Multi-task Learning

Ramtin Mehdizadeh Seraj

Jan 2014

SFU Machine Learning Reading Group

The standard methodology in machine learning

-learning one task at a time

-Large problems are broken into small, reasonably independent subproblems that are learned separately and then recombined

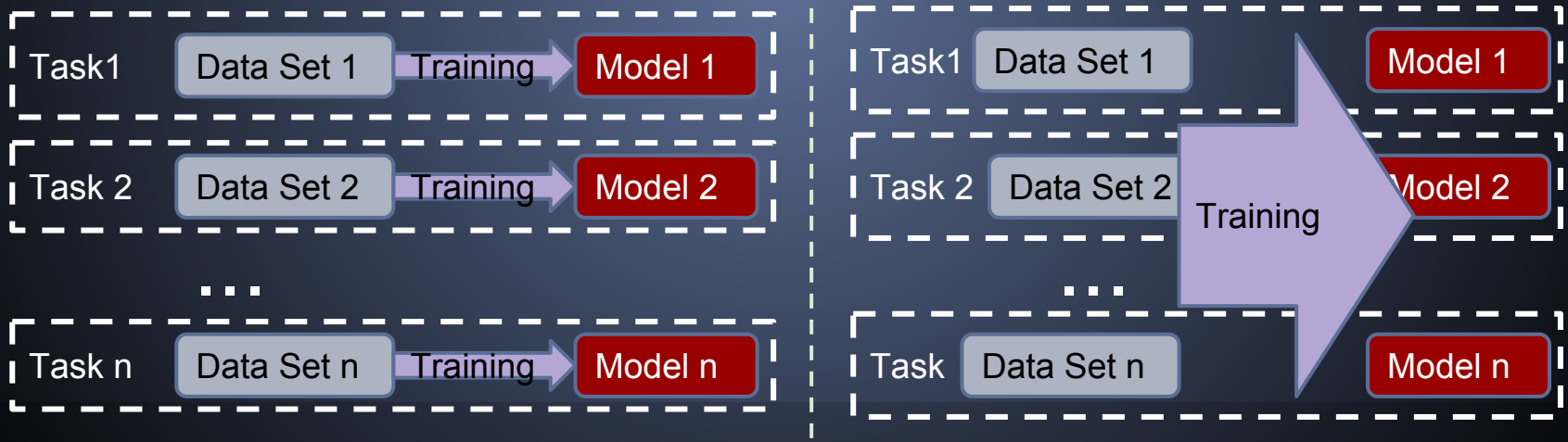
Motivation

- A net with a 1000x1000 pixel input retina is unlikely to learn to recognize complex objects in real-world scenes

- But what if we simultaneously train a net to recognize object outlines, shapes,edges, regions, subregions, textures, reflections, highlights, shadows, text, orientation,size, distance, etc.,

Concepts and General View

- According to Wikipedia :Multi-task Learning is an approach to learns a problem together with other **related** problems at the same time, using a shared representation.



