NLPCC 2018 Tutorials

復旦大學
FUDAN UNIVERSITY

自然语言处理中的多任务学习

邱锡鹏
复旦大学
2018年9月12日
http://nlp.fudan.edu.cn/xpqiu

研究组介绍

- 主要聚焦于深度学习与自然语言处理领域，包括语言表示学习、词法/句法分析、文本推理、问答系统等方面。
- 主要成果
  - 近几年发表国际顶级会议/期刊（IJCAI、ACL、AAAI、EMNLP等）论文40余篇，ACL2017杰出论文
  - 开源自然语言处理系统：FudanNLP
    - Currently, an incubating project: fastNLP

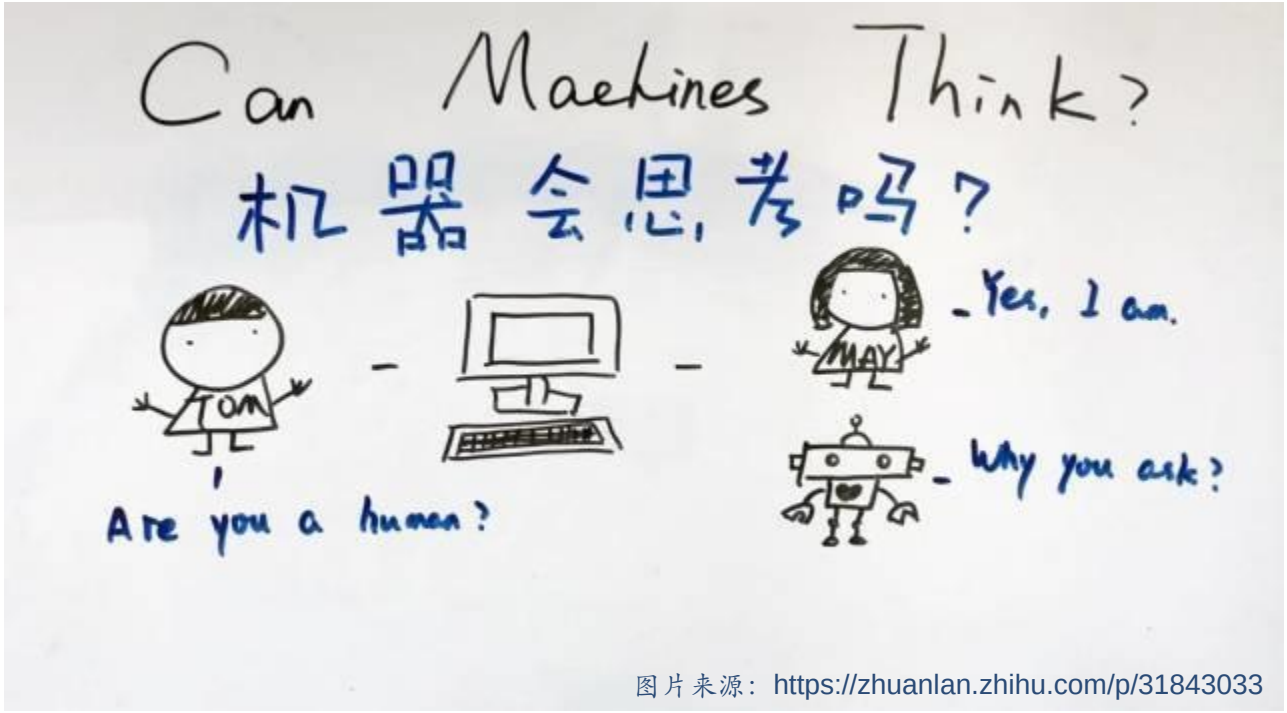
报告概要

- 自然语言处理简介
- 基于深度学习的自然语言处理
- 深度学习在自然语言处理中的困境
  - 无监督预训练
  - 多任务学习
- 自然语言处理中的多任务学习
  - 硬共享模式
  - 软共享模式
  - 共享-私有模式
  - 函数共享模式
  - 多级共享模式
  - 主辅任务模式
- 新的多任务基准平台

