Meta-learning for Bridging Labeled and Unlabeled Data in Biomedicine

Maria Brbić, Chelsea Finn and Jure Leskovec

Stanford University

Concluding remarks

Outline

- Practical tips
- Guidelines on available resources and libraries
- Demo example
- Future perspectives and concluding remarks

Practical Tips

- Data preprocessing is important
  - Scale the input properly
  - Visualize and verify data if possible
- Recommended optimizer: Adam
  - Fast, efficient and adapts the learning rate
- Usually 2-4 layers are enough
- ReLU activation function for hidden layers is usually a good choice

Practical Tips: Batch Normalization

- Batch normalization often helps

$$x_i \leftarrow \gamma \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$$

mean and standard-deviation calculated over the mini-batches

parameters

- Usually applied after a fully connected/convolutional layer and before a non-linearity
- Allows higher learning rates
- Reduces the strong dependence on initialization

Practical Tips: Prevent Overfitting

How to prevent overfitting?

- Dropout: randomly drop out neurons
- Weight regularization
- Early stopping: stop training when performance on a validation dataset starts to degrade

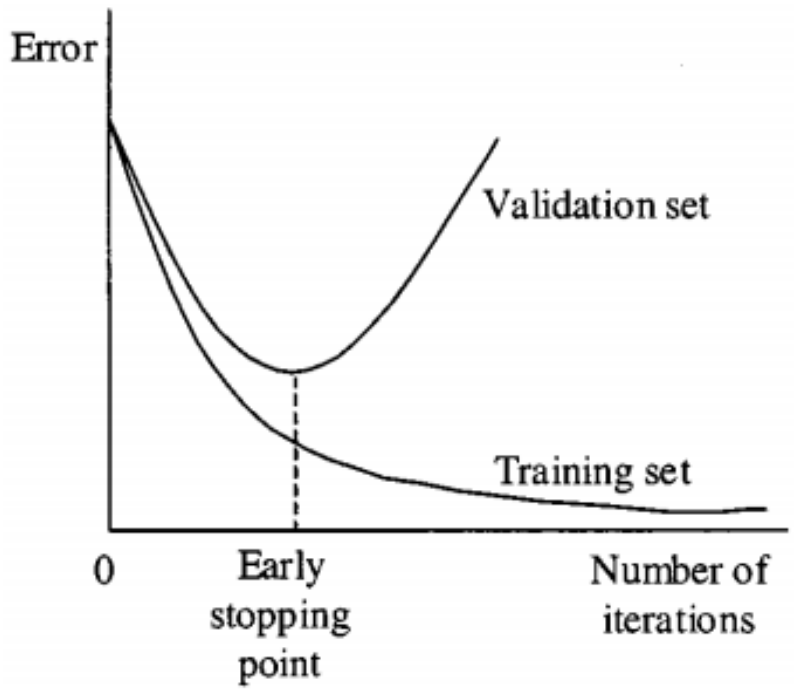


Figure from Gençay & Qi, *IEEE Trans on Neural Networks* '11

Practical Tips: Pretraining

- Pretraining often helps for low-data regime tasks and especially open-world learning

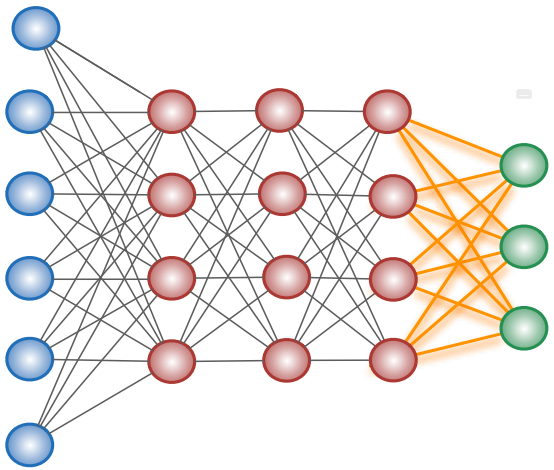
Pretraining strategies:

- Self-supervised pretraining: Need for an auxiliary task for which labels are easy to obtain
- Auto-encoder: Minimize data reconstruction loss

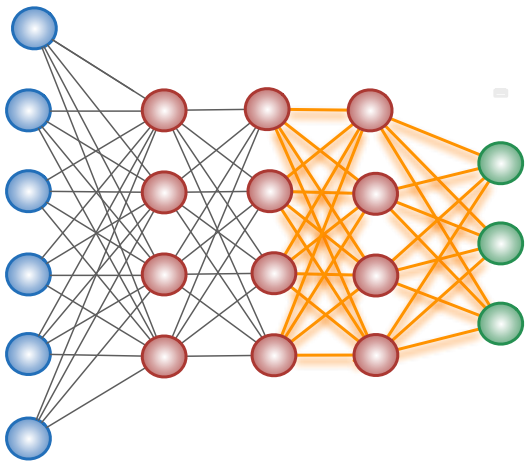
Practical Tips: Fine-Tuning

- Three ways of leveraging pretrained weights:

