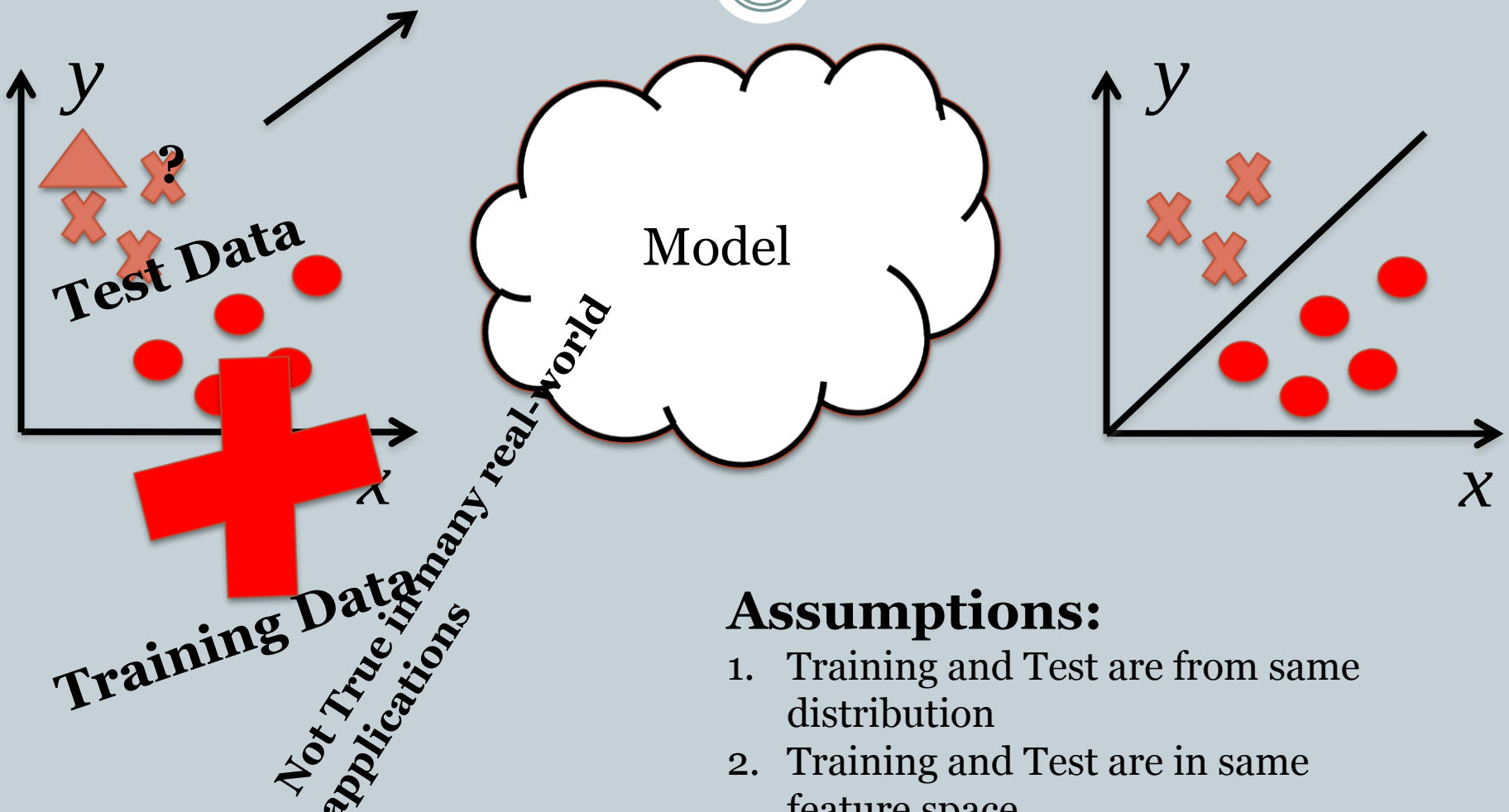
How Transferable are Neural Networks in NLP Applications

DAI YONG

SMILE LAB

Motivation



Assumptions:

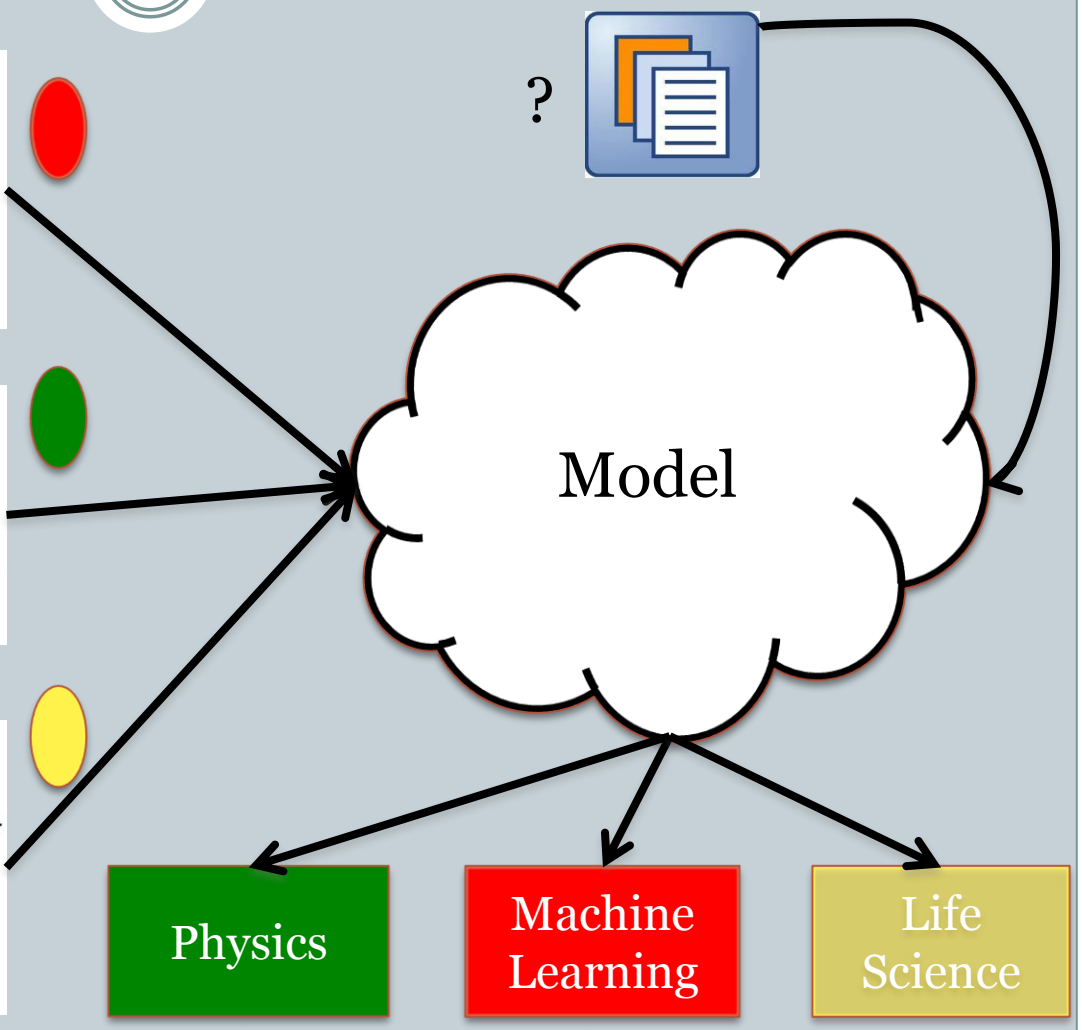
1. Training and Test are from same distribution
2. Training and Test are in same feature space

