Transfer Learning and Domain Adaptation

Sargur N. Srihari
srihari@cedar.buffalo.edu

Topics in Representation Learning

1. Greedy Layer-Wise Unsupervised Pretraining
2. Transfer Learning and Domain Adaptation
3. Semi-supervised Disentangling of Causal Factors
4. Distributed Representation
5. Exponential Gains from depth
6. Providing Clues to Discover Underlying Causes

The scenario

- *Transfer Learning* and *Domain Adaptation* refer to the situation where what has been learned in one setting (i.e., distribution P_1) is exploited to improve generalization in another (say, Distribution P_2)
- It generalizes the idea previously seen in greedy unsupervised pretraining
  - Where we transfer representations between an unsupervised learning task and a supervised learning task

Transfer Learning

- In transfer learning the learner must perform two or more different tasks
- But we assume that the factors that explain the variations in P_1 are relevant to the variations to be captured for learning P_2

Example of transfer learning

- Supervised learning context
  - Input is the same, but target is different
  - There is more data in distribution P_1 and very few in distribution P_2
- Visual classification
  - Significantly more data in distribution sampled from cats and dogs
  - Then learn to quickly generalize to ants and wasps
  - Visual categories share low-level notions of edges and visual shapes, geometric changes, lighting

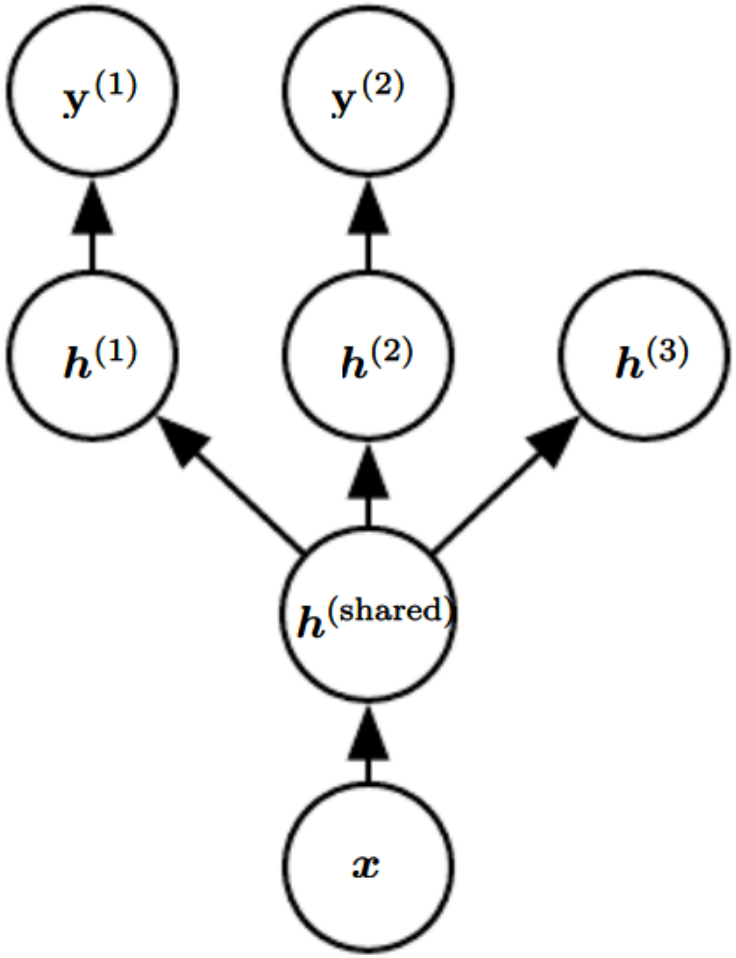
Shared Semantics of Input

- Transfer learning, multi-task learning and domain adaptation are achieved via representation learning
  - Where there exist features that are useful for different tasks or settings
  - This is illustrated next, with shared lower layers and task dependent upper layers

Multi-task learning (shared input)

– Tasks share a common input but involve different target random variables

- Task specific parameters (weights into and from $\boldsymbol{h}^{(1)}$ and $\boldsymbol{h}^{(2)}$ can be learned on top of those yielding $\boldsymbol{h}^{(\text{shared})}$
- In the unsupervised context some top level factors $\boldsymbol{h}^{(3)}$ are associated with none of the tasks