Relatedness

-Learning tasks with the aim of mutual benefit

-Assumption : All tasks are related

- Example 1 : Different classification tasks

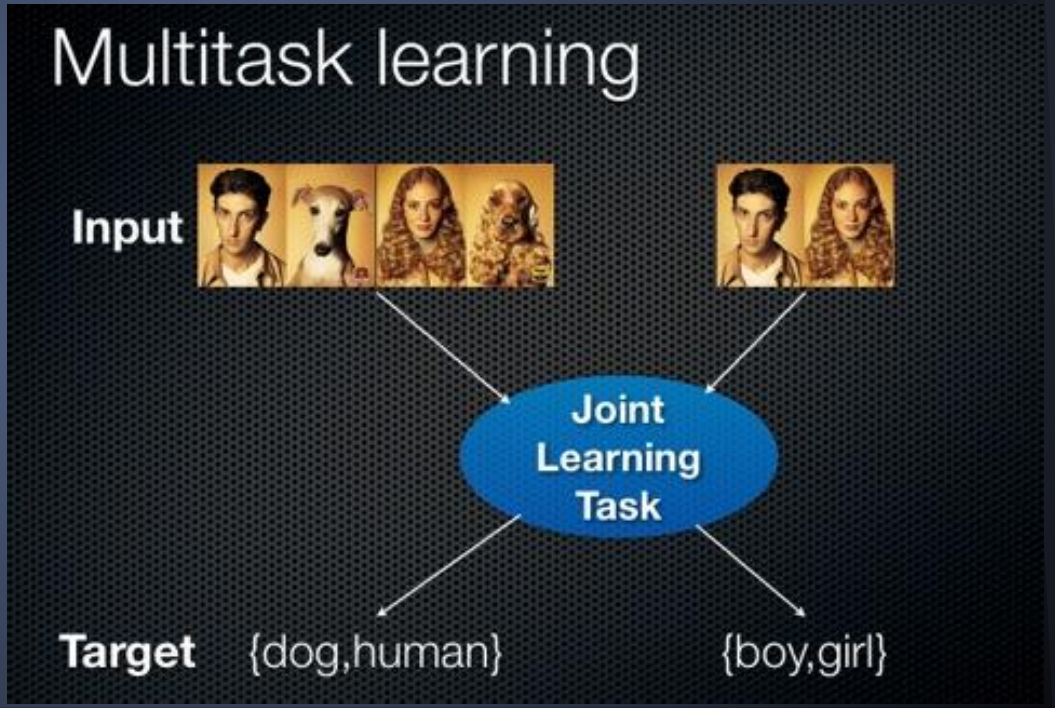
Spam filtering - Everybody Has a slightly different distribution over spam or not-spam emails

but there is a common aspect across users.

Idea : Learning together can be a good regularizer

Relatedness

Example 2 : Image Categorization



Relatedness

Other examples:

- Web Page Categorization [chen et al ICML 09]

    Page categories can be related

- Movie Ranking [Yu et. al NIPS 06]

    similar tastes between users

Learning simultaneously

- Inductions of multiple task are performed simultaneously to capture intrinsic relatedness

- The main question : How to learn ?

Learning Methods

- Joint feature learning : the simplest idea
- Mean-regularized MTL : Penalizes the deviation of each task from the mean
- Shared parameter gaussian process
- Low rank regularized
- Alternating structural optimization
- … [will discuss later]

Shared Representation

-Shared Hidden node in a Neural Network:

The simplest one can be a neural network shared hidden units among tasks .

- Shared Parameter:

  Like Gaussian process

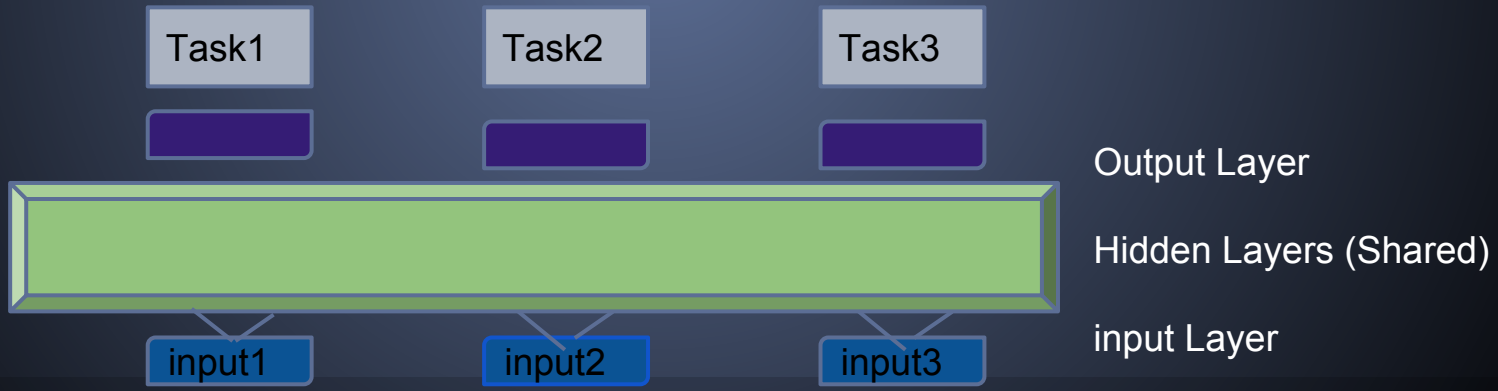
- Regularization-based :

  Mean , Joint feature table, ...