Examples: Web-document Classification

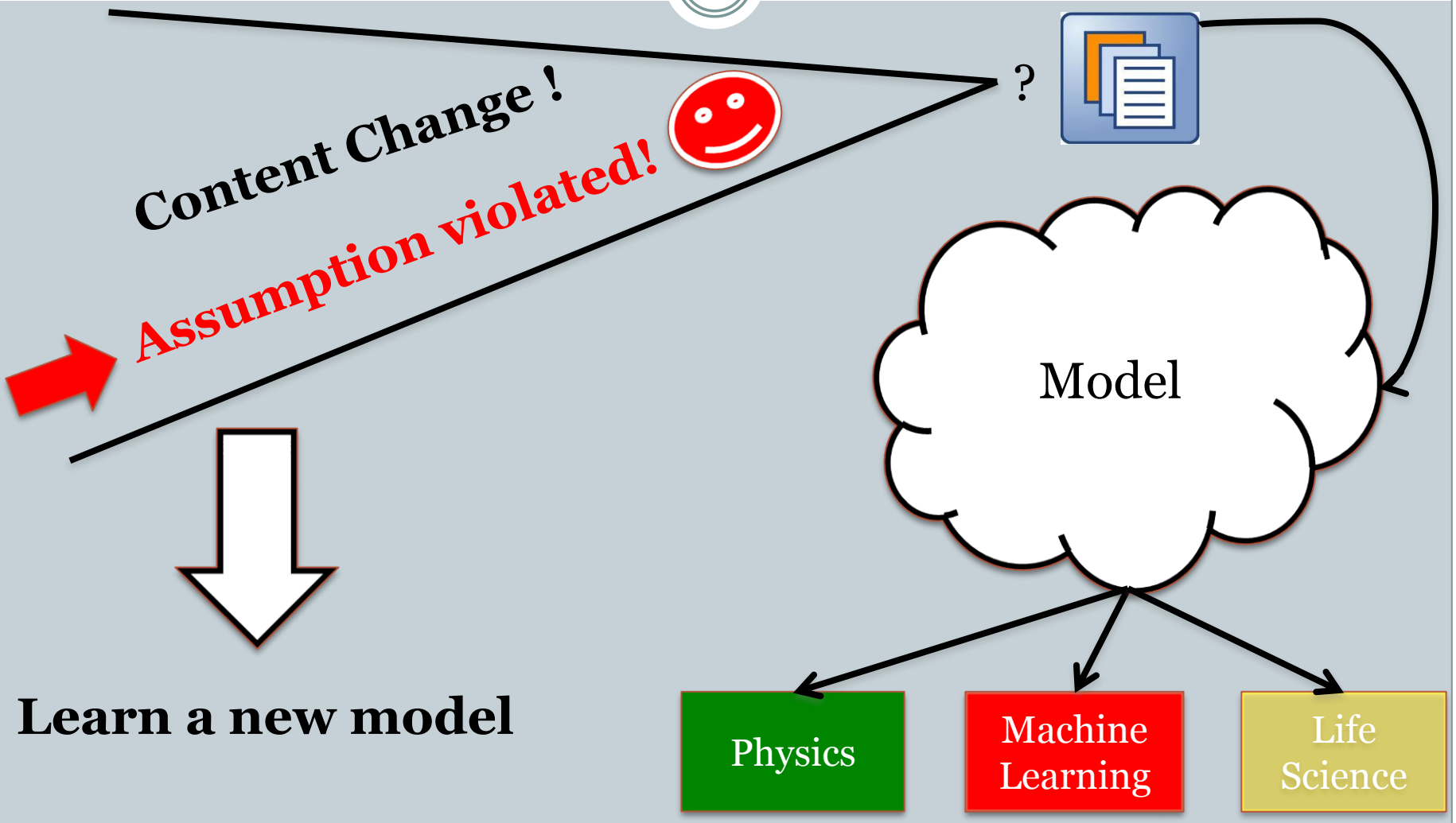
Many security applications, e.g. for access control, use face recognition as one of its components. That is, given the photo (or video recording) of a person, recognize who this person is. In other words, the system needs to **classify** the faces into one of many categories (Alice, Bob, Charlie, ...) or decide that it is an unknown face. A similar, yet conceptually quite different problem is that of verification. Here the goal is to verify whether the person in question is who he claims to be. Note that differently to before, this is now a yes/no question. To deal with different lighting conditions, facial expressions, whether a person is wearing glasses, hairstyle, etc., it is desirable to have a system which *learns* which features are relevant for identifying a person.

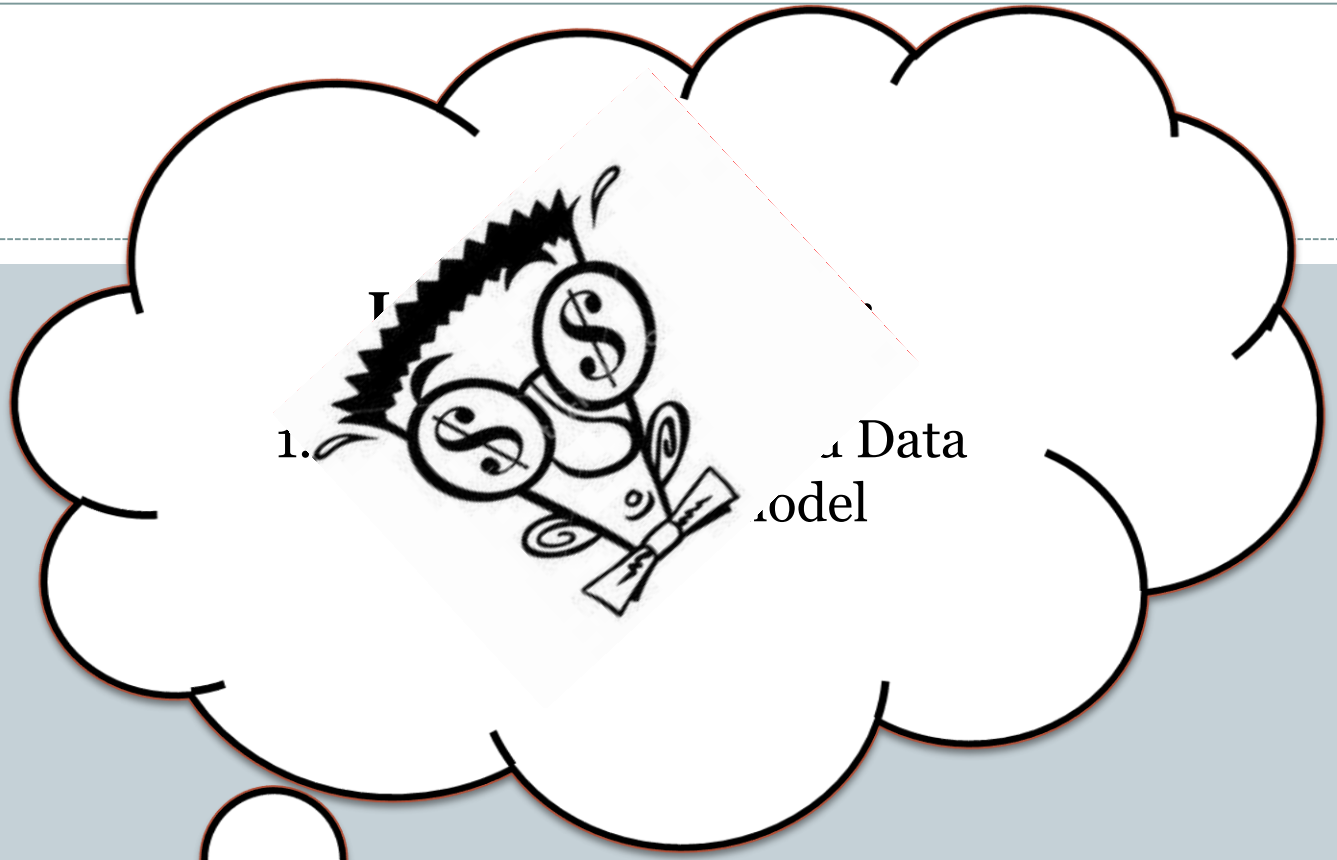
Quantum Interpretation: Let us change the way of looking at this problem and thereby turn it into a quantum mechanical experiment. You have heard at various points in your physics course that light comes in little quanta known as photons. The first time this assumption had been made was by Planck in 1900 'as an act of desperation' to be

Darwin gathered data and honed his theory for 20 years before publishing his well-known book in 1859, *The Origin of Species by Means of Natural Selection, or The Preservation of Favoured Races in the Struggle for Life*. Darwin and his fellow naturalist Alfred Wallace independently came to the conclusion that geologically older species of life gave rise to geologically younger and different species through the process of natural selection.