自然语言处理简介

从人工智能开始

Alan Turing (1912-1954)

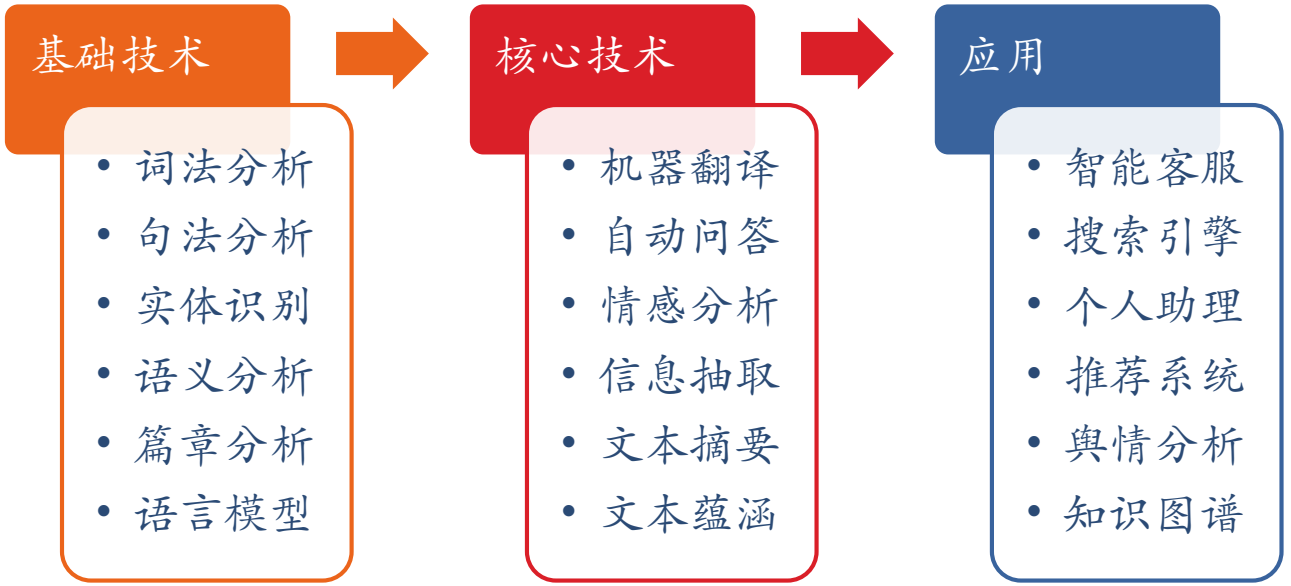


图片来源：https://zhuanlan.zhihu.com/p/31843033

自然语言处理：理解和生成

什么是自然语言处理（NLP）？

- 自然语言≈人类语言
  - 区别于人工语言（比如程序语言）
- 自然语言处理包括语音识别、自然语言理解、自然语言生成、人机交互以及所涉及的中间阶段。
  - 是人工智能和计算机科学的子学科。

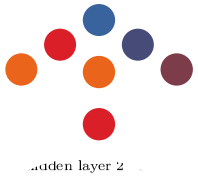


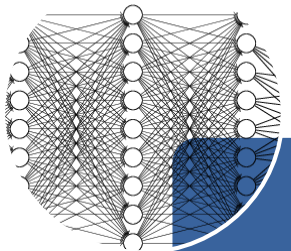
自然语言处理的难点：歧义性

- 以中文分词为例
  - 不同的语言环境中的同形异构现象，按照具体语言环境的语义进行切法。
  - 交叉歧义
    - 他／说／的／确实／在理
  - 组合歧义
    - 两个/人/一起/过去、个人/问题
    - 从马/上/下来、马上/就/来
  - 句子级歧义
    - 白天鹅在水里游泳
    - 该研究所获得的成果

