

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

Topics in Transfer Learning

- Definition of Transfer Learning
- Transfer learning in NLP and Computer Vision
- Shared semantics of input/output
- Domain Adaptation/Concept Drift
- One-shot/zero-shot learning
- Transfer learning and Generalization

Definition of Transfer Learning

- Also known as *Domain Adaptation*
- Learning in one setting (i.e., distribution P_1) is used to improve generalization in another (distribution P_2)
- Generalizes *greedy unsupervised pretraining*
 - Where we transfer representations between an unsupervised task and a supervised task

Simple example of transfer learning

- Visual classification
 - Significantly more data in distribution sampled from cats and dogs



- Then learn to quickly generalize to ants and wasps

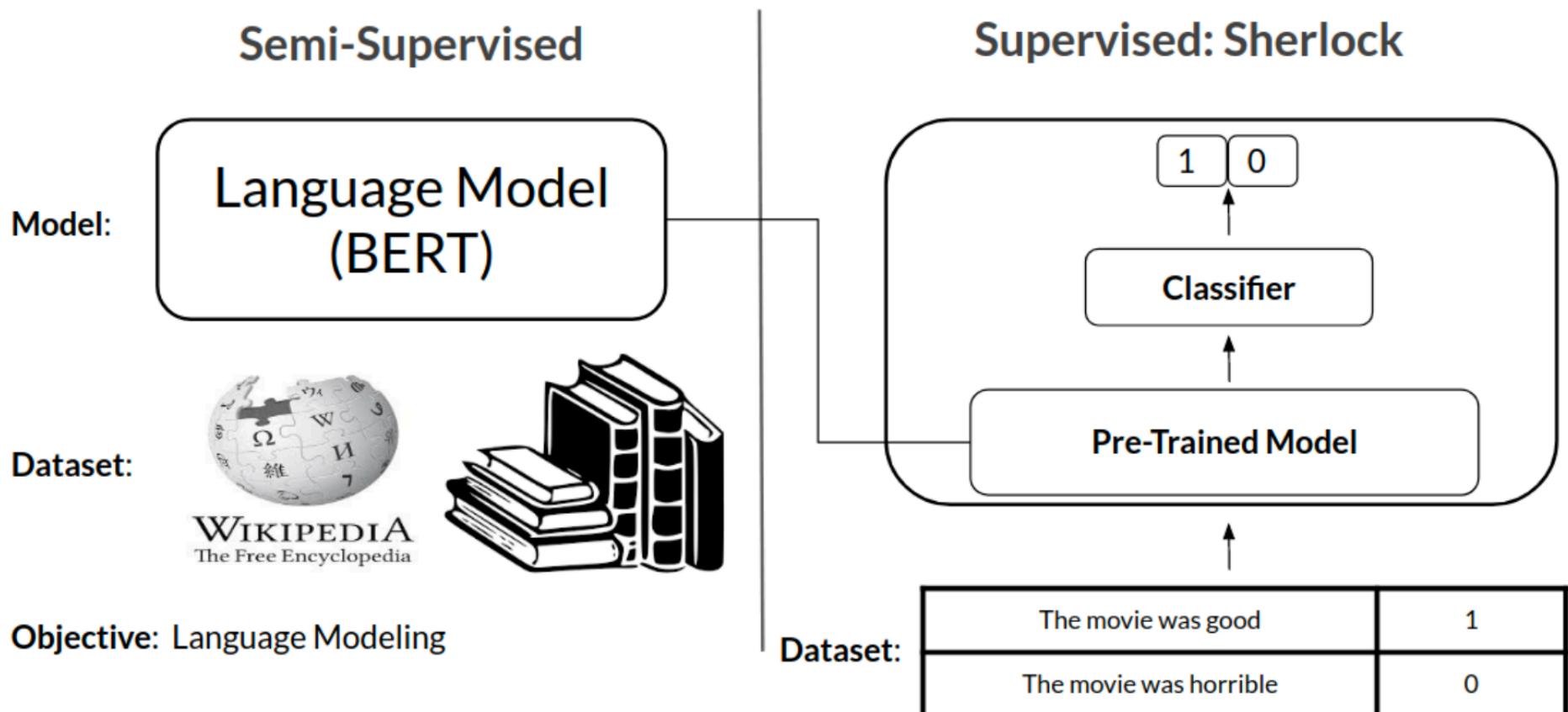


- Visual categories share
 - Low-level notions of edges and visual shapes,
 - Effects of geometric changes
 - Changes in lighting