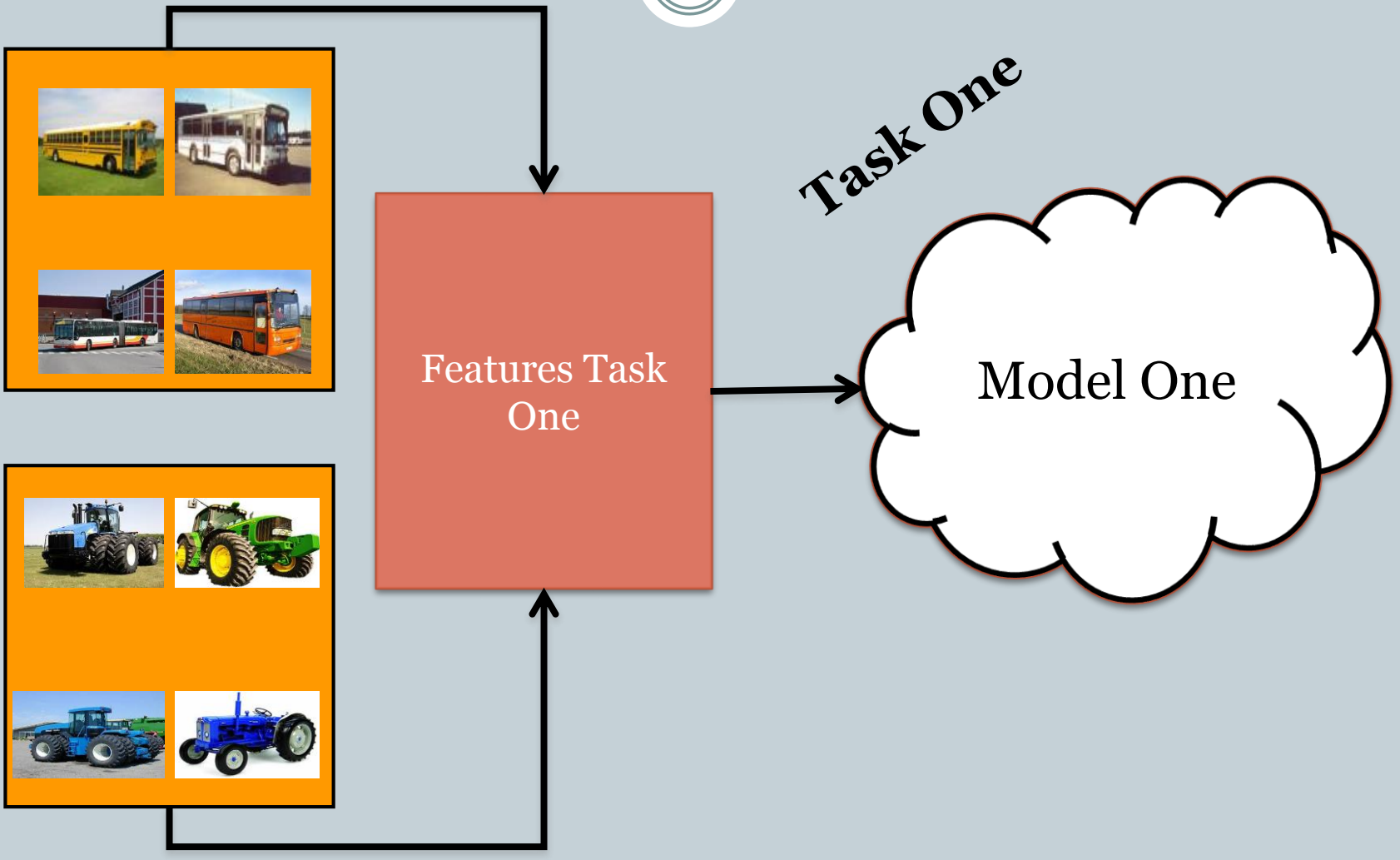
Examples: Web-document Classification



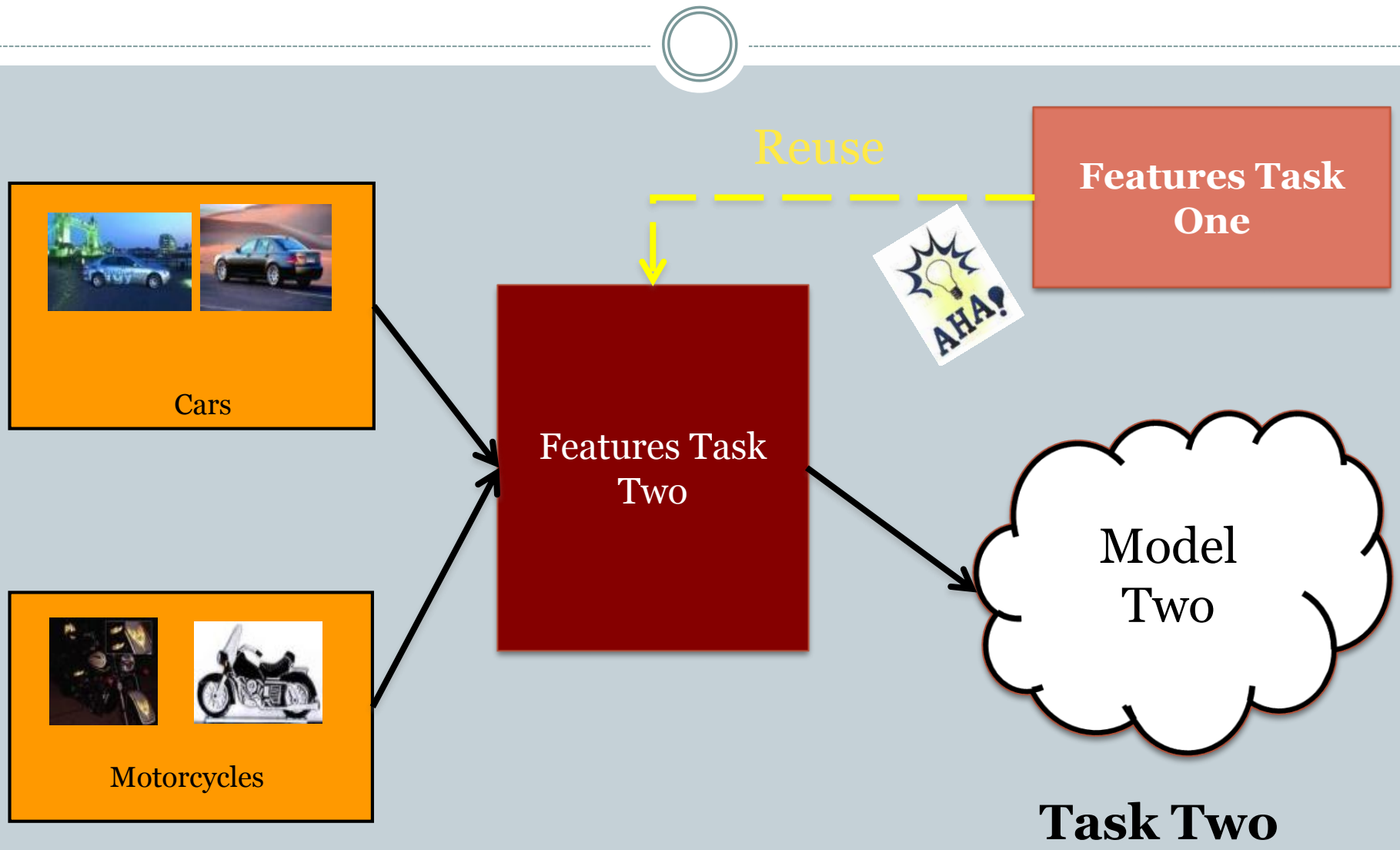


Reuse & Adapt already learned model !