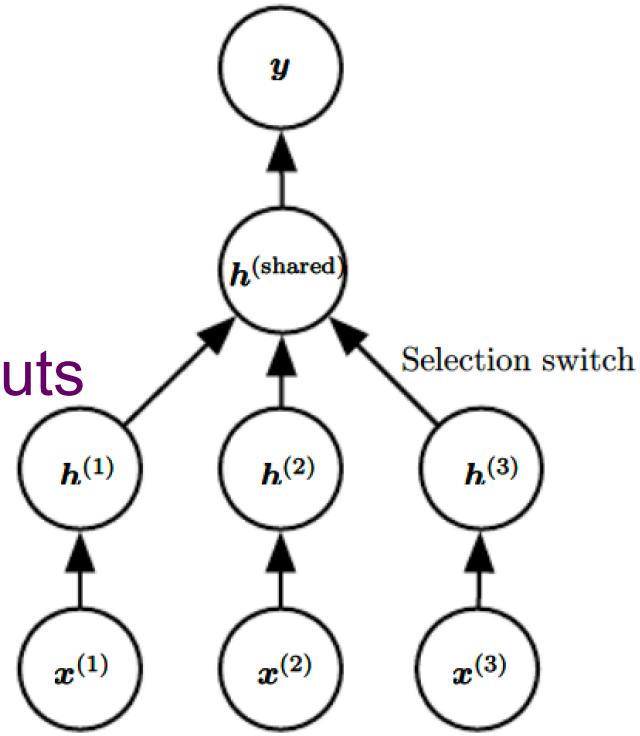


Shared semantics of Output

- What is shared among different tasks is not the semantics of the input but the semantics of the output
- Ex: speech recognition
  - Need to produce valid sentences at the output
  - Earlier layers near input need to recognize very different versions of input phonemes depending on person speaking

Transfer Learning (shared output)

- When output variable y has the same semantics for all classes
  - $\boldsymbol{x}$ has different meaning
  - dimension for each task
    - Three tasks $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$ and $\boldsymbol{x}^{(3)}$ are inputs



- Lower levels upto selection switch are task-specific
  - – Upper levels are shared
    - Semantics of output are shared, not semantics of input

      as in speech recognition where vocalizations are based on different speakers

Success of Transfer Learning

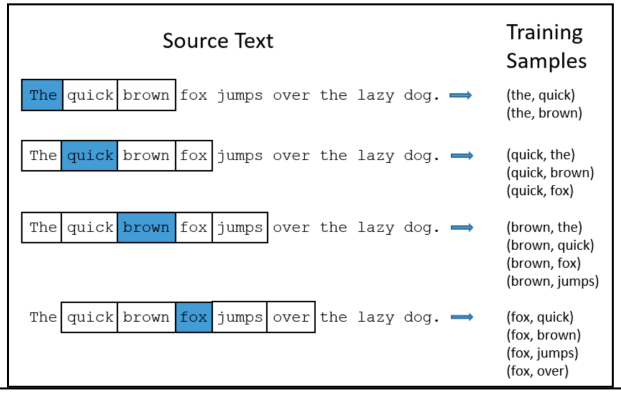
- Unsupervised deep learning for transfer learning has found success in ML competitions
  - Each participant is given data from distribution P_1 illustrating some set of categories
  - Participants learn a feature space
    - Mapping raw input to a representation space
  - This transformation is applied to samples from P_2
  - A linear classifier is trained from very few samples
- As deeper representations used (learned purely unsupervised from P_1) performance improves
  - For deeper representations fewer samples needed

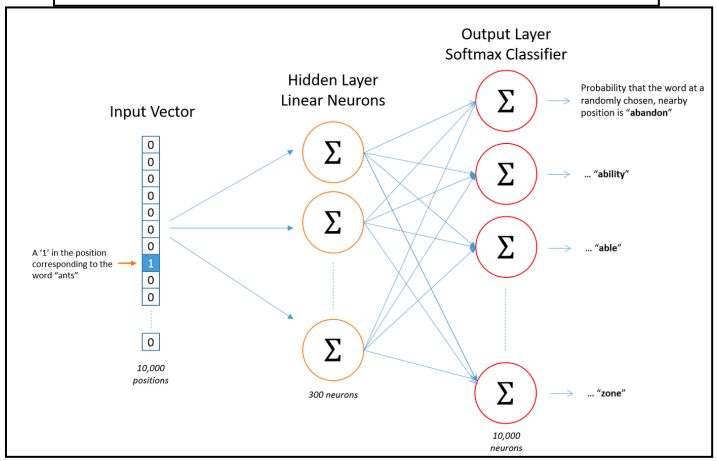
Domain Adaptation

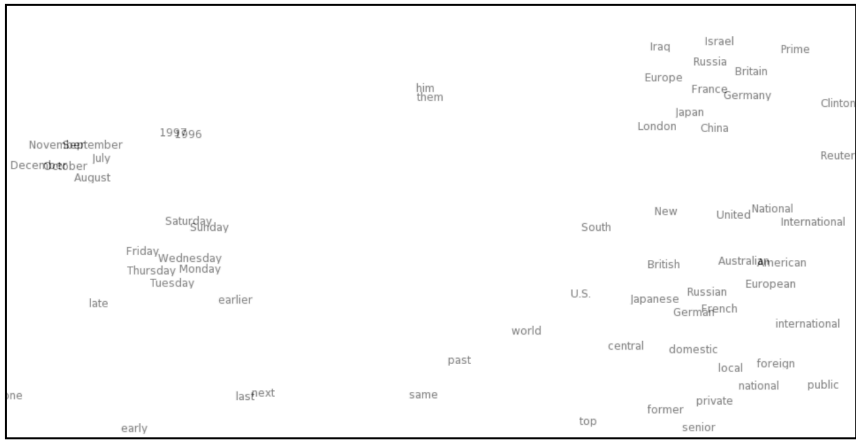
- Related to transfer learning
- Optimal input-to-output mapping remains the same between each setting
- But the input distribution is slightly different
- Ex: In *sentiment analysis*, moving from domain of media (books/music/videos) to domain of consumer electronics (TV/smartphones)