Practical Need for Transfer Learning

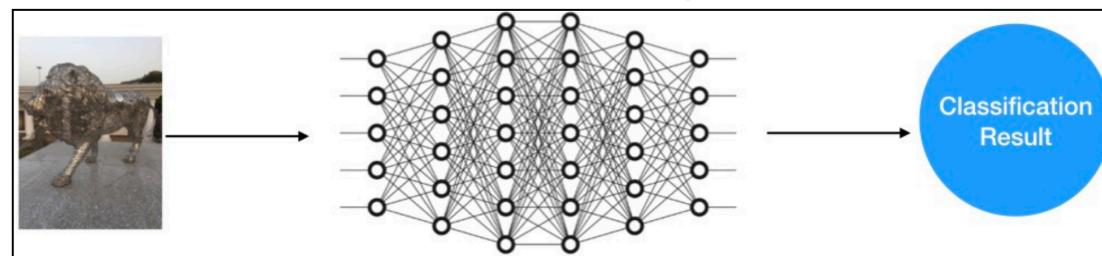
- Training a deep learning model requires
 - a lot of data and
 - more importantly a lot of time
- Useful to take advantage of pre-trained weights on huge datasets that took days/weeks to train, and leverage it towards our use case
- Depending on how much data we have at hand, here are the different ways to leverage this

