely to be unsure about
- But… this only helps us when we have an Oracle to give us labels

ATLAS: Active Transfer Learning for Adaptive Sampling



Advantages:

- Doesn't contaminate the Labeled Training Pool
- Advises on what samples should be labeled later
- Allows for some fine tuning even before Oracle gives labels

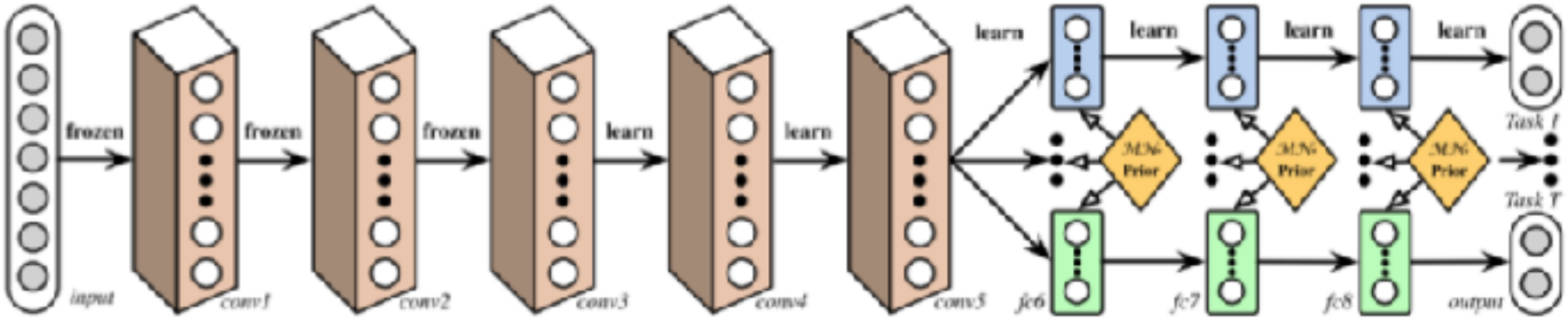
Time Period	Protocol	Expected Feedback
First Week	Homeowner provides 8-20 examples over the first week: • 1-2 Shower usages • 1 run of the dishwasher • 1 run of the laundry machine • 2 examples of each toilet • 1 example of hot and cold water use for each dual handle faucet • 1 example of hot, cold, and mixed water use for each single handle faucet (2 examples if in kitchen)	HydroSense relies on the rule based classifier for the first week. Pressure waves are saved in order to create a sparse codebook of features. Results are displayed at the fixture category for dishwashers, showers, and washing machines.
Start of Second Week	Homeowner provides 2-4 labels every other day when the system messages them on their mobile device	Results are displayed at the full fixture category level from the CoDBN-VE algorithm. Expected accuracy: • 85% at fixture category level
End of Second Week	Homeowner has supplied 9-12 examples that were flagged by active learning.	HydroSense now displays results at the Lumped Fixture level. Expected accuracy: • 82% at fixture level • 87% at fixture category level
End of Third Week	Homeowner continues to supply sparsely selected examples every other day. About 9-12 additional examples provided.	Valve level accuracy now provided. Expected accuracy: • 80% at valve level • 87% at fixture level • 92% at fixture category level
Fourth Week	Homeowner can optionally continue to provide examples to the system for increased accuracy.	Expected accuracy: • 81% at valve level • 89% at fixture level • 93% at fixture category level

Table 8-2. Expected feedback and calibration protocol for semi-supervised HydroSense system

Optional Topic: Multi-Task Model Examples

> He uses statistics like a drunken man uses a lamp post, more for support than illumination.
>
> -- Andrew Lang