Shared Representation

Sharing Hidden Nodes in Neural Network

- A set of hidden units are shared among multiple tasks.(goal :improving generalization)



Shared Representation

-Joint Feature Learning

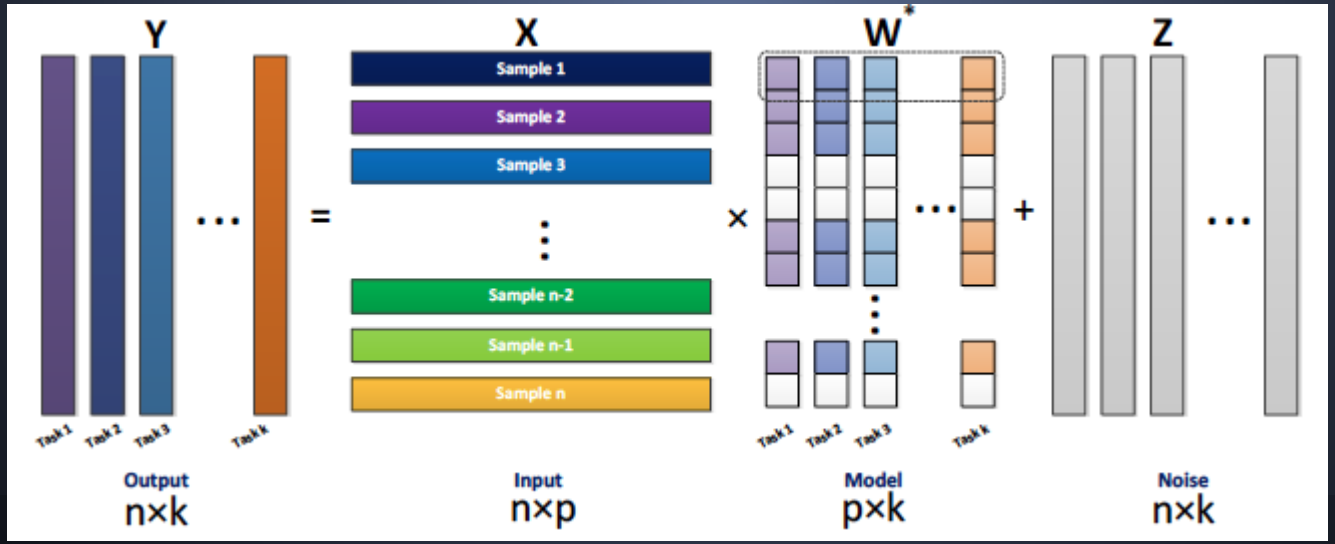
creating a common set of features

MTL with Joint Feature learning

-Using Group Sparsity

l1/l2-norm regularization

$$\min_W \frac{1}{2}\|XW - Y\|_F^2 + \lambda \sum_{i=1}^{p} \|\boldsymbol{w}_i\|_2$$



An Application In NLP

- A unified architecture for Natural Language Processing deep neural net with multi task learning (by Collobert and Watson)

-Tasks :POS, NER, Chunking, Semantic Roles,...

-Relatedness : Are these tasks related ?

-Shared Representation: NN layers

-Training : Joint training using weight sharing

An Application In NLP - Intro

- Tasks :