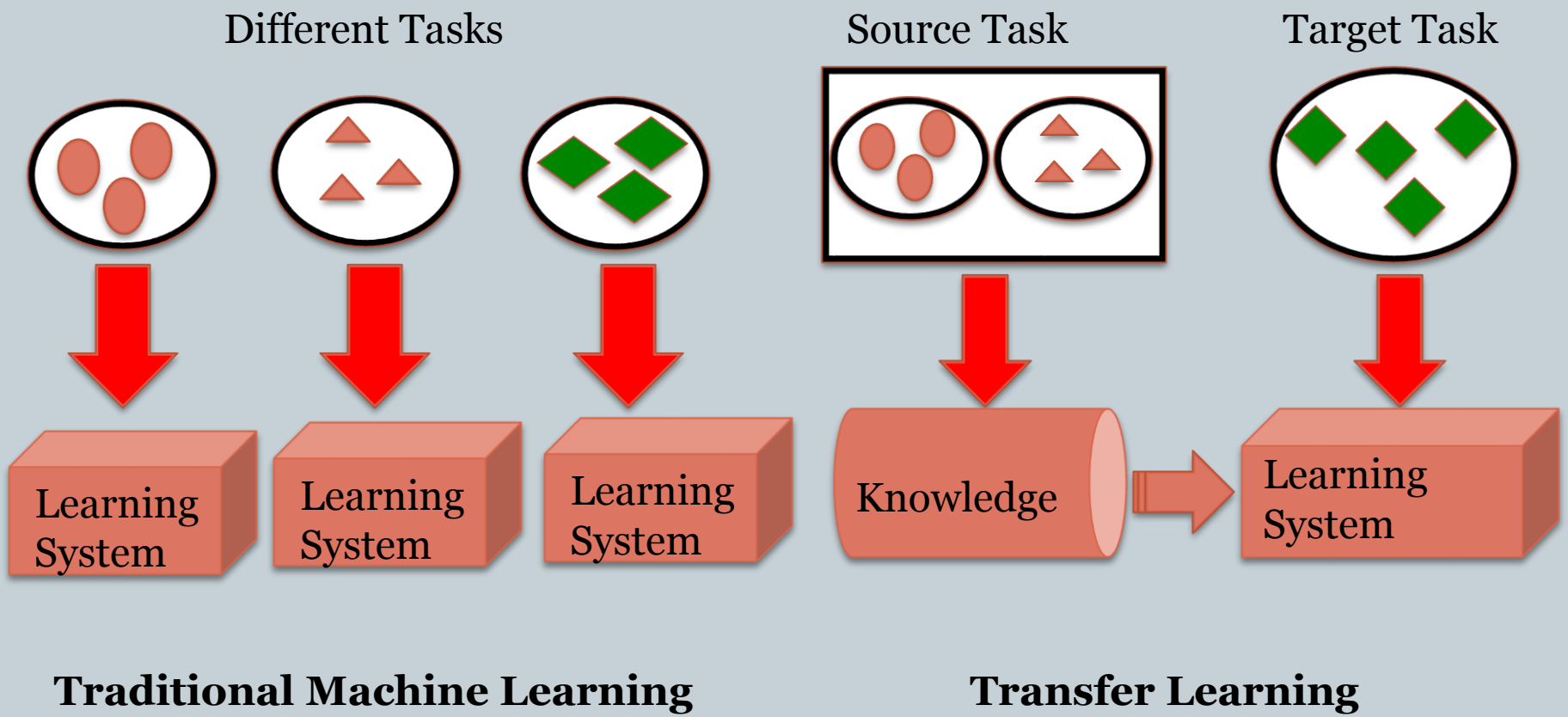
Examples: Image Classification



Examples: Image Classification



Traditional Machine Learning vs. Transfer



Other motivations

- In some domains, labeled data are in short supply.
- In some domains, the calibration effort is very expensive.
- In some domains, the learning process is time consuming.

- *How to extract knowledge learnt from related domains to help learning in a target domain with a few labeled data?*
- *How to extract knowledge learnt from related domains to speed up learning in a target domain?*

Transfer learning techniques may help!

Transfer Learning Definition

- Notation:
  - Domain $\mathcal{D}$:
    - Feature Space: $\mathcal{X}$
    - Marginal Probability Distribution: $P(X)$
      - with $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$
  - Given a domain then a task is :

$$\mathcal{T} = \{\mathcal{Y}, f(.)\}$$

Label Space

P(Y|X)

Transfer Learning Definition

Given a source domain and source learning task, a target domain and a target learning task, transfer learning aims to help improve the learning of the target predictive function using the source knowledge, where

$$\mathcal{D}_S \neq \mathcal{D}_T \quad \text{or} \quad \mathcal{T}_S \neq \mathcal{T}_T$$

Transfer Definition

- Therefore, if either :

Domain Differences

$$\mathcal{X}_S \neq \mathcal{X}_T \qquad \mathcal{P}_S(X) \neq \mathcal{P}_T(X)$$

Task Differences

$$\mathcal{Y}_S \neq \mathcal{Y}_T \qquad P(Y_S|X_S) \neq P(Y_T|X_T)$$

Examples: Cancer Data

$P_S(X)$

Age Smoking

$$\mathcal{X}_S = \{x_1^S, x_2^S\}$$

$$\mathcal{X}_S \neq \mathcal{X}_T$$

$$P_S(X) \neq P_T(X)$$

$P_T(X)$

Age Height Smoking

$$\mathcal{X}_T = \{x_1^T, x_2^T, x_3^T\}$$

Examples: Cancer Data

$P_S(X)$

$\mathcal{X}_S = \{x_1^S, x_2^S\}$

Task Source: Classify into cancer or no cancer

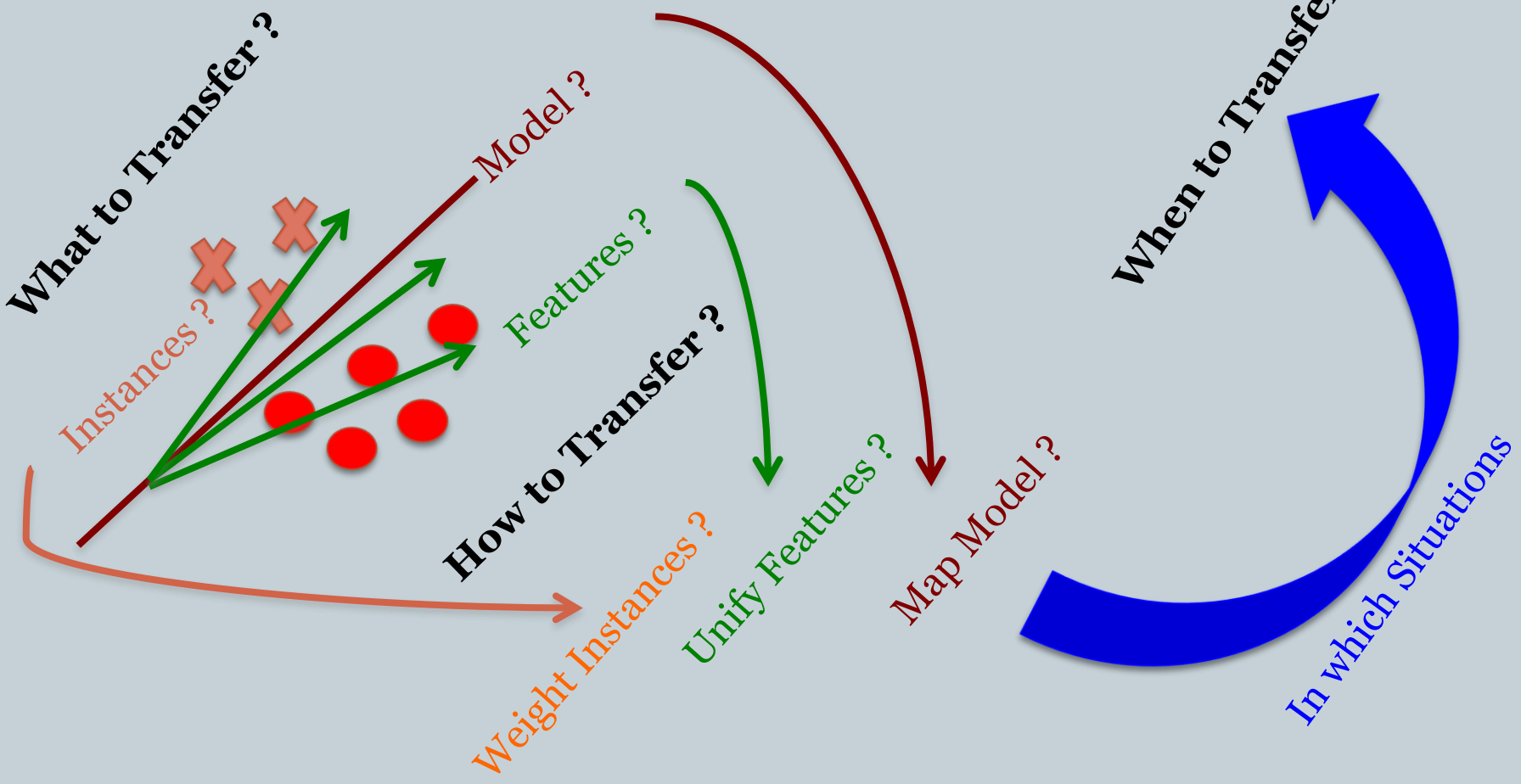
$\mathcal{Y}_S \neq \mathcal{Y}_T$

$P_T(X)$

$\mathcal{X}_T = \{x_1^T, x_2^T, x_3^T\}$

Task Target: Classify into cancer level one, cancer level two, cancer level three

Questions to answer when transferring



How Transferable are Neural Networks in NLP Applications 2016 EMNLP

This paper mainly focuses on the following research questions:

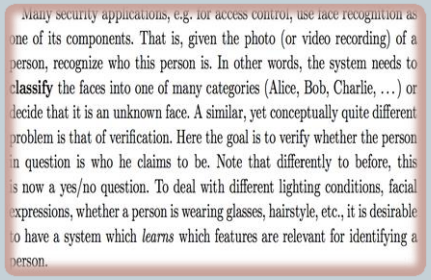
Q1: How transferable are neural networks between two tasks with similar or different semantics in NLP applications?