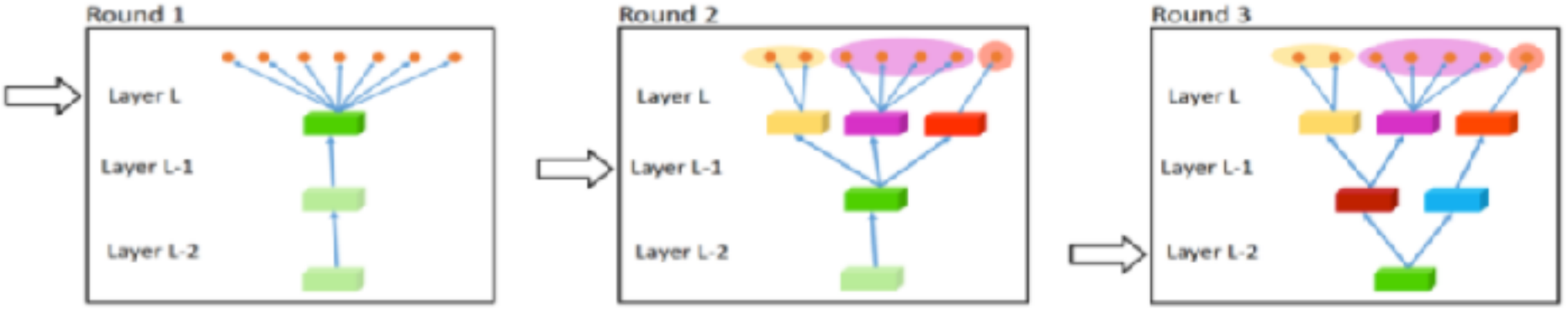
Multi-task: Deep Relationship Networks



- Start training traditionally (CCE)
- Minimize Kroenecker Product of fully connected task specific layers (here matrices are vectorized and therefore it is an outer product)
  - intuitively: make Covariances between tasks close to a given prototype Covariance
  - encourages feature maps in each task to be **less correlated** to feature maps of another task

https://arxiv.org/pdf/1506.02117.pdf

Multi-task: Adaptive Feature Sharing



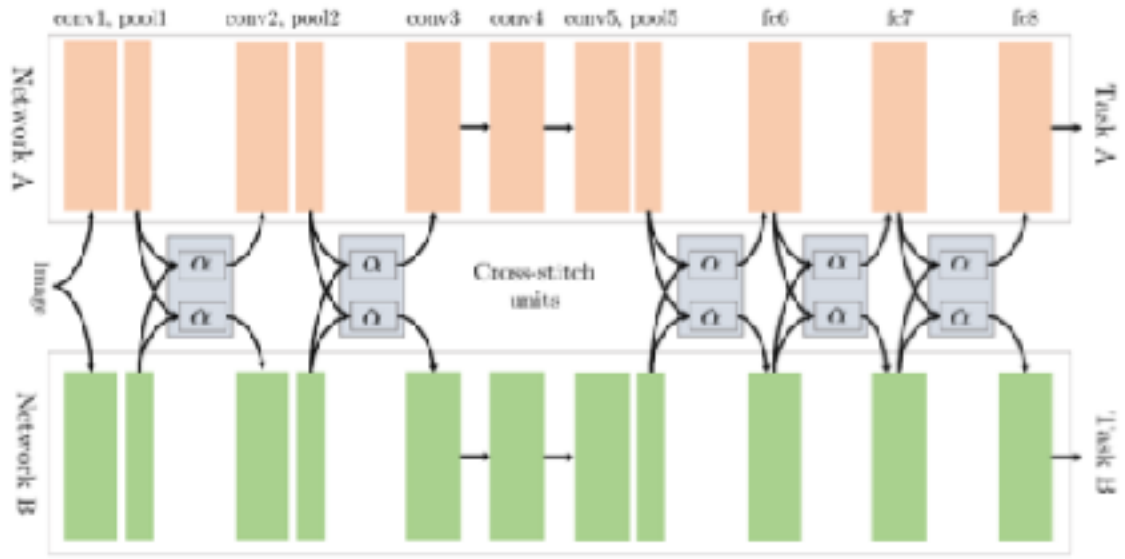
- Train
- Repe
  - D
  - Cut weights and retrain (fine tune) network
  - Decrement current layer index

