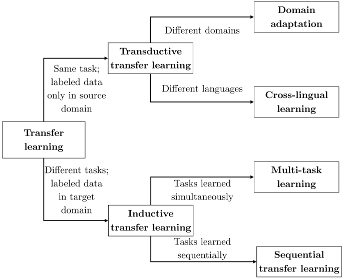
Transfer Learning

Basic definitions and challenges



Learning goals

- Differentiate the different flavors of transfer learning
- Understand the challenges we might be able to overcome by using transfer learning

FEATURE-BASED TRANSFER LEARNING

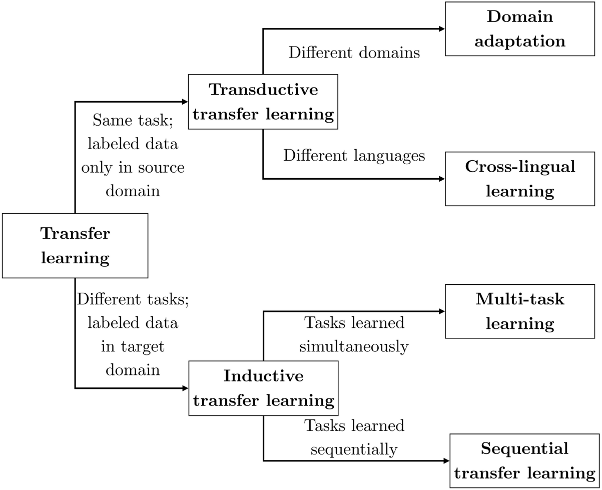
How it works with word2vec

- Train word2vec on some "fake task" (CBOW or Skip-gram)
- Extract the stored knowledge (a.k.a. embedding)
  or: Directly download embeddings from the web
- Perform a different (supervised) task using the embeddings

How it works with ELMo

- Do not *extract* the stored knowledge, but use the whole embedding model *as is*
- Only train/fine-tune task specific weights on top of ELMo

TAXONOMY OF TRANSFER LEARNING ▸ RUDER, 2019



Source: *Sebastian Ruder*

TAXONOMY OF TRANSFER LEARNING ‣ RUDER, 2019

Transductive Transfer learning

- Domain adaptation:
  $\rightarrow$ "*Transfer knowledge learned from performing task A on labeled data from domain X to performing task A in domain Y.*"

- Cross-lingual learning:
  $\rightarrow$ "*Transfer knowledge learned from performing task A on labeled data from language X to performing task A in language Y.*"

- *Important:* No labeled data in target domain/language *Y*.

TAXONOMY OF TRANSFER LEARNING ▸ RUDER, 2019

Inductive Transfer learning

- Multi-task learning:
  $\rightarrow$ "*Transfer knowledge learned from performing task A on data from domain X to performing multiple (simultaneous) tasks B, C, D, .. in domain Y.*"

- Sequential transfer learning:
  $\rightarrow$ "*Transfer knowledge learned from performing task A on data from domain X to performing multiple (sequential) tasks B, C, D, .. in domain Y.*"

- *Important:* Labeled data only for task(s) from target domain *Y*.

REMARK ON MULTILINGUALITY

Cross-lingual transfer:

- Languages can be grouped into certain families
- Patterns that a model learns for one language, might be beneficial for learning a second language (just as it is for us humans as well: For those who learned French in high school, learning Spanish afterwards might be easier)
- Again: Scarcity of resources; assume the following scenario:
  - **Large** parallel corpus for languages A and B
  - **Large** parallel corpus for languages A and C
  - *Small* parallel corpus for languages B and C

  $\rightarrow$ Training a model for B and C in isolation not the best idea

DEFINITION: SELF-SUPERVISION

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., 2013)
- Language modeling objective (cf. e.g. ▸ Bengio et al., 2003)
- *Masked language modeling (MLM)* objective (cf. BERT)
  $\rightarrow$ 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