Q2: How transferable are different layers of NLP neural models?

Q3: How transferable are INIT and MULT, respectively? What is the effect of combining these two methods?

Difficulties in NLP

Image processing


Many security applications, e.g. for access control, use face recognition as one of its components. That is, given the photo (or video recording) of a person, recognize who this person is. In other words, the system needs to **classify** the faces into one of many categories (Alice, Bob, Charlie, ...) or decide that it is an unknown face. A similar, yet conceptually quite different problem is that of verification. Here the goal is to verify whether the person in question is who he claims to be. Note that differently to before, this is now a yes/no question. To deal with different lighting conditions, facial expressions, whether a person is wearing glasses, hairstyle, etc., it is desirable to have a system which *learns* which features are relevant for identifying a person.

Natural language processing

- Pixels are low-level signals
- Continuous
- Less related to semantics

Less clear

- Tokens are discrete
- Each word well reflects the thought of humans
- Neighboring words do not share as much information as pixels in images do

Experiment settings

Two scenarios

(1)transferring knowledge to a semantically similar/equivalent task but with a different dataset; (2)transferring knowledge to a task that is semantically different but shares the same neural topology/architecture so that neural parameters can indeed be transferred.

Two methods

(1) using the parameters trained on *S* to initialize *T* (INIT), and (2) multi-task learning (MULT)

Datasets

Experiment I: Sentence classification	Experiment II: Sentence-pair classification
- **IMDB**. A large dataset for binary sentiment classification (positive vs. negative) - **MR**. A small dataset for binary sentiment classification. - **QC**. A (small) dataset for 6-way question classification (e.g., location, time, and number).	- **SNLI**. A large dataset for sentence entailment recognition. The classification objectives are entailment, contradiction, and neutral. - **SICK**. A small dataset with exactly the same classification objective as SNLI. - **MSRP**. A (small) dataset for paraphrase detection. The objective is binary classification : judging whether two sentences have the same meaning.

Statistics and examples of the datasets

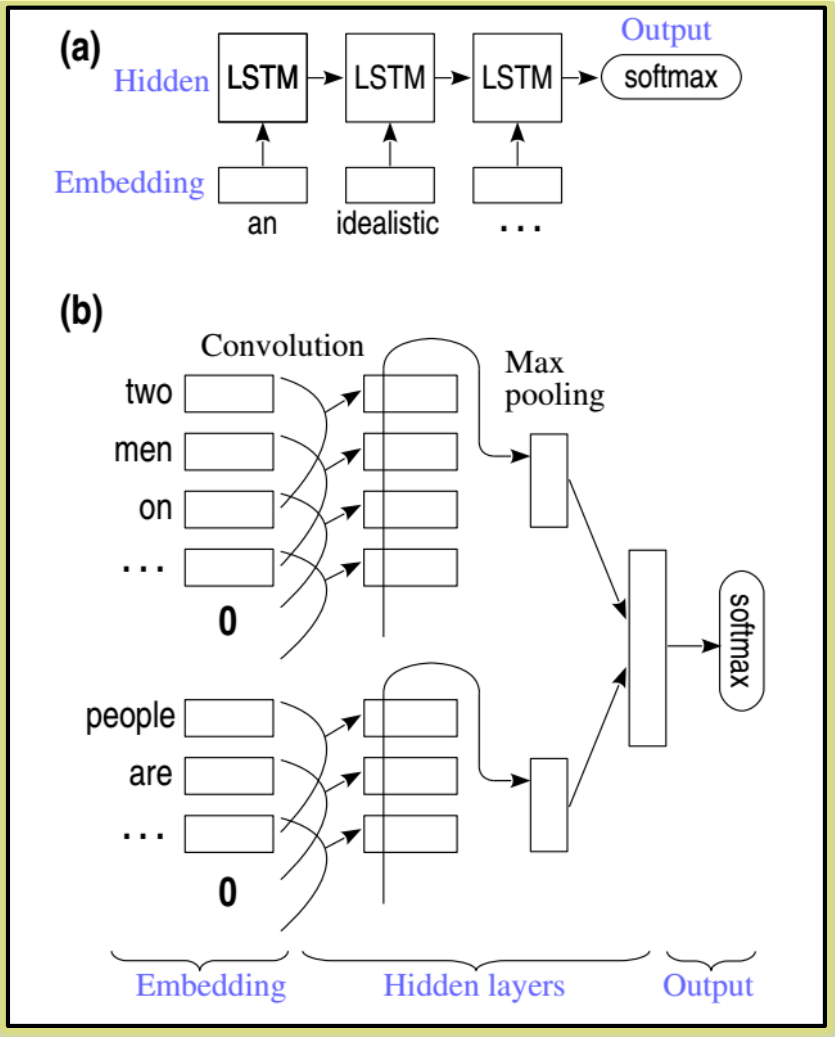
Statistics (# of Samples)						
	Experiment I			Experiment II		
	IMDB	MR	QC	SNLI	SICK	MSRP
#Train	550,000	8,500	4,800	550,152	4,439	3,575
#Val	50,000	1,100	600	10,000	495	501
#Test	2,000	1,100	500	10,000	4,906	1,725

Examples in Experiment I	
Sentiment Analysis (IMDB and MR)	
An idealistic love story that brings out the latent 15-year-old romantic in everyone.	+
Its mysteries are transparently obvious, and its too slowly paced to be a thriller.	−
Question Classification (QC)	
What is the temperature at the center of the earth?	number
What state did the Battle of Bighorn take place in?	location

Examples in Experiment II		
Natural Language Inference (SNLI and SICK)		
Premise	Two men on bicycles competing in a race.	
Hypothesis	People are riding bikes.	E
	Men are riding bicycles on the streets.	C
	A few people are catching fish.	N
Paraphrase Detection (MSRP)		
The DVD-CCA then appealed to the state Supreme Court.		Paraphrase
The DVD CCA appealed that decision to the U.S. Supreme Court.		
Earnings per share from recurring operations will be 13 cents to 14 cents.		Non-Paraphrase
That beat the company's April earnings forecast of 8 to 9 cents a share.		

Two scenarios of transfer regarding semantic similarity:
(1) Semantically equivalent transfer (IMDB to MR, SNLI to SICK)
(2) semantically different transfer (IMDB to QC, SNLI to MSRP)

Neural Models and Settings



(a) Experiment I: RNNs with LSTM units for sentence classification. (b) Experiment II: CNN for sentence pair modeling.

(b) a convolutional neural network (CNN, Figure b) with a window size of 5 to model local context, and a max pooling layer gathers information to a fixed-size vector. Then the sentence vectors are concatenated and fed to a hidden layer before the softmax output.

Transfer Methods

Embedding space

Parameter space

INIT : Parameter initialization	**MULT** : Multi-task learning	**MULT+INIT : Combination**
first trains the network on S, and then directly uses the tuned parameters to initialize the network for T - Fix the parameters in the target domain - fine-tune the parameters if labeled data are available in T	trains samples in both domains $J = \lambda J_T + (1-\lambda) J_S$	first pretrain on the source domain S for parameter initialization, and then train S and T simultaneously

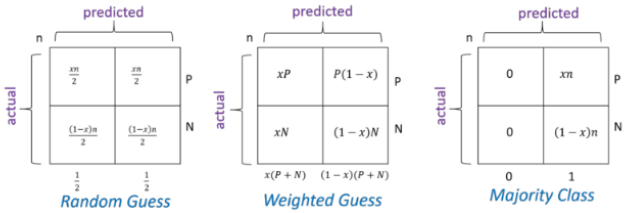
Results of Transferring by INIT

Experiment I		
Setting	IMDB→MR	IMDB→QC
Majority	50.0	22.9
E⊠ H□ O□	75.1	90.8
E🔒 H□ O□	78.2	93.2
E🔒 H🔒 O□	78.8	55.6
E🔒 H🔒 O🔒	73.6	–
E🔓 H□ O□	78.3	92.6
E🔓 H🔓 O□	81.4	90.4
E🔓 H🔓 O🔓	80.9	–
Experiment II		
Setting	SNLI→SICK	SNLI→MSRP
Majority	56.9	66.5
E⊠ H□ O□	70.9	69.0
E🔒 H□ O□	69.3	68.1
E🔒 H🔒 O□	70.0	66.4
E🔒 H🔒 O🔒	43.1	–
E🔓 H□ O□	71.0	69.9
E🔓 H🔓 O□	76.3	68.8
E🔓 H🔓 O🔓	77.6	–

