Forms of weak supervision

I. Small data

Data augmentation
- ex: image classification task
- add transformations
- label: invariant/semantically equivalent

-use simulator
- generate lots of data
- how realistic?

Multi-tasking
- one real task + one auxiliary task

real task: few labeled data
auxiliary task: lots of labeled data

Transfer learning
- sequential training: first on auxiliary task, then on real task
- pick a pre-trained model

conv: RegNet/VGG
ImageNet
2 pre-trained layers
100
- analysis from [Rethinking ImageNet pre-training]

Small data:
- helps in getting good features

Big data:
- 📝 From scratch: a $5x$ boost in training time

$$D \rightarrow \text{DNN} \times \text{DNN} \rightarrow S$$

Freeze train from scratch

If needed: unfreeze freeze-valid

IV. Forms of weak supervision

- Amount of data, but few are labeled

Semi-supervision:

 Semi-supervision: 

\[ \begin{align*} 
\text{Data set} & \quad \rightarrow \quad \text{labels} \rightarrow \mathbb{D}_2 \\
& \quad \rightarrow \quad \text{without labels} \rightarrow \mathbb{D} \\
\end{align*} \]

Ex: when labeling is costly (requires time, expertise, …)

Several approaches:

1) unsupervised training on full dataset $\mathbb{D} \rightarrow$ good representation $\rightarrow$ supervised task on $\mathbb{D}_2$

Ex: autoencoder

\[ \begin{align*} 
\text{unsupervised} & \quad \rightarrow \quad \text{supervised} \\
\times \mathbb{D} & \quad \rightarrow \quad \mathbb{D}_2 \rightarrow \mathbb{D} \times \mathbb{D}_2 \\
\end{align*} \]

2) supervised training $\rightarrow$ label some of the unlabeled samples $\rightarrow$ bigger dataset $\mathbb{D}_L \rightarrow \mathbb{D}_{\text{new}} = \mathbb{D}_L \cup \mathbb{D}_{\text{new labels}}$ 

- Issue: what if mistakes?
3) Supervised framing → apply to full dataset → check some properties 
→ adjust parameters

\( D \)

Properties:
- Bias
- Density
- Margin (SVM)

\( \mathbf{y} : \) Global bias

\( \text{Goal: } A \to B \)

\( \mathbf{y} = 50\% \)

\( \text{Go to } 70\% \)

\( \delta \) adjust the decision threshold

\( f_\theta(x) < 0 \)

\( f_\theta(x) > 2 \)

Weak supervision

Go more general:

\( \mathbf{y} : \) labels and numerical

Self-supervision

\( \) used for pre-training

\( \) no supervision (no label given by hand)

\( \) unsupervised task formulated as a supervised one

\( \) with automatic labels

\( \mathbf{y} : \) image classification

- Image puzzle

- Image rotation

\( \mathbf{y} : \) video classification

- Predict next frame

- Give 3 frames: ask temporal order

\( f_\theta(x) \)
Building on teacher-student techniques

- "ClusterFit":

```
\[ X \rightarrow \text{DDD} \rightarrow \hat{g} \]  dummy task
```

\[ \text{clustering} \rightarrow \text{add labels} \]

\[ \text{teacher} \]

Train a new network:

```
\[ X \rightarrow \text{DDD} \rightarrow \hat{g} \]  task: predict this label, i.e., which cluster \( x \) belongs to
```

\[ \text{student} \]

- DINO:

```
\[ \text{student} \rightarrow \text{DDD} \rightarrow \text{DDD} \]
\[ \text{randomly initialized} \]
```

\[ \text{teacher} \rightarrow \text{D} \rightarrow \ldots \rightarrow \text{D} \]

average of the past student (recent history) moving average

- Active-learning

```
\[ \text{same setting as semi-supervision} \rightarrow \text{ask some samples to be labeled} \]
\[ \text{costly labeling} \]
```

\[ \text{goal: increase global accuracy as fast as possible} \]
\[ \text{(in terms of # of labeled samples)} \]

\[ \text{large dataset: } D = \{x_i\} \]

- Costs for now: \( D_L = \{x_1, \ldots, x_p\} \)

\[ \text{with } p < |D| \]

- Which \( x_i \in D \) to pick? \( i \in [p, |D|] \)

\[ \text{to ask to be labeled} \]

* apply the current model \( f_\theta \) to all samples \( \rightarrow \) predictions \( \hat{g}_i = f_\theta(x_i) \)

\[ \text{classification task} \]

Local methods

\[ \rightarrow \text{quantify the impact of the choice of } x_i \text{ on the prediction} \]

for that point only
uncertainty: pick $x_i$ for which $f_0$ is the most uncertain

$$\arg\min_{x \in \mathcal{D}^*} \sup_{c \in C} \hat{y}_i^c$$

margin:

entropy: $H(\hat{y}_i^c) = \sum_{c \in C} \hat{y}_i^c \log \hat{y}_i^c$

query by committee:
if predictor = ensemble of $K$ models $\sum_{k=1}^{K} \hat{S}_{i,k}$

$\sum$ do models agree? pick $i$ where models disagree most

Global methods
$\sum$ quantify impact of the sample choice over all dataset examples

Expected model change
$f_\theta \rightarrow$ retrain $\rightarrow f_{\theta_{new}} \rightarrow$ apply to $\rightarrow$ impact?

just one

quantify the information gain as $\| \Theta_{e+1} - \Theta_{e} \|$

True label not known $\Rightarrow$ average over possibilities

$$E_{\hat{y}_i^c} \| V_0 \text{loss}(\hat{y}_i^c, \hat{\delta}_e^c) \|$$

$$\sum_{c \in C} \hat{y}_i^c \| V_0 \text{loss}(\hat{y}_i^c, \hat{\delta}_e^c) \|$$

$p(c^*) = \hat{y}_i^c$
Expected error reduction

\[ \sum_{i} \sum_{j \neq i} \text{prediction variation for sample } x_j \text{ if trained also with } (x_i, c) \]

\[ \theta_t \rightarrow \theta_{t+1} = -\eta \nabla_\theta \text{Loss}(x_i, c) + \eta \nabla_\theta \nabla_{\theta_0} f_\theta(x_j) + O(\theta^2) \]

\[ f_{\theta_{t+1}}(x_j) = f_{\theta_t}(x_j) + \frac{\delta \theta}{\theta_0} \nabla_{\theta} f_{\theta_t}(x_j) \]

Cons: Computational power
Post: Avoid choosing outliers
Both focus on yet-unclustered clusters of similar points

\[ \nabla_{\theta} f_{\theta}(x_i) \frac{dL}{d\theta} \nabla_{\theta} f_{\theta}(x_i) \]

(chose role)

\[ \nabla \text{influence functions} \]

\[ h(x_i, x_j) = \nabla_{\theta} f_{\theta}(x_i) \cdot \nabla_{\theta} f_{\theta}(x_j) \]

similarity