Transfer Learning in NLP

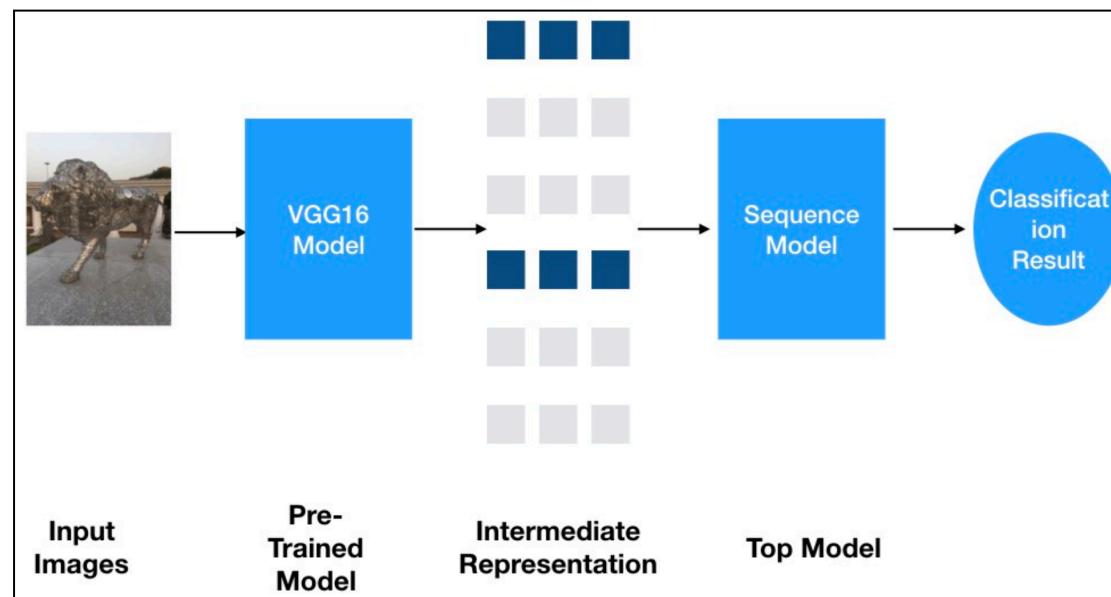


Transfer Learning in Computer Vision

Direct Learning



Transfer Learning



Transfer Learning

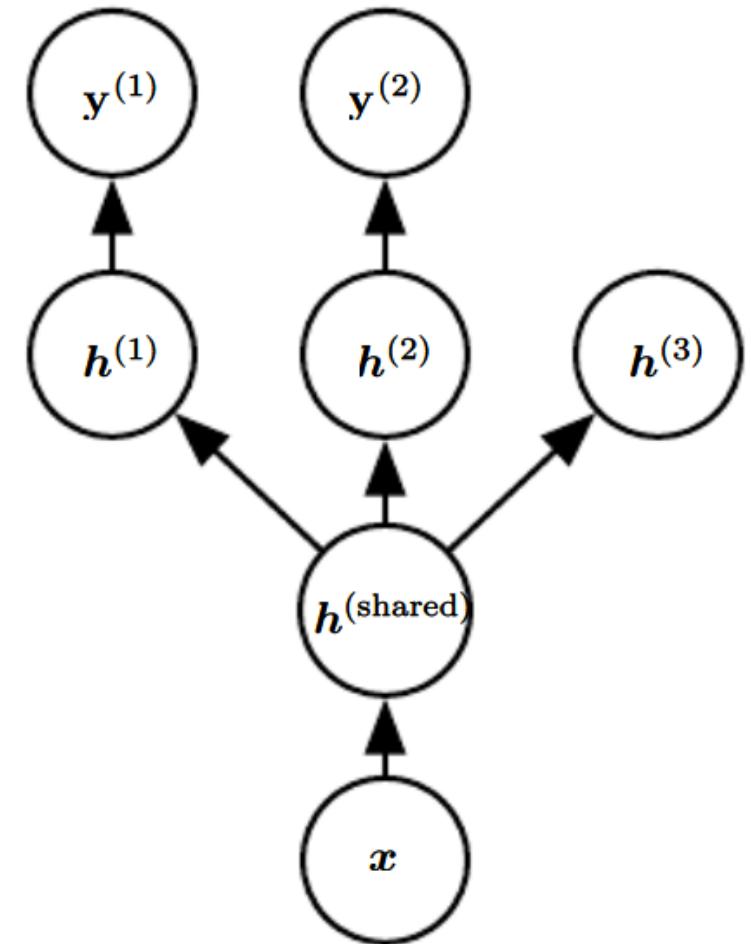
- In transfer learning the learner must perform two or more different tasks
- But we assume that the factors that explain the variations in P_1 are relevant to the variations to be captured for learning P_2
- Supervised learning context
 - Input is the same, but target is different
 - There is more data in distribution P_1 and very few in distribution P_2

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
 - Corresponding to underlying factors that appear in more than one setting
 - 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 $h^{(1)}$ and $h^{(2)}$ can be learned on top of those yielding $h^{(\text{shared})}$)
 - In the unsupervised context some top level factors $h^{(3)}$ are associated with none of the tasks

