Domain Adaptation $\subset$ Transfer Learning

Saptarshi Sengupta, 2nd Year PhD, Informatics, Penn State

Definitions

I feel that for both domain & task, it should be called prior dist. since with marginal prob. we assume more than 1 independent variable…

1. Domain ($\mathbb{D}$): $\{\chi, P(X)\}$ where,
    a. χ is the feature space
    b. $P(X)$ is the marginal probability distribution over the feature space
    c. $X = \{x_1, \cdots, x_n\} \in \chi$ where x_i *is the* i^{th} training instance (without label)
2. Task (T): $\{\mathcal{Y}, P(Y), P(Y|X)\}$ where,
    a. $\mathcal{Y}$ is the label space
    b. $P(Y)$ is the prior distribution over the label space
    c. $P(Y|X)$ is learned from the training data consisting of pairs of $\{(x_i \in X, y_i \in \mathcal{Y})\}$

src: [Neural Transfer Learning for Natural Language Processing (PhD thesis)]

Example:

- $\mathcal{T}$ = Sentiment Classification for tweets
- $\mathcal{Y}$ = {Happy, Sad, Angry}
- χ = Embedding space for tweets
- $X = \{x_1, \cdots, x_n\}$ = Collection of tweets
- Training instances = {(tweet_1, happy), ..., (tweet_n, sad)}

Got enough data?

- What happens when you want to apply a ML/DL model to a very ***niche space*** like vaccine literature, studies on coral reefs, etc.?
- You could,
  - Train a model from scratch in this space
    - But you most likely won't have enough data to train a decent model
  - Try creating a larger dataset with more & annotations
    - Expensive & time consuming
  - Leverage domain specific knowledge bases/graphs?
    - But they are not always available for all domains

Solution Please?

Transfer Learning

> Main Idea -> To leverage/repurpose the knowledge of the task & data acquired in the source domain for the target domain

> Advantages

1. **Avoiding cold start**
   a. No need to train a new model every time (analogous to human learning)
   b. i.e. Less expensive + Time consuming
2. **Avoiding data sparsity**
   a. Not starting with a randomly initialized model to learn the target task with limited data
   b. i.e. More resourceful
3. **Retraining possibility?**
   a. The target domain *might* contain examples that reteach the model about what it knows
   b. i.e. Support for lifelong learning

Transfer v/s Ordinary Learning setups

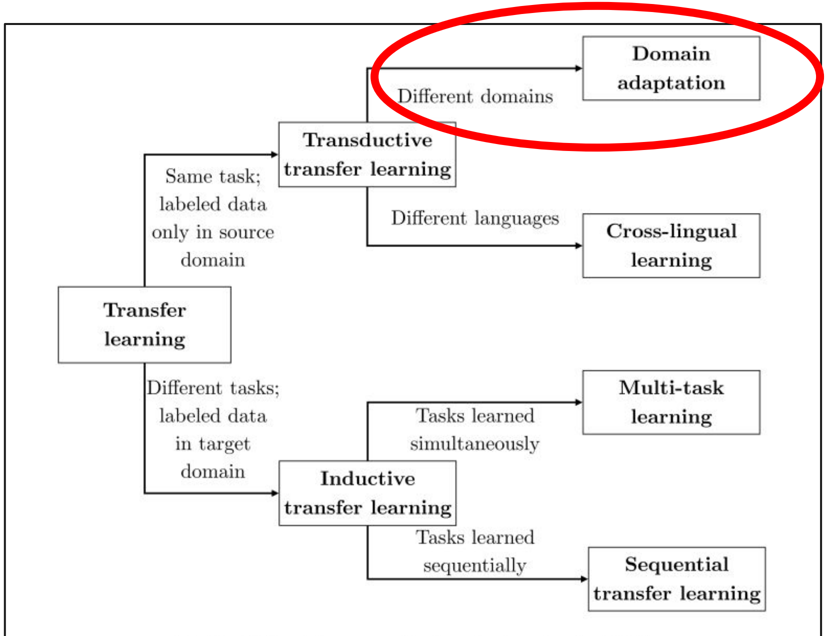

FIGURE 3.3: A taxonomy for transfer learning for NLP.

$\mathcal{D}_{source} \neq \mathcal{D}_{target}$

$\mathcal{T}_{source} \neq \mathcal{T}_{target}$

Usual Learning Setting is where, Source & Target,

1. Domains are the same
2. Tasks are the same

src: [Neural Transfer Learning for Natural Language Processing (PhD thesis)]

Why do we need Domain Adaptation? - Consider this

Dataset	Evaluation	EM	F1
SQuAD	Human Baseline	86.831	89.452
	RoBERTa	86.820	89.795
COVID-QA	Human Baseline	N/A	N/A
	RoBERTa	52.9	27.8
	BioBERT	50.4	27.5
	SciBERT	**53.7**	**28.6**

Comparison b/w the 2 datasets,

Attribute	SQuAD	COVID-QA
Language	English	English
Domain	General/open domain knowledge (wikipedia, etc.)	Biomedical/COVID specific
$\text{Performance}_{\text{RoBERTa}}$	**Near human level!**	**Terrible!**

➡️ **Terrible** w.r.t "acceptable" standards

What could be the issue here?

1. Are TLMs like BERT/RoBERTa **overfitting** on their training datasets?
    - Evidence seems to suggest it (to a certain degree)
2. Do we need to **specifically pretrain a domain specific TLM?**
    - Even then we see barely any improvements as soon as a new dataset comes along?
3. Could it be an architectural flaw?
    - Do we need additional layers/different tokenization schemes/etc.?
4. How different could the *language* of the domains be?
    - Is the language b/w wikipedia & scientific articles **that different**?
5. How much data do TLMs need to see in order to gain *general language understanding*?
    - Basic TLMs like BERT are trained on wikipedia data & wikipedia contains a lot of domain knowledge. So, shouldn't they be acquiring *some knowledge of these concepts*?
6. Problems with Fine-Tuning (FT)?
    - As shown by https://arxiv.org/abs/2006.04884, FT has stability issues. Even with the same seed and hyperparameters, we might observe drastically different scores across runs

Big journeys begin with small steps...

Trying to chase after unsupervised domain adaptation is fine, from the standpoint of scientific progress. However, if we can't even solve the simpler task of supervised DA, why attempt a more difficult task?