    1. POS (Part of Speech Tagging): labeling each word with a unique tag that shows its tactic roles, ex. adverb, noun,...

    2. Chunking: labeling segments of a sentence with syntactic constituents

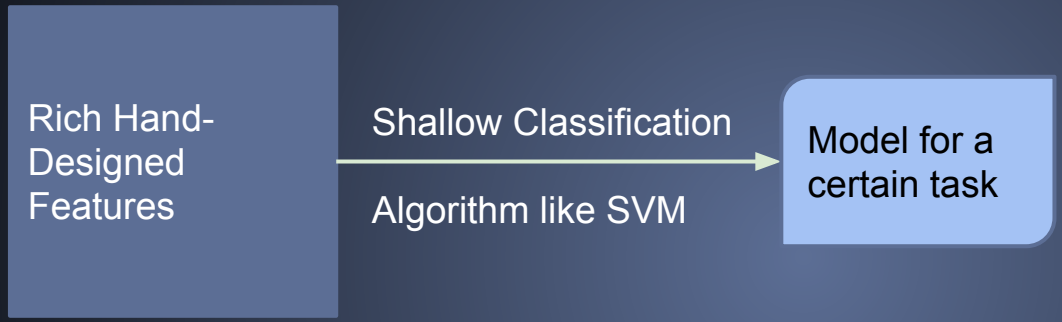
An Application In NLP - Intro

3. Named Entity Recognition: Labeling atomic elements in the sentence into categories such that “Location”, “Person”

4. Semantic Role Labeling: Giving a semantic role to a syntactic constituent of a sentence.

Example: [John]Arg0 [ate]Rel [the apple]Arg1

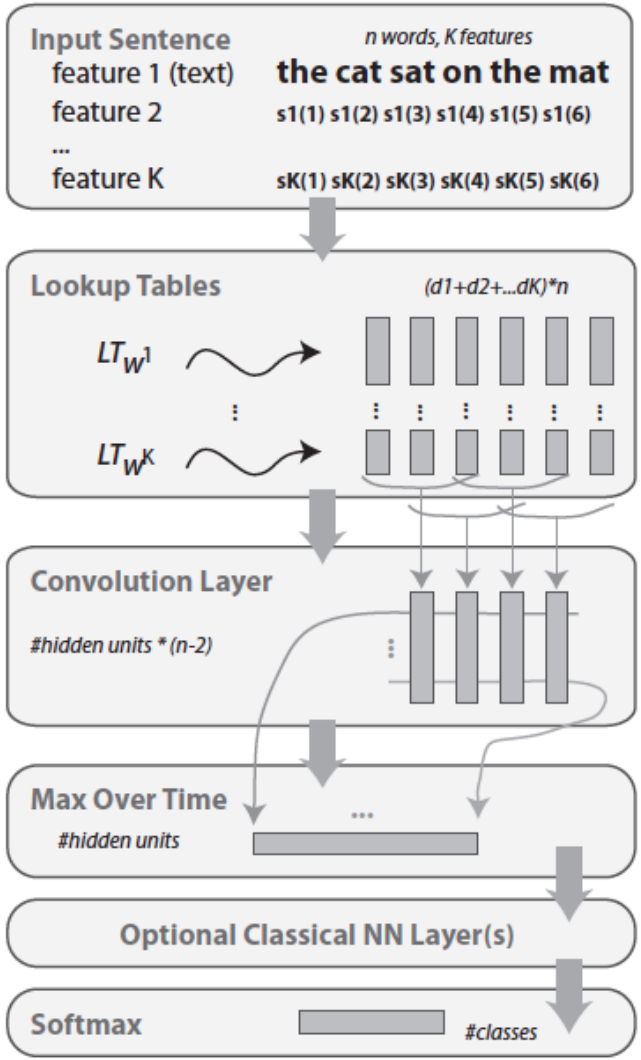
An Application In NLP - Regular approaches



Selecting features by empirical process (trial and error)
Task-based algorithm selection