We can define some simple non-machine learning classifiers that assign labels based simply on the proportions found in the training data:

- **Random Guess Classifier** : randomly assign half of the labels to P and the other half as N.
- **Weighted Guess Classifier** : randomly assign x% of the labels to P, and the remaining (1 − x)% to N
- **Majority Class Classifier** : assign all of the labels to N (the majority class in the data)

The confusion matrices for these trivial classifiers would look like:



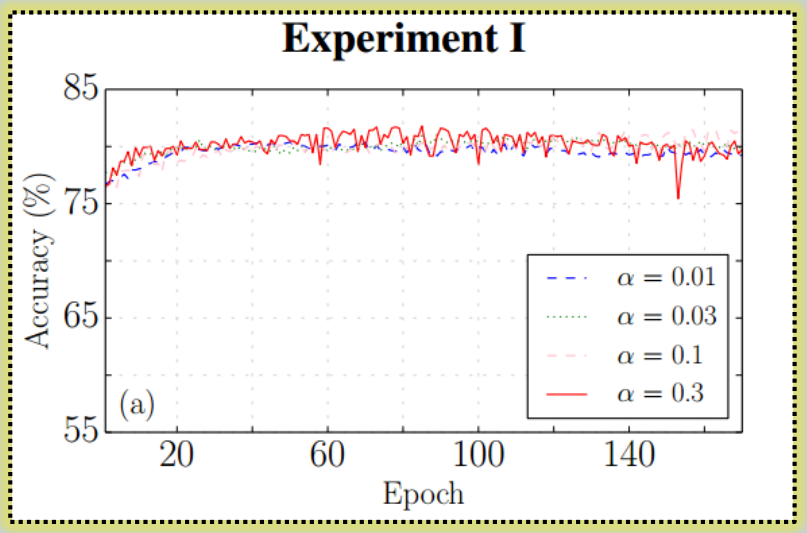
Link

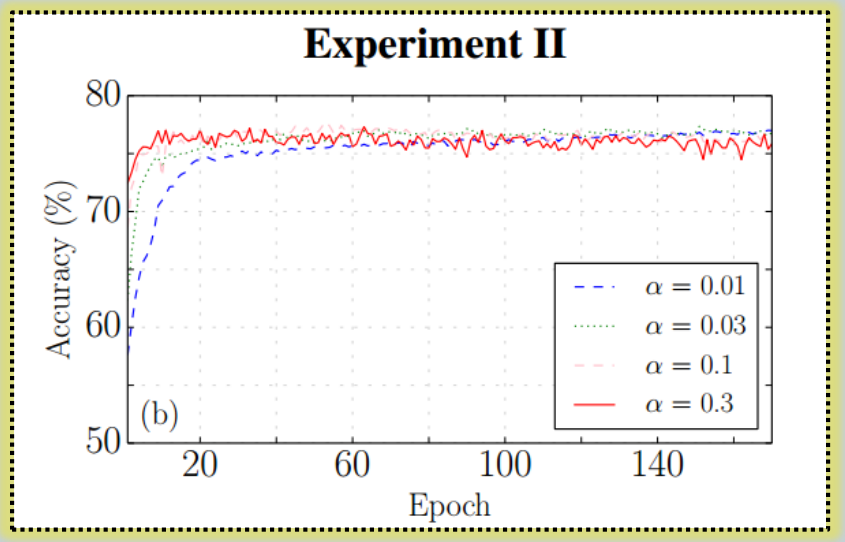
Table 3: Main results of neural transfer learning by INIT. We report test accuracies (%) in this table. E: embedding layer; H: hidden layers; O: output layer. ⊠: Word embeddings are pretrained by `word2vec`; □: Parameters are randomly initialized); 🔒: Parameters are transferred but frozen; 🔓: Parameters are transferred and fine-tuned. Notice that the E🔒H🔒O🔒 and E🔓H🔓O🔓 settings are inapplicable to IMDB→QC and SNLI→MSRP, because the output targets do not share same meanings and numbers of target classes.

Layer-by-Layer Analysis

- The output layer is mainly specific to a dataset. Transferring the output layer's parameters yields little (if any) gain.
- For semantically similar tasks, both of embeddings and the hidden layer play an important role
- For semantically different tasks (IMDB *to* QC and SNLI *to* MSRP), the embeddings are the only parameters that have been observed to be transferable
- Neural networks may not be transferable to NLP tasks of different semantics

How does learning rate affect transfer





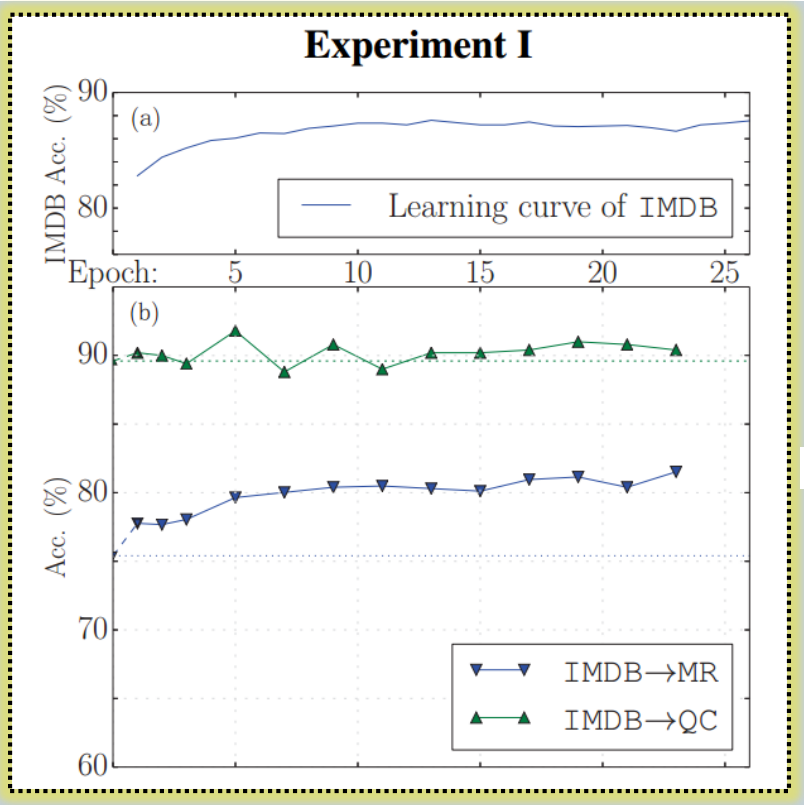
(a) Experiment I: IMDB *to* MR

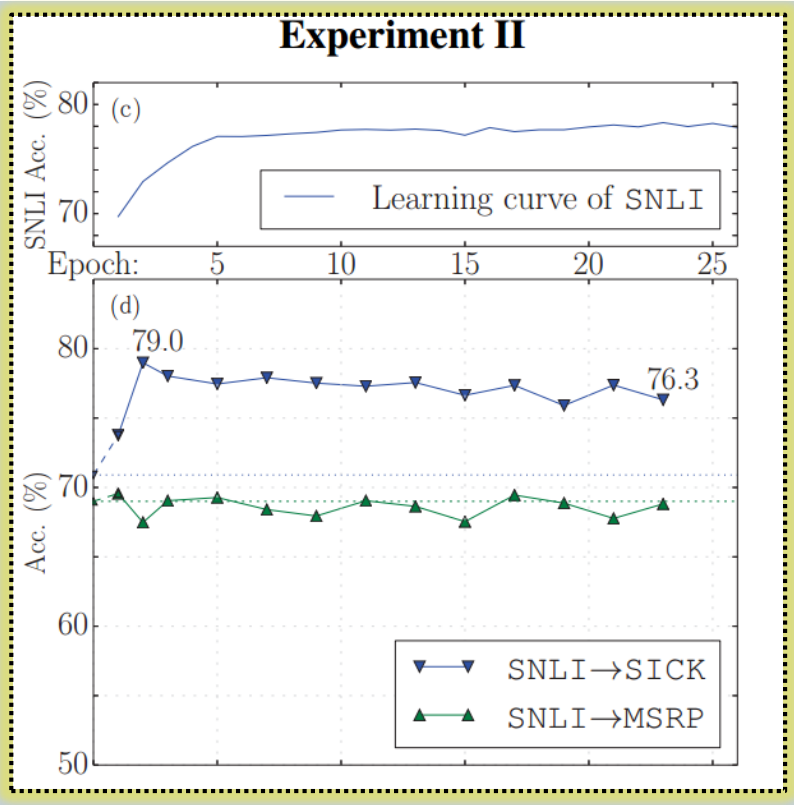
E↶H↶O□

(a) Experiment II: SNLI *to* SICK

A large learning rate does not damage the knowledge stored in the pretrained hyperparameters, but accelerates the training process to a large extent