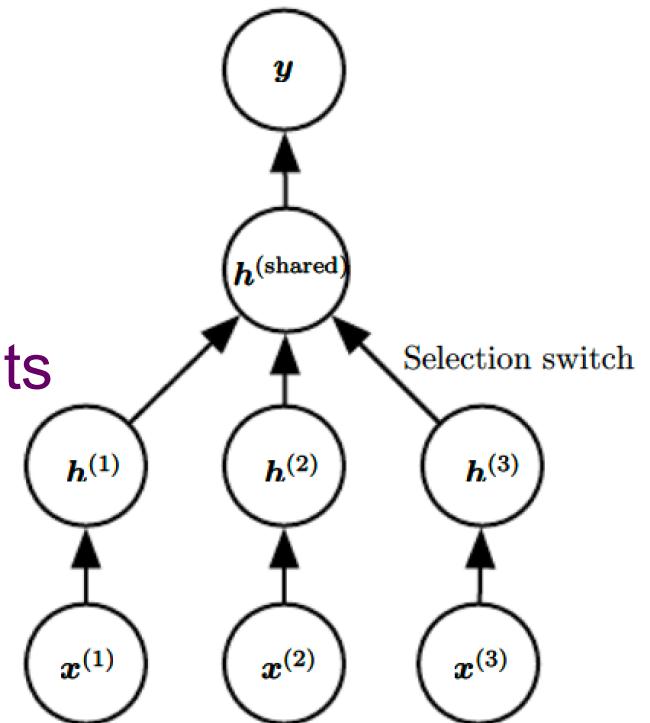


Shared semantics of Output

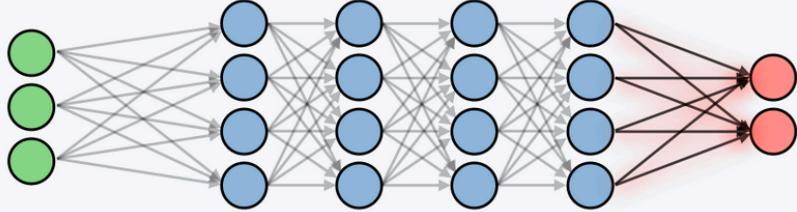
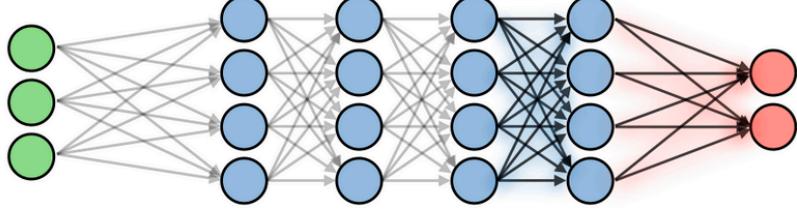
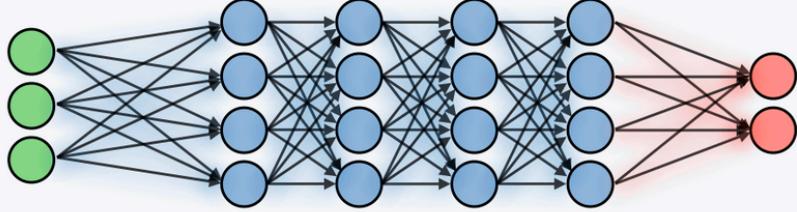
- Sometimes what is shared among different tasks is not the semantics of the input but the semantics of the output
- Ex: speech recognition
 - Need to produce valid sentences at the output
 - Earlier layers near input need to recognize very different versions of input phonemes depending on person speaking

Transfer Learning (shared output)

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



Ways of leveraging Transfer Learning

Training size	Illustration	Explanation
Small		Freezes all layers, trains weights on softmax
Medium		Freezes most layers, trains weights on last layers and softmax
Large		Trains weights on layers and softmax by initializing weights on pre-trained ones

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)

Domain Adaptation: Sentiment Analysis

- Determine if comment is positive/negative
 - Sentiment predictor is trained on
 - Customer reviews of media: books, videos and music
 - Used for comments on consumer electronics
 - Televisions and smartphones
- Vocabulary and style varies between domains
- 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

- Single labeled example of the transfer task
 - Possible because representation learnt separates classes during Stage 1
 - During Stage 2, only one labeled example needed to infer label of test examples that cluster around same point
- Works to extent that factors of variation corresponding to invariances have been separated from other factors in 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 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 $p(y|x, T)$

Type of Representation of T

- Zero-shot requires T 's representation allows generalization
 - T cannot be just a one-hot code indicating 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 X to word B in language Y
- Can guess a translation for word A because
 - We have learned representations for words in X and for words in Y then created a link relating the two spaces via training examples of matched pairs of sentences
 - Best when two representations and relations are learned jointly

Zero-shot as a form of transfer learning

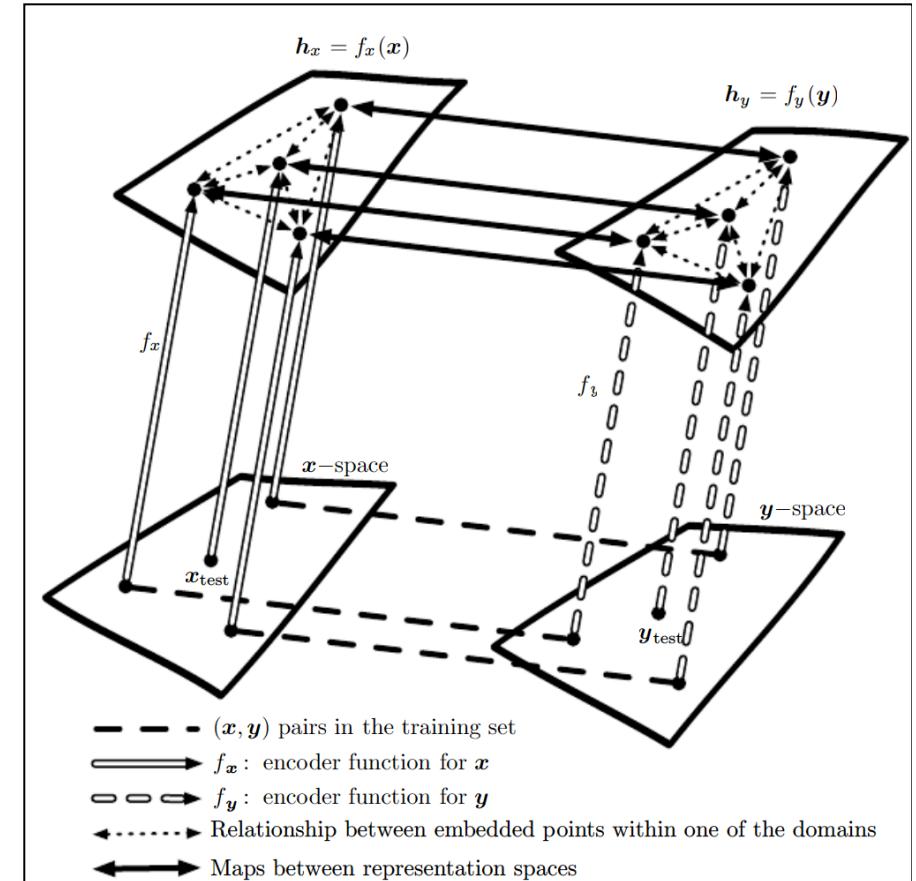
- Same principle explains multi-modal learning
 - capturing a representation in one modality, a representation in the other, and the relationship (in general a joint distribution) between pairs (x, y) consisting of one observation x in one modality and another observation y in the other modality
- By learning all three sets of parameters
 - (x to its representation, y to its representation, and relationship between the two representations),
 - concepts in one representation are anchored in the other, and vice-versa, allowing one to meaningfully generalize to new pairs