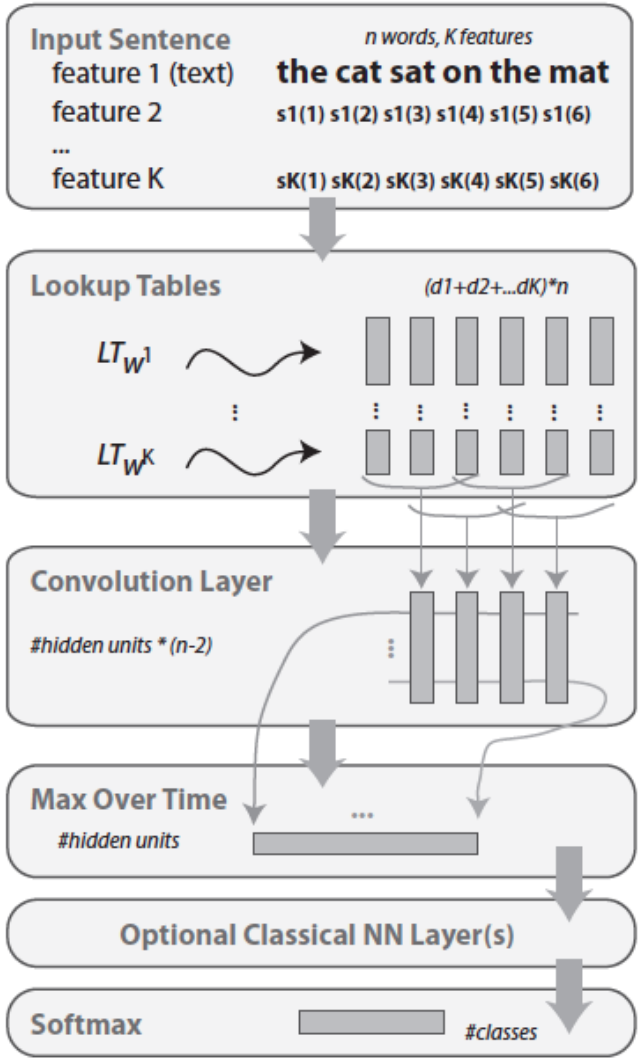
An Application In NLP 1 - new approach

- Deep Neural Network
- Feature extraction in several layers using back propagation



An Application In NLP 2 - new approach

- First Layer : features for each words
- Second Layer : features for the input sentence (sequenced based)
- Following layers : Classical NN layers



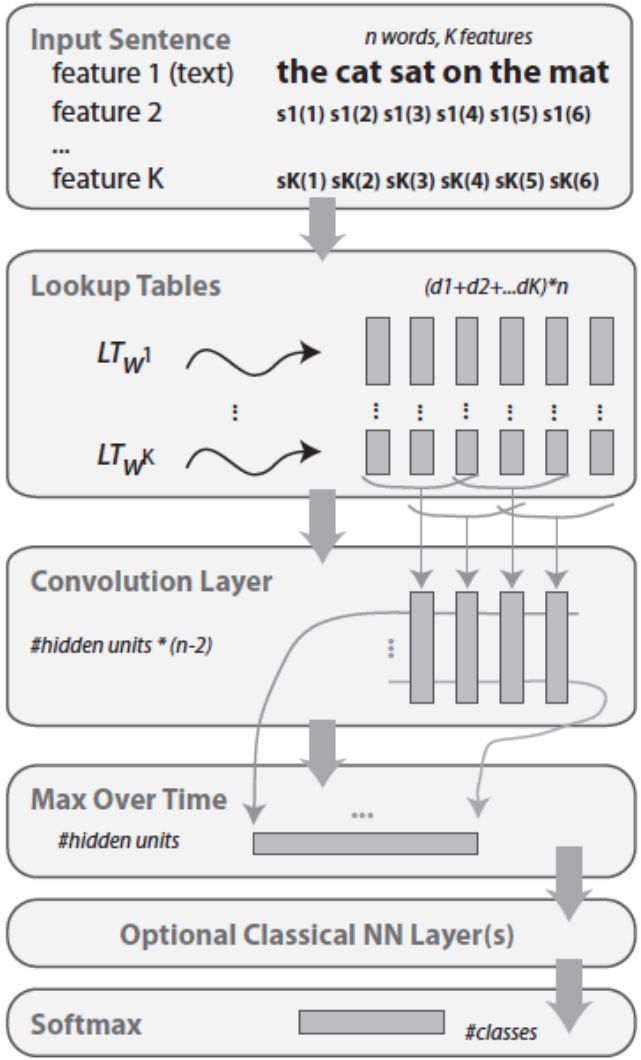
An Application In NLP 3- Look up tables layer

- for word i in the Dictionary considering a d-dimensional space

LTw(i) = Wi

-W : parameters to be learnt

- For solving variable sentence length: Considering fixed size window size around each word.