伪歧义

发展历程





数据驱动

2012年~

基于深度学习的方法

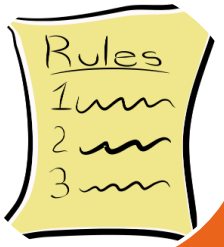
- 神经网络
- 分布式表示
- 非线性模型

规则驱动

1990年~2012年

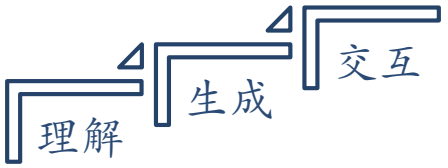
基于统计学习的方法

- 离散表示
- 线性模型

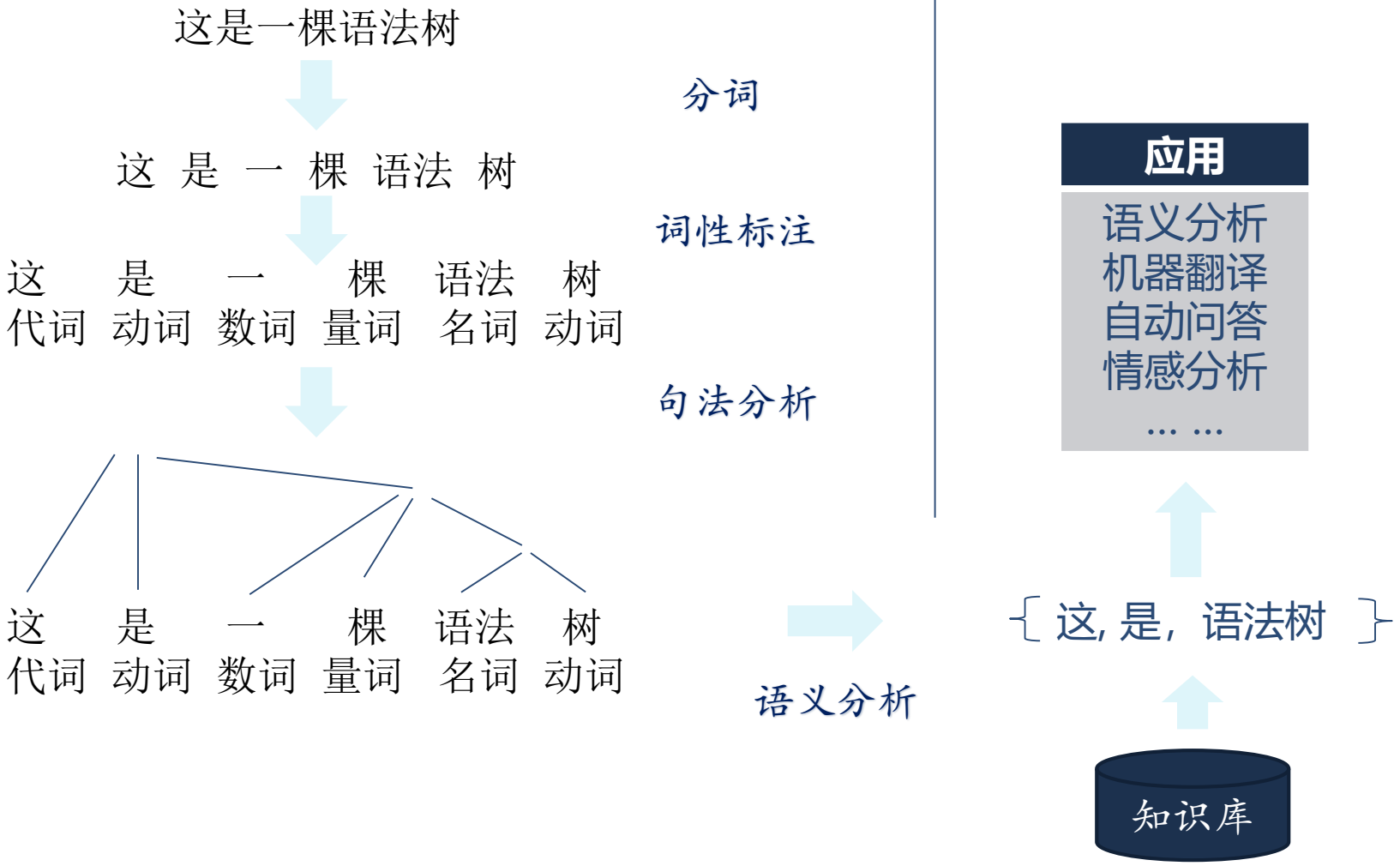


1990年以前

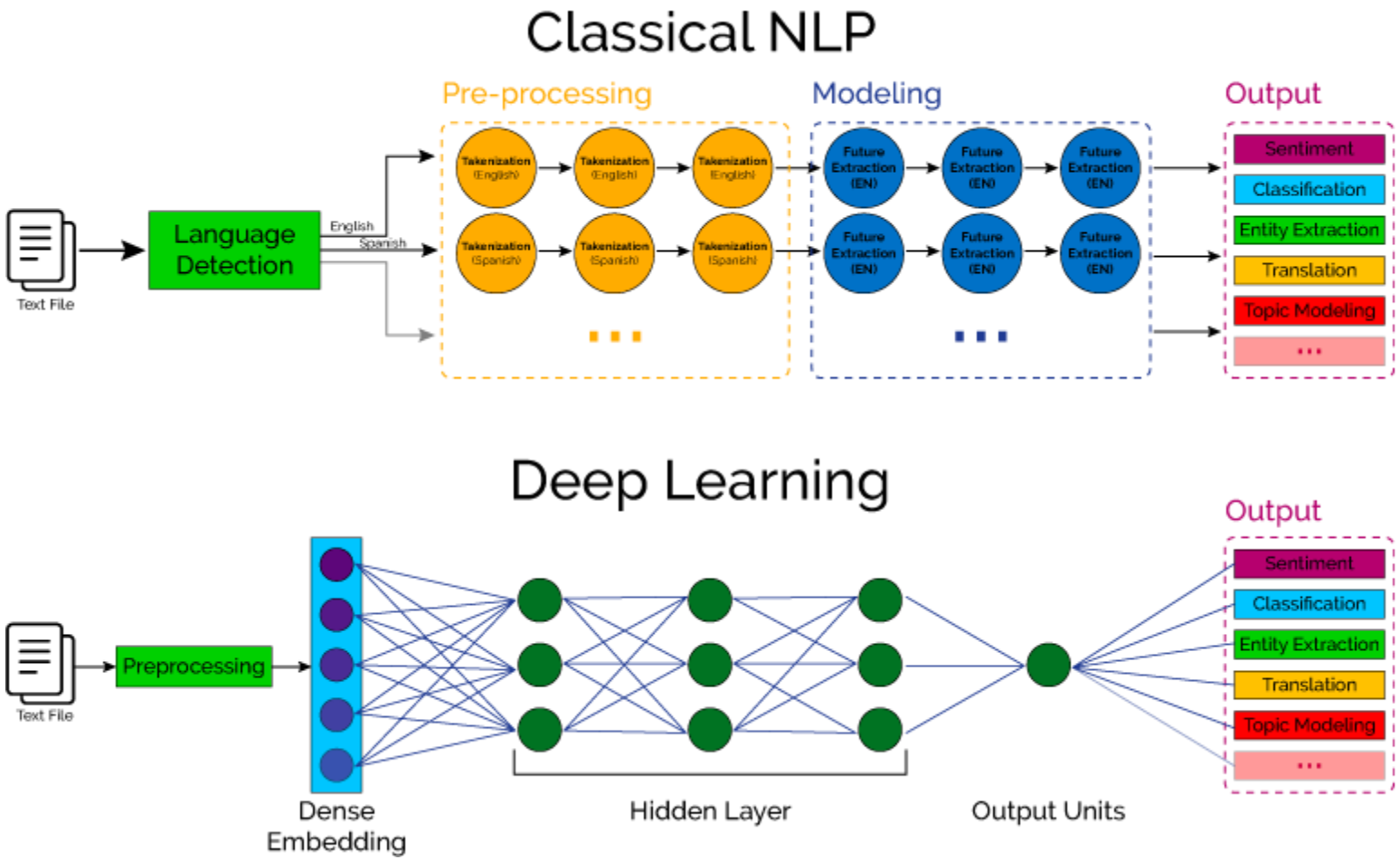
