Forms of weak supervision

Small data

Data augmentation:
- ex: image classification
  - add transformations
    - Grid: inverse equivalent

Use a simulator:
- generate a lot of data
- how realistic? (mind the gap)

Multi-tasking
- one real task + one (auxiliary) auxiliary tasks

Transfer learning
- sequential training: first: on auxiliary task then: on real task
- unfreeze after convergence of new layers
- fine-tuning: minutes / attempts trying & hyper-parameter tuning

\[ \text{DDDDDD} \rightarrow \text{Y} \quad \text{aux task} \]
\[ \text{DD} \rightarrow \gamma \quad \text{new task} \]
\[ \text{DD} \rightarrow \gamma \quad \text{new task} \]

\[ \text{DDDDDDDD} \rightarrow \text{DD} \quad \text{real task} \]

\[ \text{DD} \rightarrow \gamma \quad \text{aux task} \]

\[ \text{DD} \rightarrow \gamma \quad \text{new task} \]

\[ \text{DD} \rightarrow \gamma \quad \text{new task} \]

\[ \text{DD} \rightarrow \gamma \quad \text{new task} \]
Analysis from [Rethinking ImageNet pre-training]

- Big data:
  - From training from scratch
  - Semi-Full Accuracy
  - Big boost in time

- Small data:
  - Helps in getting good features

---

Forms of weak supervision:

- Amount of data, but only few are labeled

  - Labeled $D_L$
  - Unlabeled $D_U$

  - when labeling is costly (requires time, expert knowledge...)
  - $\Rightarrow$ "semi-supervised"

Several approaches:

1) **Unsupervised training on full dataset $D$** → good representation → **Supervised task on $D_L$**

   - Un-supervised: $\rightarrow$ rebuild $x$
   - $D_L \rightarrow$ supervised task on $D_L$

2) **Supervised training** → label some of the non-labeled data → bigger dataset $D' = D_L \cup$ new labels

   - Iterate
   - Issue: what don't we know matches?

3) **Supervised training on $D_L$** → apply network to full dataset → check statistics / properties of your estimator / predictor

   - Adjust global parameters
   - Example: global bias, class imbalance (loss), binary class $P^*$: $A/A+B$ (40%, 60% bias, 55% correct, 65% estimated)
Self-supervision (approach 1)

used for pre-training

- no supervision => design supervised task so that labels are automatically assigned (aux)

ex. imageclassify:

\[ \text{real} \quad \text{pred} \rightarrow \text{classify} \]

Puzzle:

\[ \text{scramble} \quad \text{permutated} \rightarrow \text{image} \rightarrow \text{dummie task:} \]

where was the middle piece? 9 classes

Transformation:

\[ \text{rotation} \rightarrow \text{applying random rotation} \rightarrow \text{what was the angle \( \theta \)?} \]

(ex. video classify:

- predict next frame

- give 3 frames: DINO: asks temporal order?)

Teacher-student techniques (distillation)

"ClusterPhi":

\[ \text{student} \rightarrow \text{DINO} \rightarrow \text{y cluster} \]

DINO:

\[ \text{student} \rightarrow \text{DINO} \rightarrow \text{average of the last p student's} \]

\[ \text{randomly} \quad \text{(square architecture bias)} \]

\[ \text{initialized} \]

\[ \text{ask student} (x') \text{ to be close to} \text{teacher} (x) \]

\[ \text{learning to be invariant} \quad \text{to the group of transform} \]

\[ \text{ask adapt to have high variance} \text{(to prevent collapse)} \]

train new network to predict cluster class

\[ \text{activity space} \]

\[ \text{y cluster} \rightarrow \text{DINO} \rightarrow \text{cluster index} \]
Active learning

- Large database of unlabeled data
- Ask someone to label more samples (costly)
- Need: pick the most informative samples to label
  - less improvement over all dataset

Modeling 2 quantities:
- How non-confident are current predictions for that sample?
- How similar the sample is to many other unlabeled samples?

Noisy data

1) Noisy inputs
   - Denoising autoencoder
   - Remove noise
   - Learn to get rid of the noise
   - Model the noise

2) Classification with noisy labels
   - Fully labeled set: labels are often wrong
     - If labels are wrong: ok
     - So:
     - Random:
       - noise cancels out on average

3) Regression with noisy labels
   - Noise: central (noisy)
   - Quantify similarity
   - Slow: NTK
   - \[ \sum_i \| \hat{f}(x_i) - y_i \|^2 \]
   - \[ \hat{f}(x) = 5 \pm \sqrt{N} \]