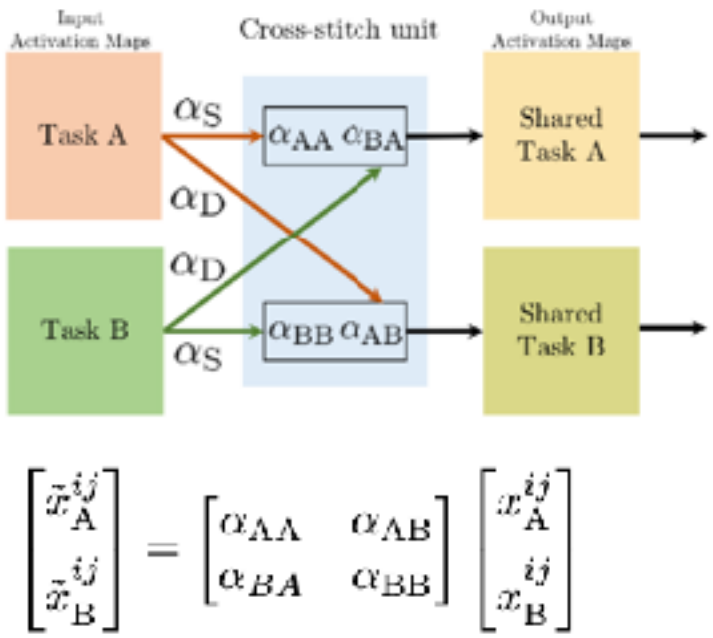
$$A^\star, \omega^\star(l) = \underset{A \in \mathbb{R}^{d \times d'}, |\omega| = d'}{\arg\min} ||W^{p,l} - AW^{p,l}_{\omega:}||_F, \qquad (2)$$

where $W^{p,l}_{\omega:}$ is a truncated weight matrix that only keeps the rows indexed by the set ω. This problem is NP-hard, however, there exist approaches based on convex relaxation

http://openaccess.thecvf.com/content_cvpr_2017/papers/Lu_Fully-Adaptive_Feature_Sharing_CVPR_2017_paper.pdf

Multi-task: Cross Stitch Networks

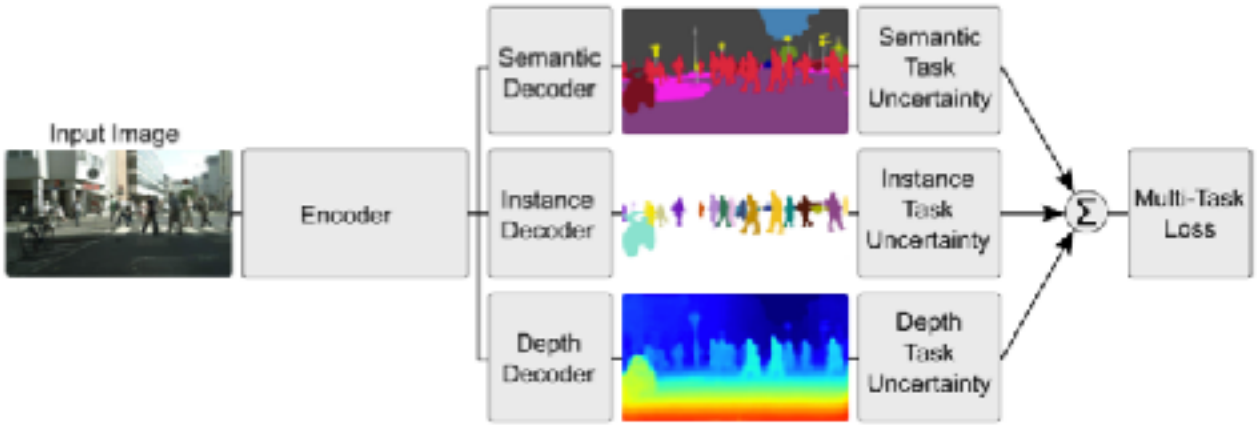




- Only works for simultaneous multi-label problems
  - like semantic segmentation and surface normal segmentation (clustering similarly facing objects)
- Take a learned weighted sum of the activations
- Works a little better than single task, but no worse