NLP: Word to Vec

- Training Data
- Word-to-vec
  - One-hot vector of words
    - With 30,000 elements
  - mapped to $\boldsymbol{h}$
    - vector of 300
- Word embedding
  - Similar words are close together







Examples of Words-to-vecs

Represent noun by co-occurrences with 25 verbs*

Semantic feature values:
"**celery**"
0.8368, eat
0.3461, taste
0.3153, fill
0.2430, see
0.1145, clean
0.0600, open
0.0586, smell
0.0286, touch
…
…
0.0000, drive
0.0000, wear
0.0000, lift
0.0000, break
0.0000, ride

Semantic feature values:
"**airplane**"
0.8673, ride
0.2891, see
0.2851, say
0.1689, near
0.1228, open
0.0883, hear
0.0771, run
0.0749, lift
…
…
0.0049, smell
0.0010, wear
0.0000, taste
0.0000, rub
0.0000, manipulate

* in a trillion word text collection

Example of Domain Adaptation

- Sentiment Analysis Task
- Determine if comment is positive/negative
  - Sentiment predictor is trained on customer reviews of media content such as books, videos and music
  - Later used to analyze comments about consumer electronics such as televisions and smartphones
  - Vocabulary and style may vary from one domain to other
  - Simple unsupervised pretraining (with denoising autoencoders) found useful with domain adaptation

Concept Drift

- Related to Domain Adaptation is Concept Drift
  - A form of transfer learning where there are gradual changes in data distribution over time
- Both concept drift and transfer learning can be regarded as different forms of multi-task learning
  - Multitask Learning typically refers to supervised learning
  - Transfer Learning is applicable to unsupervised learning and reinforcement learning as well

Success in transfer learning

- Unsupervised deep learning for transfer learning: successful in ML competitions
  - Participant given data set from first setting (from distribution P_1) illustrating examples from some set of categories to learn a good feature space
  - Learned transformation applied to inputs from transfer setting (distribution P_2), a linear classifier trained to generalize well from few labeled samples
- With deeper representations in learning from P_1, learning curve on P_2 becomes much better

Two extreme forms of transfer learning

- One-shot learning
  - Only one labeled sample example of the transfer task is given
- Zero-shot learning
  - No labeled samples are given for the transfer task

One-shot learning

- Only one labeled example of the transfer task
  - Possible because the representation learns to cleanly separate underlying classes during Stage 1
  - During transfer learning, only one labeled example is needed to infer the label of many possible test examples that cluster around the same point in representation space
- Works to the extent that factors of variation corresponding to these invariances have been cleanly separated from the other factors in the learned representation space

Zero-shot learning

- No labeled examples
- Ex: A learner reads a large collection of text and then solves object recognition problems
  - Having read that a cat has four legs and pointed ears, learner guesses that an image is a cat without having seen a cat before

Zero-data learning explained

- Possible because additional data exploited
- Zero-data learning scenario includes three random variables
  1. Traditional inputs $\boldsymbol{x}$
     - Unlabeled text data containing sentences such as “cats have four legs”, “cats have pointy ears’)
  2. Traditional outputs y (y=1 indicating yes, y=0 for no)
  3. Description of task T (represents questions to be answered)
     - Is there a cat in this image?
- Model trained to determine conditional $p(y|\boldsymbol{x}, T)$

Type of Representation of T

- Zero-shot learning requires T to be represented in a way that allows some sort of generalization
  - T cannot be just a one-hot code indicating an object category
  - Instead a distributed representation of object categories by using a learned word embedding for the word associated with each category

Similar phenomenon in Machine Translation

- We have words in one language
  - Word relationships learned from a unilingual corpus
- We have translated sentences that relate words in one language with words in the other
- No labeled word translations available
  - i.e., word A in language $\boldsymbol{X}$ to word B in language $\boldsymbol{Y}$
- Can guess a translation for word A because
  - We have learned distributed representations for words in $\boldsymbol{X}$ and for words in $\boldsymbol{Y}$ then created a link relating the two spaces via training examples of matched pairs of sentences
    - Works best when two representations and relations are learned jointly

Transfer learning enables zero-shot

- Labeled or unlabeled examples of $\boldsymbol{x}$ allow:
  - Learning a representation function f_x and similarly with examples of $\boldsymbol{y}$ to learn f_y
    - Each application of f_x and f_y appears as an upward arrow
    - Distances in h_x and h_y space provide a similarity metric
    - Image $\boldsymbol{x}_{\text{test}}$ is associated with word $\boldsymbol{y}_{\text{test}}$ even if no image of that word was ever presented