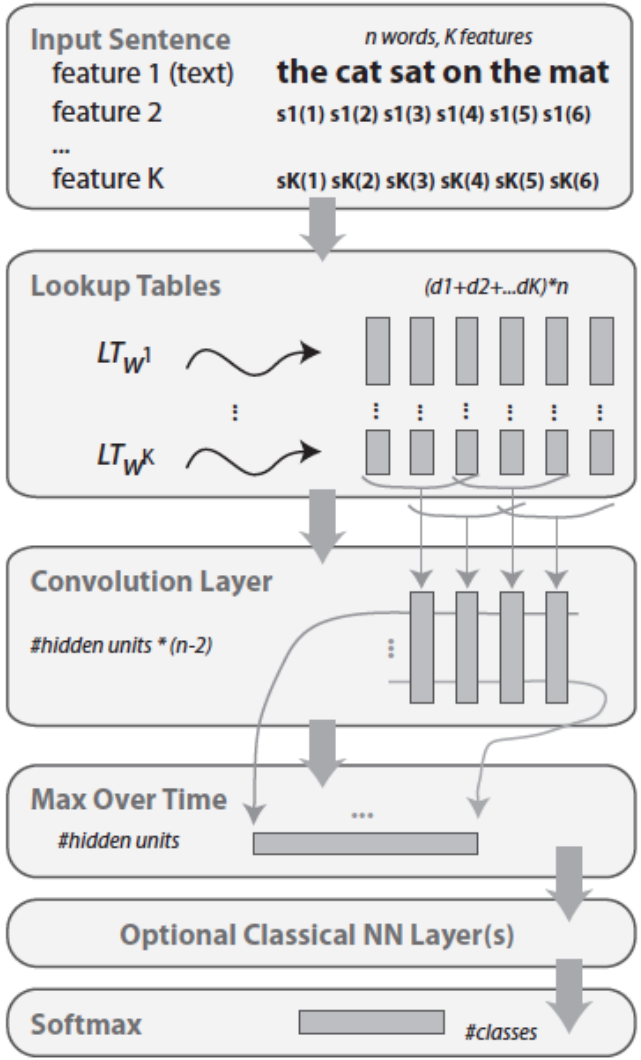


An Application In NLP 4- NN and Max Layer

- Time Delay Neural Network :

perform linear operation over the input words.

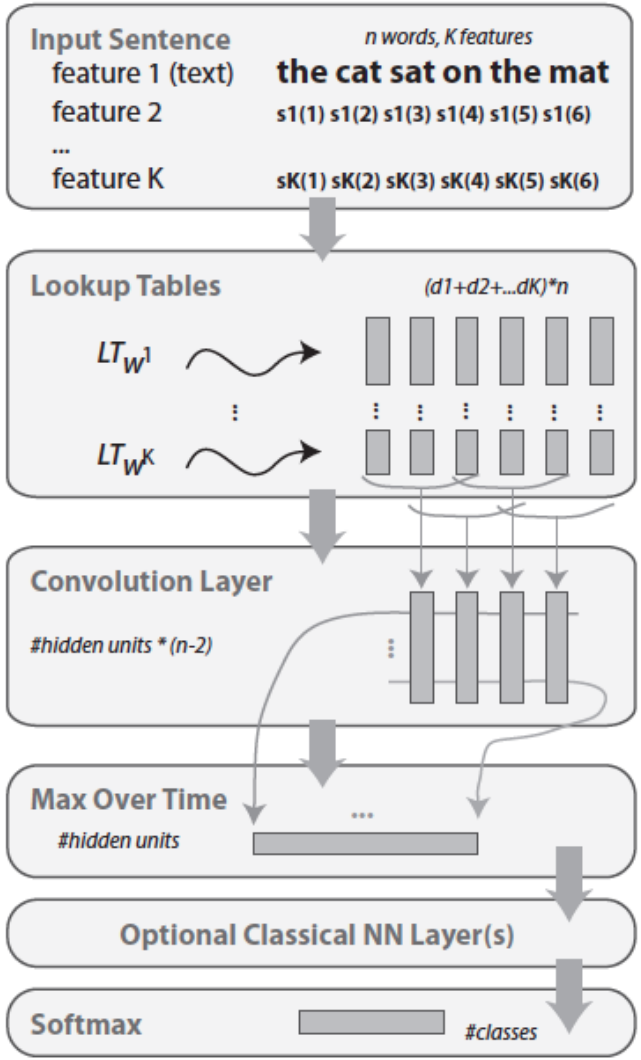
- Max Layer : Captures the most relevant features over the sentence.