https://arxiv.org/pdf/1604.03539.pdf

Multi-task: Uncertainty Weighting



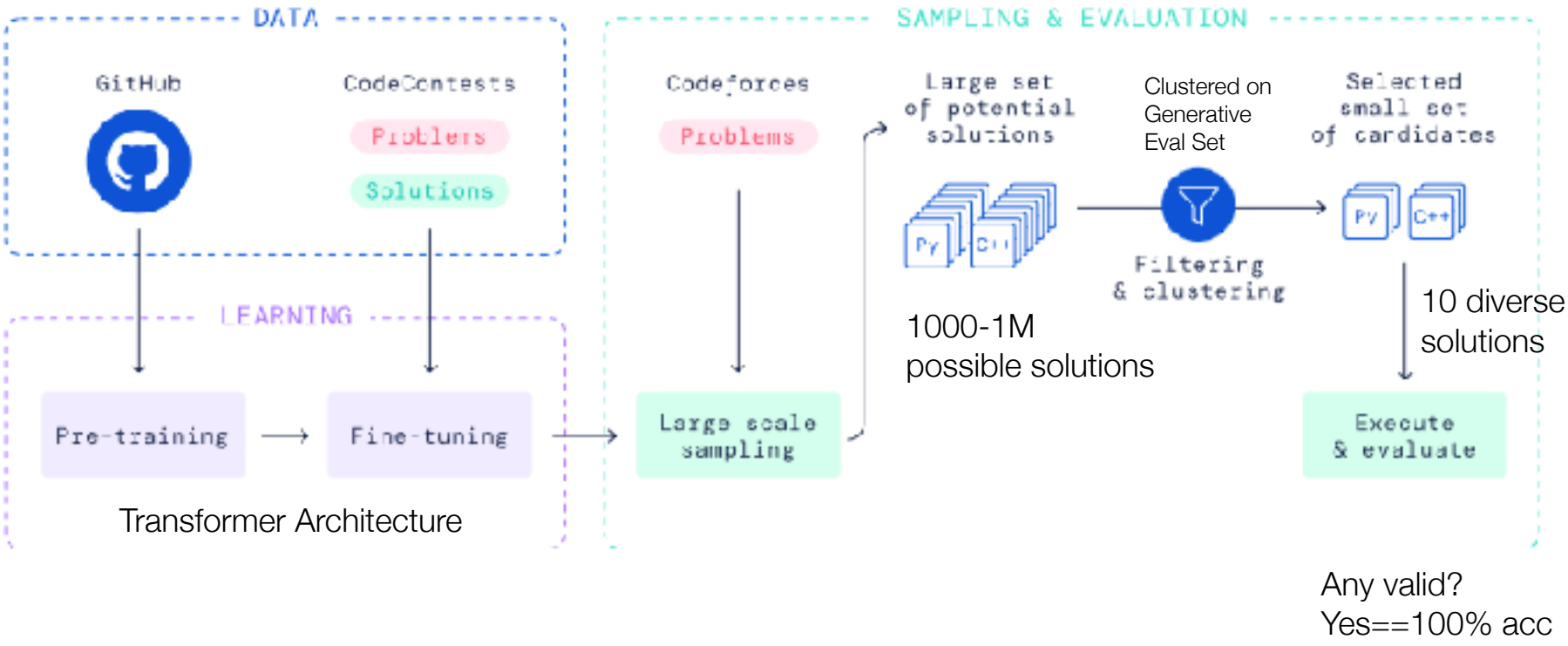
- Use variance of each loss function from each task to normalize
  - Which is a great idea, tasks with more uncertainty become less influential to loss function
  - call it homoscedastic without sound reasoning because that feels better than "normalized variance"
    - talk about homoscedasticity for no reason
- Write an entire paper in a "mathy" way to make contribution less clear
- Reviewers give you a pass because you have a history of doing good work

https://arxiv.org/pdf/1705.07115.pdf

Current Multi-task Research

- Incredibly diverse sets of solutions
- Mostly not evaluated on similar datasets
- Reasoning given is mostly ad-hoc…
- Theory is wildly under developed
  - because the problem is incredibly difficult
- Neural architecture search is an option…

AlphaCode



https://www.deepmind.com/blog/article/Competitive-programming-with-AlphaCode