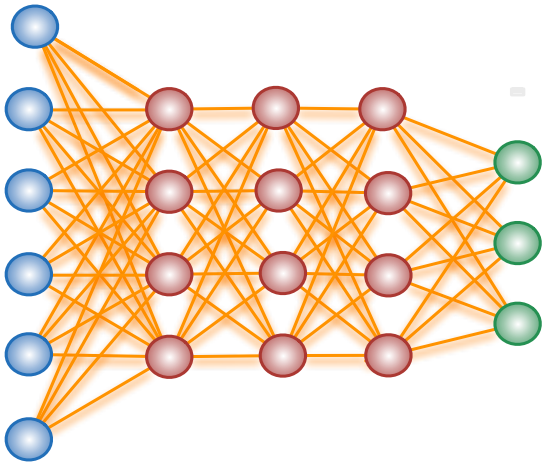
1) Fine-tune only last layer; freeze all others



2) Fine-tune last layers; freeze most others



3) Fine-tune all layers



Which strategy to use?

Depends how much labelled data you have.

Resources

- Deep learning code bases:
  - Examples in PyTorch, Tensorflow, Keras (high-level)
- Multi-task biomedical data:
  - Cross-tissue cell type discovery: http://snap.stanford.edu/projects.html: MARS, COMET
  - Smaller scale single-cell data (cross-species and cross—technology): http://snap.stanford.edu/mars/
  - Drug response prediction: https://github.com/idekerlab/TCRP/
  - MIMIC-III dataset: https://physionet.org/content/mimiciii-demo/1.4/ [demo], https://mimic.mit.edu/docs/gettingstarted/ [access]

Resources

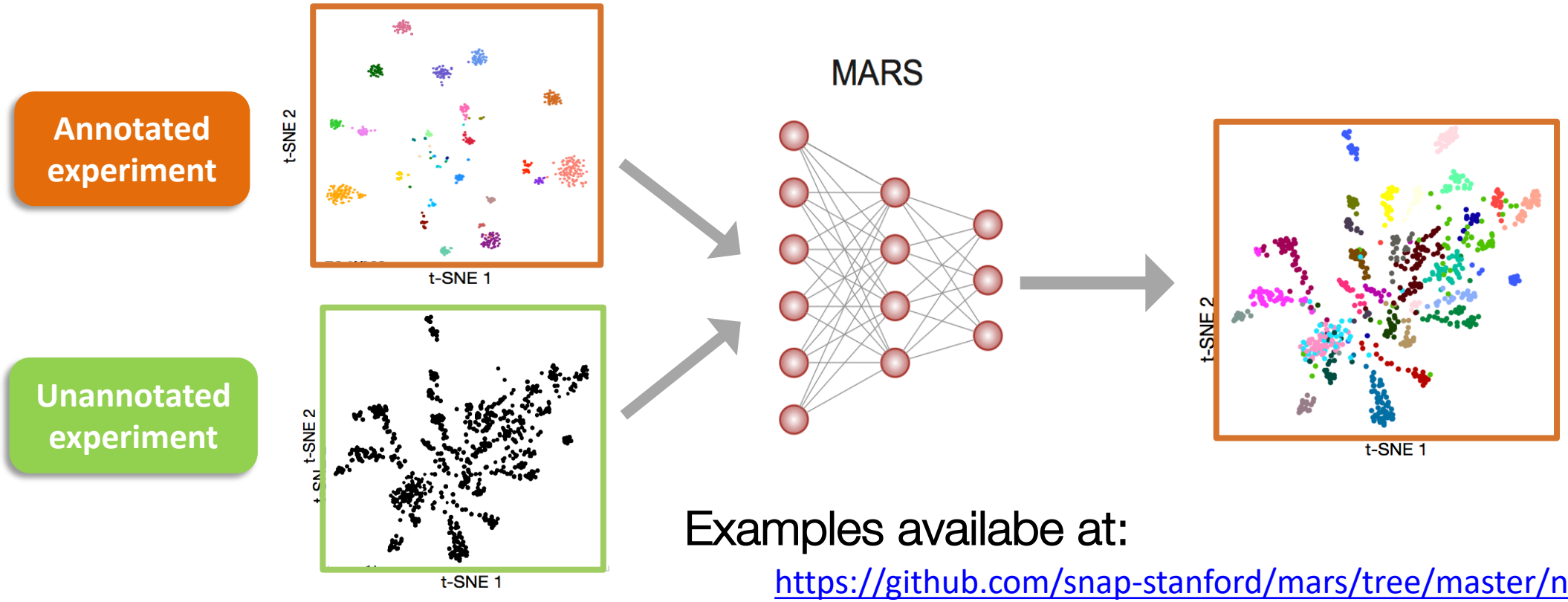
- Databases and ontologies for incorporating prior knowledge:
  - **Gene Ontology:** http://geneontology.org/docs/download-ontology/
  - **Cell Ontology:** http://www.obofoundry.org/ontology/cl.html
  - **Disease Ontology:** https://disease-ontology.org/downloads
  - Many ontologies available at **OBO Foundry**: http://www.obofoundry.org/
- **GOATOOLS:** Python library for Gene Ontology
  - https://github.com/tanghaibao/goatools