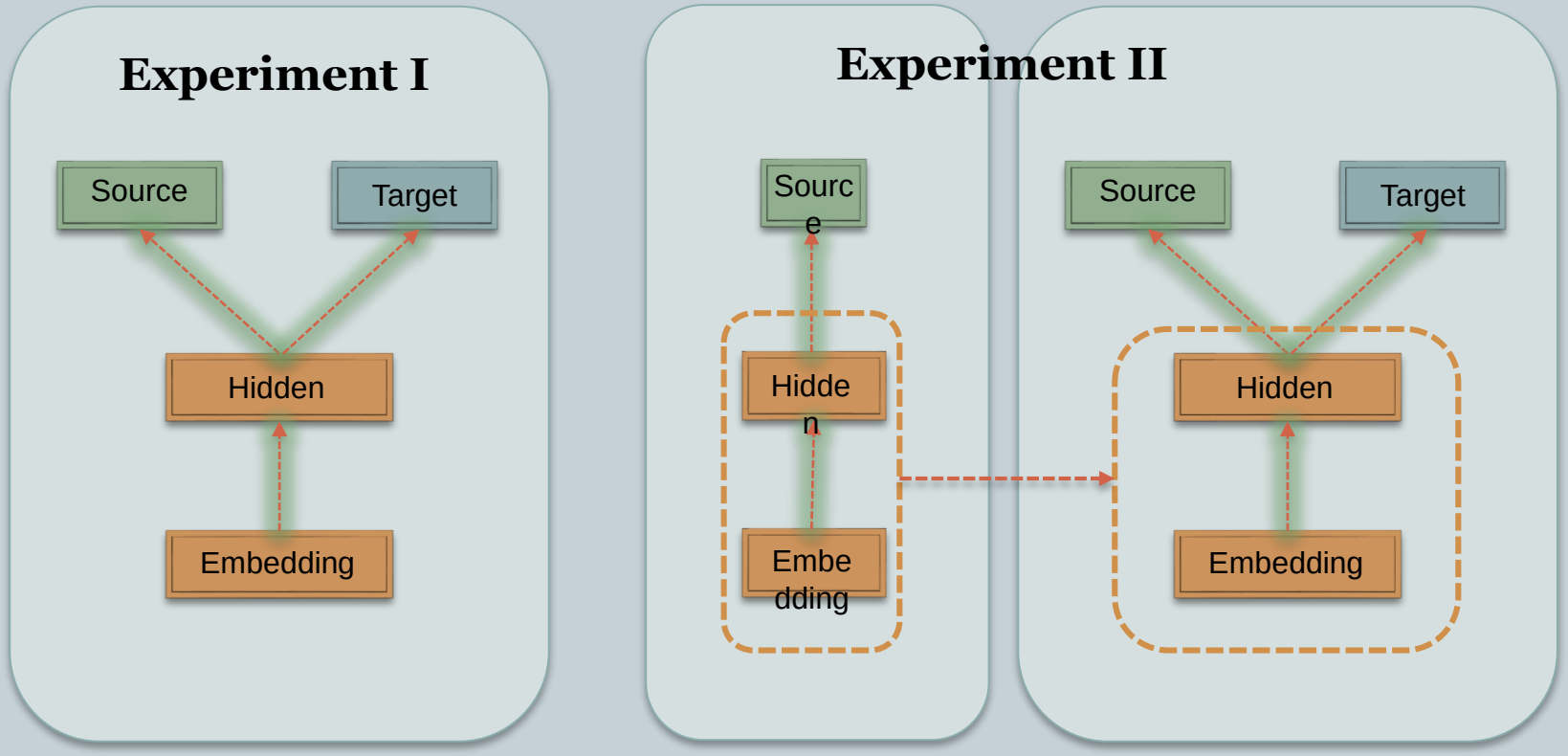
When is it ready to transfer



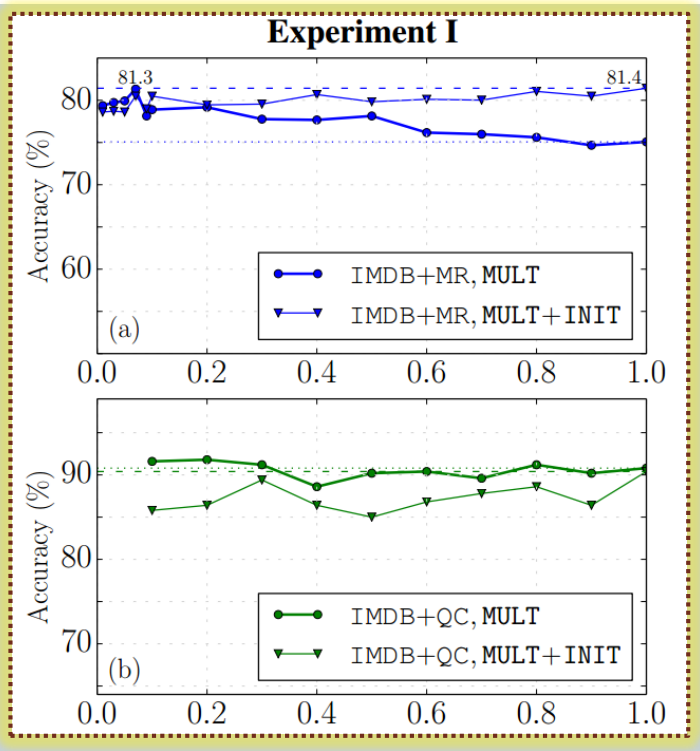


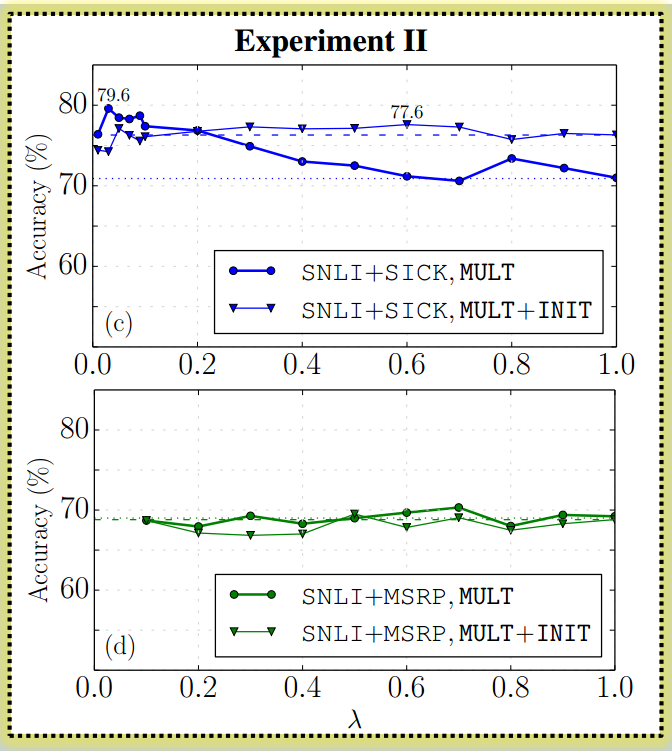
The results in these two experiments are inconsistent and lack explanation

MULT, and its Combination with INIT



MULT, and its Combination with INIT





- a small λ yields high performance of MULT in the IMDB+MR and SNLI+SICK
- we do not obtain further gain by combining MULT and INIT

Concluding Remarks

- Q1 :How transferable are neural networks between two tasks with similar or different semantics in NLP applications?

Whether a neural network is transferable in NLP depends largely on how semantically similar the tasks are.

- Q2 :How transferable are different layers of NLP neural models?

The output layer is mainly specific to a dataset. Transferring the output layer's parameters yields little (if any) gain

- Q3 :How transferable are INIT and MULT, respectively? What is the effect of combining these two methods?

MULT appears to be slightly better than (but generally comparable to) INIT in our experiment; combining MULT and INIT does not result in further gain.

Additional findings

- Q :How does learning rate affect transfer?