An Application In NLP 5- Output and Algorithm

- Using softmax for joint learning
- Algorithm (training in the stochastic manner) :

1. select the next task
2. select a random training example for this task
3. Use gradient for updating NN
4. go to step 1



Results

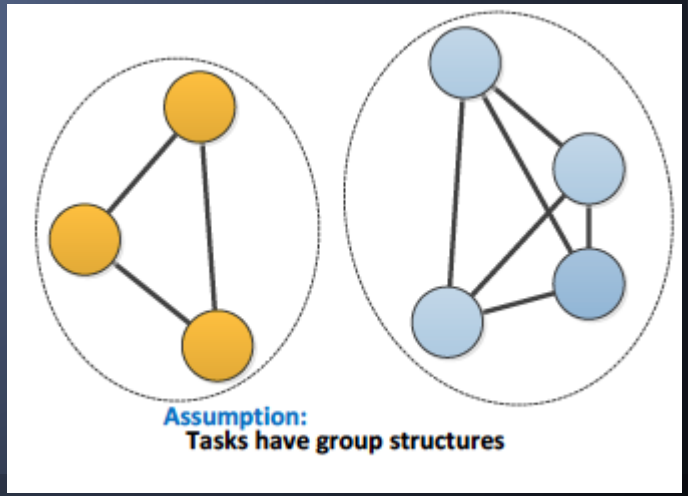
	wsz=15	wsz=50	wsz=100
SRL	16.54	17.33	18.40
SRL + POS	15.99	16.57	16.53
SRL + Chunking	16.42	16.39	16.48
SRL + NER	16.67	17.29	17.21
SRL + Synonyms	15.46	15.17	15.17
SRL + Language model	14.42	14.30	14.46
SRL + POS + Chunking	16.46	15.95	16.41
SRL + POS + NER	16.45	16.89	16.29
SRL + POS + Chunking + NER	16.33	16.36	16.27
SRL + POS + Chunking + NER + Synonyms	15.71	14.76	15.48
SRL + POS + Chunking + NER + Language model	14.63	14.44	14.50

What if tasks are not totally related

- If the tasks have a group structures
  => Clustered Multi-task learning

e.g. tasks in the yellow group are predictions of heart related diseases and in the blue group are brain related diseases.

more information : Bakker and Heskes JMLR 2003



What if tasks are not totally related

- If the tasks have a tree structures

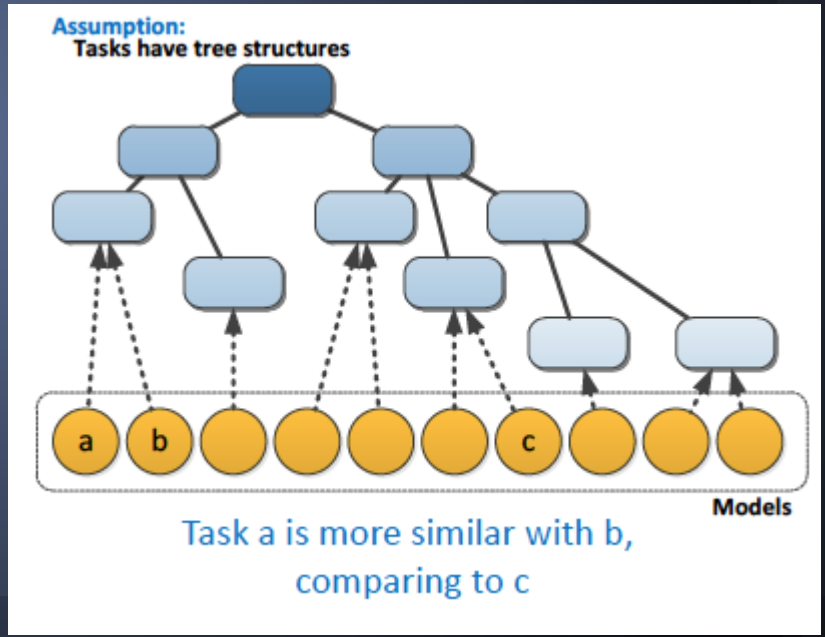
  => Multi-task Learning

  with Tree Structures

more information :