Code Bases

- MAML:
  - https://github.com/cbfinn/maml
  - https://github.com/dragen1860/MAML-Pytorch
  - MAML for cell lines prediction: https://github.com/idekerlab/TCRP/
- Matching Networks:
  - https://github.com/AntreasAntoniou/MatchingNetworks
  - https://github.com/BoyuanJiang/matching-networks-pytorch
- Prototypical Networks:
  - https://github.com/jakesnell/prototypical-networks
  - https://github.com/orobix/Prototypical-Networks-for-Few-shot-Learning-PyTorch
- COMET
  - https://github.com/snap-stanford/comet
- AM3
  - https://github.com/ElementAI/am3
- Integrated code of MAML, Matching Networks, Prototypical Networks, Relational Network:
  - https://github.com/wyharveychen/CloserLookFewShot
- MARS:
  - https://github.com/snap-stanford/mars

MARS Tool: Usage Example

```
mars = MARS(n_clusters, params, [annotated], unannnotated,
            pretrain_data, hid_dim_1=1000, hid_dim_2=100)
adata, landmarks = mars.train(evaluation_mode=False)
```



Examples availabe at:

https://github.com/snap-stanford/mars/tree/master/notebooks

Prototypical Networks for Cell Type Annotation

Prototypical networks for few-shot cell type annotations

This notebook is a part of the ISMB tutorial Meta-learning for bridging labeled and unlabeled data in biomedicine.

In this notebook, we wil use Prototypical networks to learn cell type annotations from only few-labeled as examples. We will use meta-learning approach to learn across different tissues and then generalize to a new, yet unseen tissues given only few labels.

Loading Tabula Muris dataset

We will use Tabula Muris mouse cell atlas single-cell data. This dataset can be downloaded at http://snap.stanford.edu/mars/data/tms-facs-mars.tar.gz.

In [5]:
```
import scanpy.api as sc
import numpy as np
```

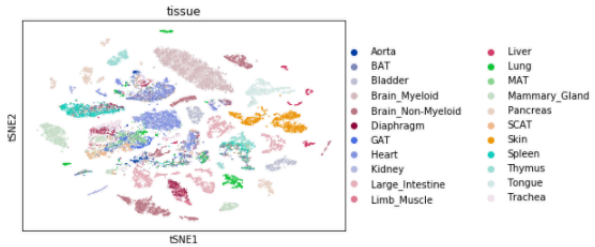
In [9]:
```
# set the path to the data
dataset_root = '../tabula-muris-senis-facs_mars.h5ad'
```

In [14]:
```
def load_sc_data(root):
    """Preprocess single cell data"""
    sc_data = sc.read_h5ad(dataset_root)
    sc_data = sc_data[sc_data.obs['age']=='3m'] # reduce file size by taking only 3months old mouse to be able to local
    sc.pp.filter_cells(sc_data, min_counts=5000)
    sc.pp.filter_cells(sc_data, min_genes=500)

    sc.pp.normalize_per_cell(sc_data, counts_per_cell_after=1e4)
    sc_data = sc.pp.filter_genes_dispersion(sc_data, subset = False, min_disp=.5, max_disp=None,
                                min_mean=.0125, max_mean=10, n_bins=20, n_top_genes=None,
                                log=True, copy=True)
    sc_data = sc_data[:,sc_data.var.highly_variable]
    sc.pp.log1p(sc_data)
    sc.pp.scale(sc_data, max_value=10, zero_center=True)
    sc_data.X[np.isnan(sc_data.X)] = 0

    return sc_data
```

In [15]:
```
tm_data = load_sc_data(dataset_root)
```

```
Trying to set attribute `.obs` of view, making a copy.
```

In [676]:
```
sc.pp.pca(tm_data)
sc.tl.tsne(tm_data)
sc.pl.tsne(tm_data, color = 'tissue')
```



How to get started?

