Transfer Learning

Self-Supervision

Learning goals

- Understand the difference to other learning paradigms
- Learn to recognize self-supervision when you see it

DEFINITION

Unsupervised Learning:

- No labels attached to the data
- Learn patterns / clusters from the features only

Supervised Learning:

- (Gold) Labels attached to the data
- Learn from the association between features and labels

Self-Supervised Learning:

- No *external* labels attached to the data
  $\rightarrow$ Samples with suitable labels can be generated from the known structure of the data itself
- *Technically* supervised learning, but *no labeling effort* + simultaneous ability to generate massive amounts of labeled data points

SELF-SUPERVISED OBJECTIVES

Recap: Language modeling

- *Training objective:* Given a context, predict the next word

Illustration (context size $= 2$)

| The | quick | brown | fox | jumps | over | the | lazy | dog |

$\Rightarrow$ (the, quick)

| The | quick | brown | fox | jumps | over | the | lazy | dog |

$\Rightarrow$ ([the, quick], brown)

| The | quick | brown | fox | jumps | over | the | lazy | dog |

$\Rightarrow$ ([quick, brown], fox)

| The | quick | brown | fox | jumps | over | the | lazy | dog |

$\Rightarrow$ ([brown, fox], jumps)

SELF-SUPERVISED OBJECTIVES

Recap: Skip-gram

- *Training objective:* Given a word, predict the neighbouring words
- *Generation of samples:* Sliding fixed-size window over the text

Illustration

The	quick	brown	fox	jumps	over	the	lazy	dog

$\Rightarrow$ (the, quick); (the, brown)

The	quick	brown	fox	jumps	over	the	lazy	dog

$\Rightarrow$ (quick, the); (quick, brown); (quick, fox)

The	quick	brown	fox	jumps	over	the	lazy	dog

$\Rightarrow$ (brown, the); (brown, quick); (brown, fox); (brown, blue)

The	quick	brown	fox	jumps	over	the	lazy	dog

$\Rightarrow$ (fox, quick); (fox, brown); (fox, jumps); (fox, over)

SELF-SUPERVISED OBJECTIVES

Self-supervised objectives:

- Skip-gram objective (cf. word2vec ▸ Mikolov et al. (2013a))
- Language modeling objective (cf. ▸ Bengio et al. (2003))
- *Masked language modeling (MLM)* objective (cf. chapter 6)
  $\rightarrow$ Replace words by a [MASK] token and train the model to predict
- *Permutation language modeling (PLM)* objective (cf. chapter 6)
  $\rightarrow$ Autoregressive objective of XLNet
- *Replaced token detection* objective (cf. chapter 6)
  $\rightarrow$ Requires two models: One performing MLM & the second model to discriminate between actual and the predicted tokens

PRE-TRAINING VS. FINE-TUNING

Pre-training:

- Using unlabeled corpora in conjunction with self-supervised objectives is commonly referred to as *Pre-Training* the model
- Generation of samples for pre-training basically effortless, exploiting the inherent structure of the text
- Construction of different self-supervised objectives, which are assumed
    - to cover different phenomena better than the others
    - to work more efficiently for learning

PRE-TRAINING VS. FINE-TUNING

Fine-tuning:

- The second phase of transfer learning, i.e. adapting the pre-trained model to a labeled data set for a specific downstream task is referred to as *Fine-Tuning*
- Far less labeled data required compared to a scenario w/o pre-training
- Also possible (cf. chapter 7): No fine-tuning, but ..
    - *Zero-Shot Transfer* w/o ANY labeled data
    - *Few-Shot Transfer* w/ FEW labaled data points
- In both of the latter cases, good pre-training becomes even more important