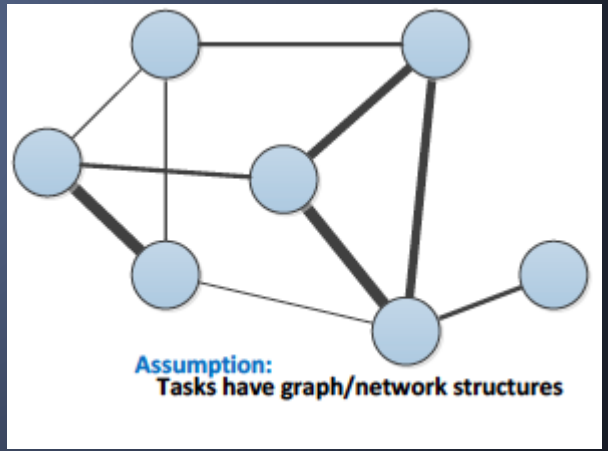
Tree-Guided Group Lasso

(Kim and Xing 2010 ICML)



What if tasks are not totally related

- If the tasks have a graph structures
  => Multi-task Learning
  with Graph Structures

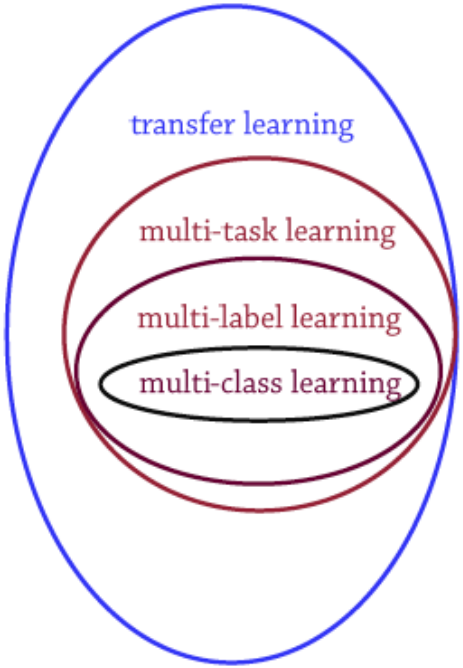


more information :

Graph-guided Fused Lasso (Chen et. al. UAI11)

Connection to other ML topics

Learning Methods



- Transfer Learning
  - Define source & target domains
  - Learn on the source domain
  - Generalize on the target domain
- Multi-task Learning
  - Model the task relatedness
  - Learn all tasks simultaneously
  - Tasks may have different data/features
- Multi-label Learning
  - Model the label relatedness
  - Learn all labels simultaneously
  - Labels share the same data/features
- Multi-class Learning
  - Learn the classes independently
  - All classes are exclusive

Software Packages

MALSAR: Multi-tAsk Learning via StructurAl Regularization

-Implemented by Biodesign Institute of Arizona State University

Main References

- Caruana, R. (1997). Multitask Learning. Machine Learning, 28(1), 41–75. doi: 10.1023/A:1007379606734

- Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. Presented at the Proceedings of the 25th international conference ….

- Lounici, K., Pontil, M., Tsybakov, A. B., & van de Geer, S. (2009, March 8). Taking Advantage of Sparsity in Multi-Task Learning. arXiv.org.

- Zhang, Y., & Yeung, D.-Y. (2012, March 15). A Convex Formulation for Learning Task Relationships in Multi-Task Learning. arXiv.org.

- Zhou, J., Chen, J., & Ye, J. (2012) Multi-Task Learning , Theory, Algorithms, and Applications, SDM

Thanks for you attention

Any Question ???