Transferring learning rate information is not necessarily useful. A large learning rate does not damage the knowledge stored in the pretrained hyperparameters, but accelerates the training process to a large extent. In all, we may need perform validation to choose the learning rate if computational resources are available.

- Q :When is it ready to transfer?

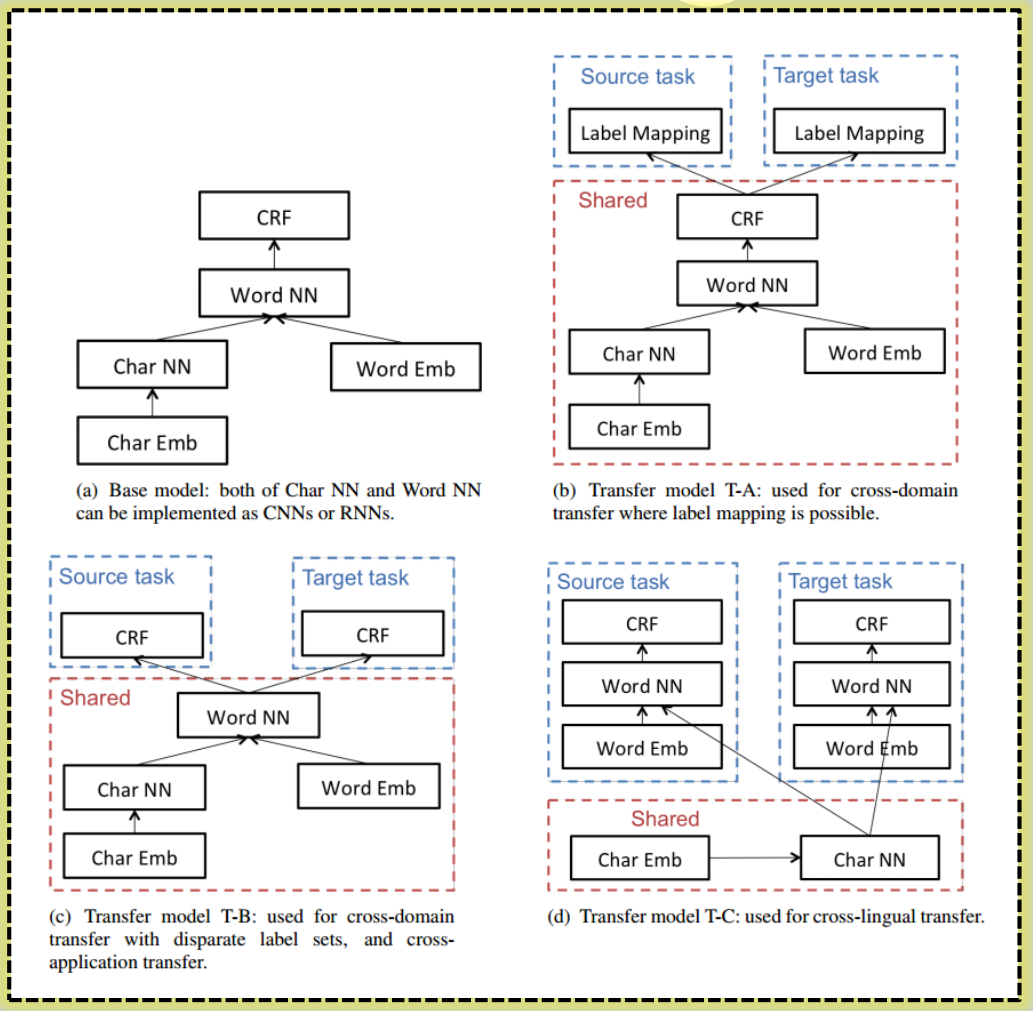
Results are not consistent.

But!

- The architecture is simple, and I don't know if there are some hierarchical features in NLP like in CV
- The experiment data sets are separated into two scenarios(semantically similar or different), belonged to qualitative analysis, and it is also necessary to do more quantitative analysis
- The results need more analysis in depth

- Further reading:
- Transfer Learning for Named-Entity Recognition with Neural Networks(MIT17)
- Transfer learning for sequence tagging with hierarchical recurrent networks(CMU17)
- Multi-task Domain Adaptation for Sequence Tagging(JHU17)

CMU17-ICLR



(a) Base model: both of Char NN and Word NN can be implemented as CNNs or RNNs.

(b) Transfer model T-A: used for cross-domain transfer where label mapping is possible.

(c) Transfer model T-B: used for cross-domain transfer with disparate label sets, and cross-application transfer.

(d) Transfer model T-C: used for cross-lingual transfer.

Char_level

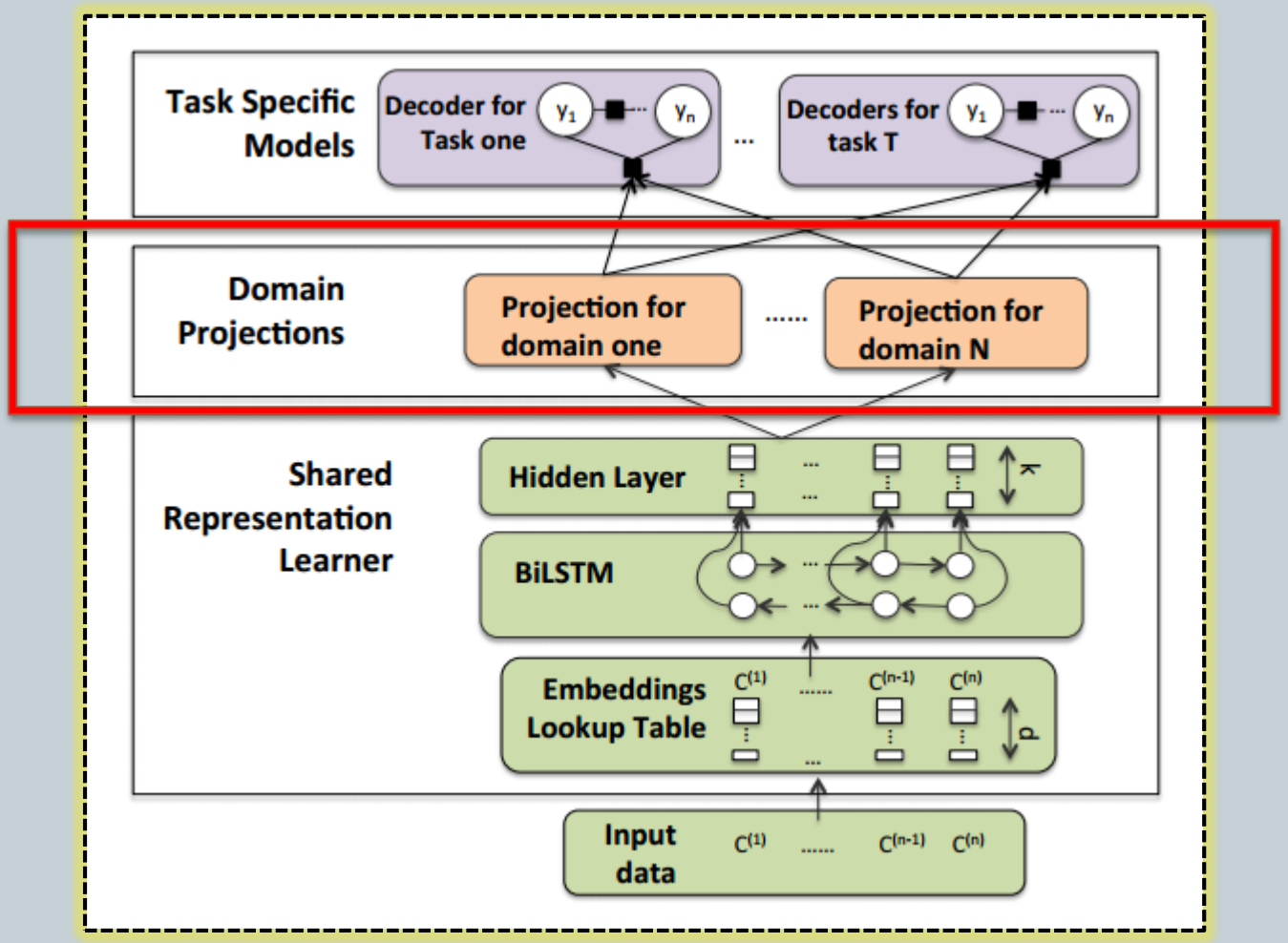
+

Word_level

+

CRF

JHU17



Domain Projections

+

Domain Masks


hmm

Thanks !