基于规则的方法



理想中的自然语言处理流程



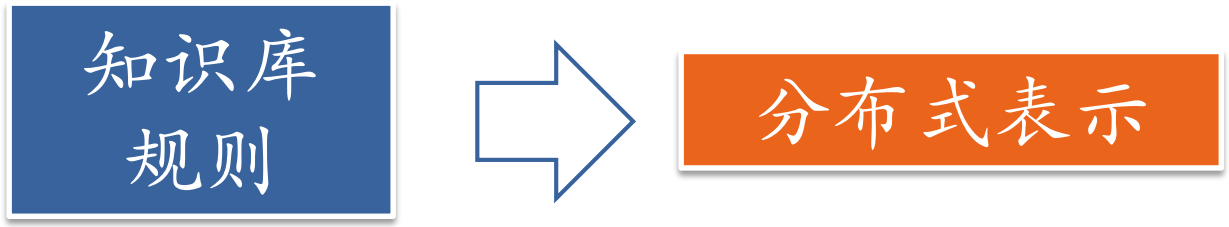
NLP技术路线


Classical NLP
Pre-processing
Modeling
Output
Text File
Language Detection
English
Spanish
Tokenization (English)
Tokenization (Spanish)
Future Extraction (EN)
Sentiment
Classification
Entity Extraction
Translation
Topic Modeling
Deep Learning
Output
Text File
Preprocessing
Dense Embedding
Hidden Layer
Output Units
Sentiment
Classification
Entity Extraction
Translation
Topic Modeling

基于深度学习的自然语言处理

语言表示

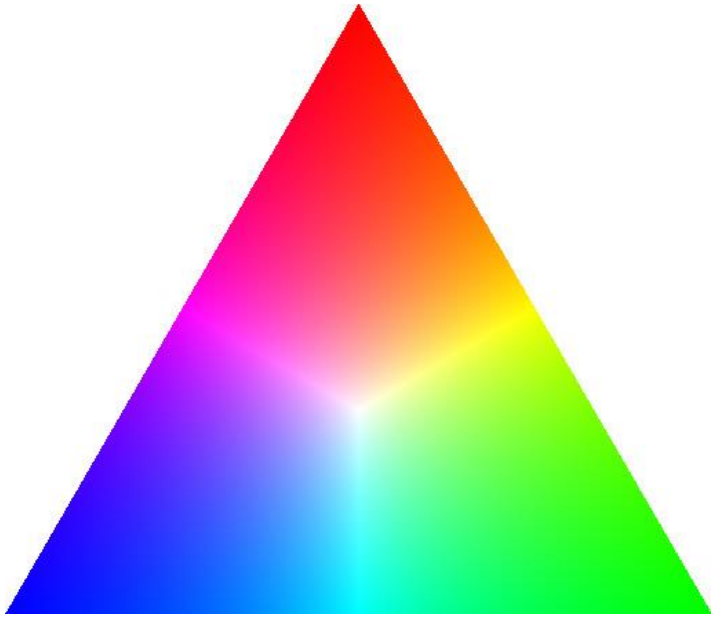
- 如何在计算机中表示语言的语义?