Transfer learning enables zero-shot

Labeled/unlabeled examples of x allow learning a representation function f_x

Also, with examples of y to learn f_y

- Each application of f_x and f_y appears as upward arrows
- Distances in h_x and h_y space provide a similarity metric
- Image x_{test} is associated with word y_{test} even if no image of that word was ever presented



Transfer Learning and Generalization

- Current ML methods are weak to generalize beyond the training distribution— which is what is needed in practice
- What has been learned in one setting needs to generalize well in other related distributions

Fast transfer learning

- Previous learning is a rich base to adapt to a new but related distribution
- Some new concept may have to be learned but because most of the other relevant concepts have already been captured, learning can be very fast on the transfer distribution

Assumption of changes

- Changes are sparse when the knowledge is represented in an appropriately modularized way, with only one or a few of the modules having changed
- Distributional change is due to actions by one or more agents, where a single causal variable is clamped to a particular value

Need for inferring causal structure

- A possible change to the joint distribution of the variables of interest due to an intervention, even one that has never been observed before
- Insufficient to learn joint distribution of observed variables
- Should also learn enough about the underlying high-level variables and their causal relations to be able to properly infer the effect of an intervention

Cause, Effect, Intervention

- Causal relations are inferred by intervention
 - A = Raining causes B =Open Umbrella (not vice-versa)
 - Changing probability of raining (because weather changed) doesn't change relation between A and B (captured by $P(B/A)$), but has an impact on $P(B)$
 - Conversely, an agent's intervention on B (Open umbrella) will have no effect on $P(A)$
- Asymmetry not visible from $(A;B)$ training pairs
 - Until an intervention
 - Motivates setup where one learns from a set of distributions arising from not necessarily known interventions, not simply to capture a joint distribution but to discover some underlying causal structure

Advantage of Correct Causal Model

- A and B are discrete with N possible values
- Consider two parametrizations ($A \rightarrow B$ model and $B \rightarrow A$ model) to estimate their joint:
 - $P_{A \rightarrow B}(A;B) = P_{A \rightarrow B}(A)P_{A \rightarrow B}(B | A)$
 - $P_{B \rightarrow A}(A;B) = P_{B \rightarrow A}(B)P_{B \rightarrow A}(A | B)$
 - Both models have $O(N^2)$ parameters
 - Maximum likelihood estimation leads to indistinguishable test set performance (where the test set is sampled from the training distribution)

Comparing ($A \rightarrow B$ vs $B \rightarrow A$)

- How fast models adapt on a transfer distribution after being trained on training distribution
 - Simple SGD is used
- Let $A \rightarrow B$ be the correct causal model.
- Consider change between the two distributions is a random change in parameters of the true $P(A)$ for the cause A (because this will have an impact on the effect B , which can be picked up and reveal the causal direction)

Adaptation to transfer distribution

As more transfer distribution examples are seen by the learner (horizontal axis), in terms of the log-likelihood on the transfer distribution (on a large test set from the transfer distribution, tested after each update of the parameters)

$N = 10$. Curves are the median over 10,000 runs.

Correct causal model adapts faster. Most informative part of trajectory is first 10-20 samples