- We prepared notebook with example of few-shot cell type annotation using Prototypical Networks
- Meta-learning across tissues on the single-cell mouse cell atlas data
- Available at: http://snap.stanford.edu/metalearning-ismb/

Future directions

Future Directions

- **So far:** Very few applications in biomedicine
  - Recent work shows that meta-learning requires magnitude less training examples compared to direct training to predict clinical outcome of cancer patients
  - Opportunity for reducing the amount of labeled data

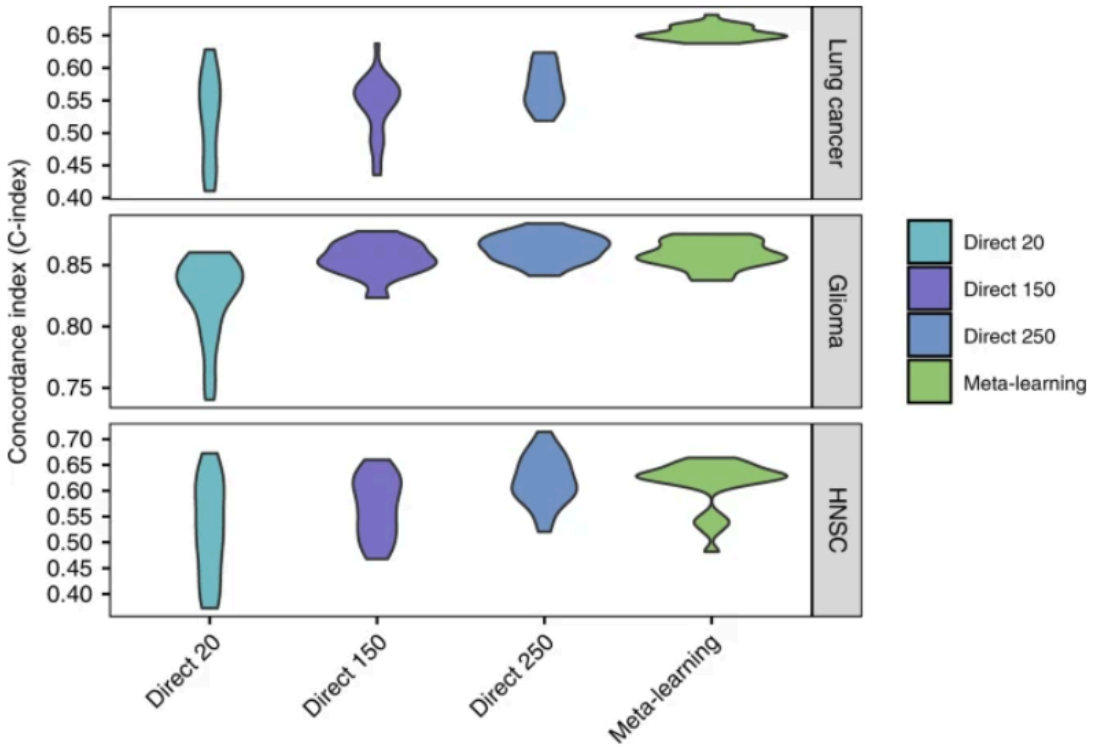


Figure from Qiu et al., *Nat. Commun. '20*

Future Directions

- Next goal: Broadening the range of applications

What is needed to get started?

1. Meta-dataset: Labeled dataset of related but different tasks
2. New dataset with unknown annotations

- Challenge: How to validate the model when used in practical settings?

Future Directions

- So far: Very few explainable methods

  How to trust a model based on only few-labeled predictions, or without any labeled examples?

  Interpretable models are essential for successful biomedical applications!

- Recent works (Rudin, Nat Mach Intell '19) emphasize issues with post-hoc interpretability models and advocate for building inherently interpretable methods

Tutorial Summary

- Few-shot learning: how to learn from only few-labeled examples
  - Meta-learning concept
  - Optimization-based and non-parameteric methods
  - How to incorporate prior knowledge in these methods
  - How to apply these methods in biomedical applications: drug response prediction, drug discovery, cell type annotation, disease prediction

Tutorial Summary

- Open-world learning: learning without any labels
  - How to discover novel classes
  - How to simultanously discover novel and recognize previously seen classes
  - How to apply these methods in biomedical applications: cell type annotation in single-cell genomics and CODEX imaging datasets

Meta-learning for Bridging Labeled and Unlabeled Data in Biomedicine

- Tutorial survey:

  https://forms.gle/y5s1Lm5yjn5pGxHJ8

- Slides with demo example:

  http://snap.stanford.edu/metalearning-ismb/