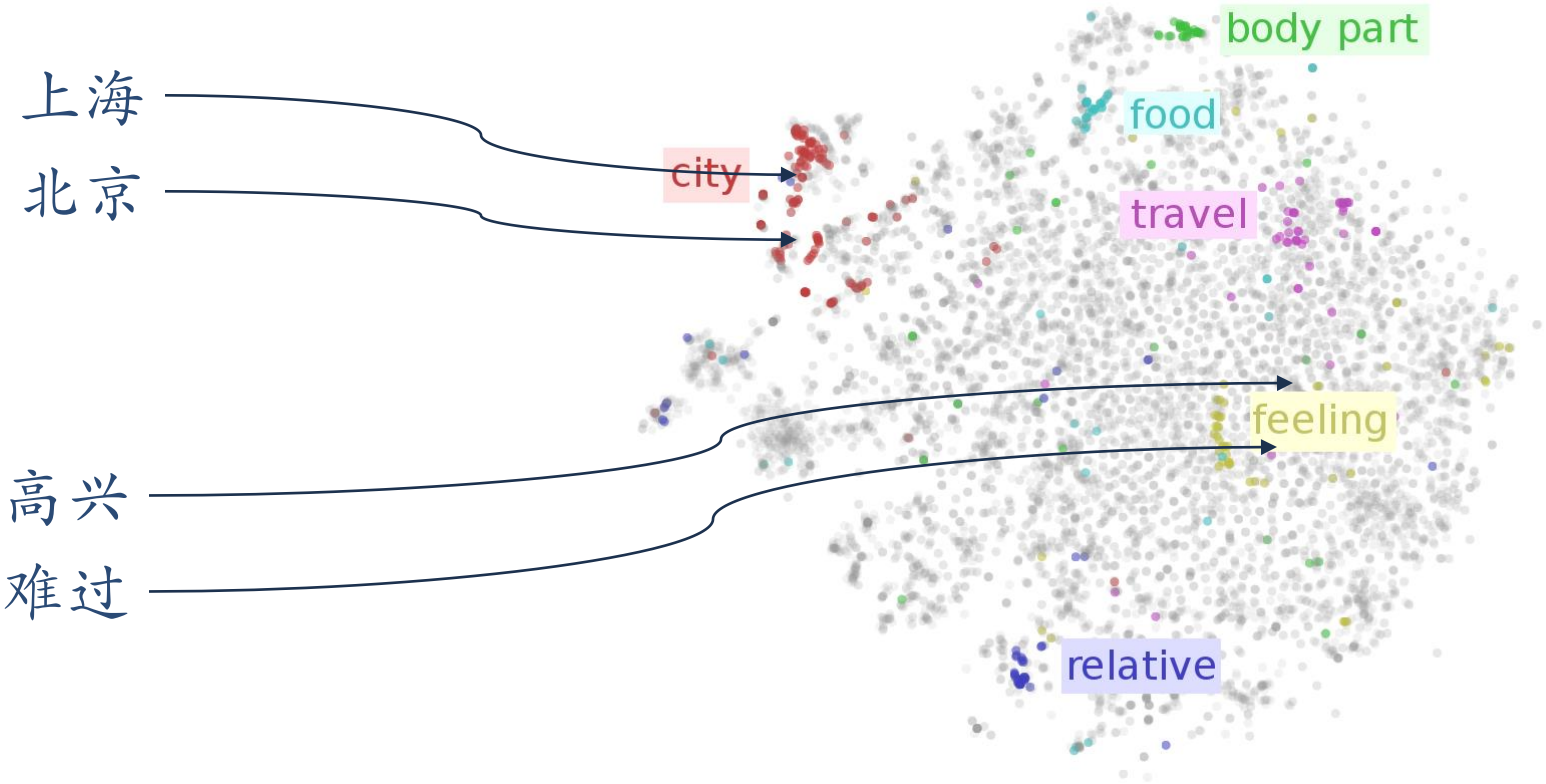
- 压缩、低维、稠密向量
- 用O(N)个参数表示 $O(2^k)$区间
  - k为非0参数，$k<N$

一个生活中的例子：颜色

命名	RGB值
红	[1,0,0]
绿	[0,1,0]
蓝	[0,0,1]
中国红	[0.67, 0.22, 0.12]
咖啡色	[0.64, 0.16,0.16]



词嵌入（Word Embeddings）



https://indico.io/blog/visualizing-with-t-sne/

分布式表示
--来自神经科学的证据

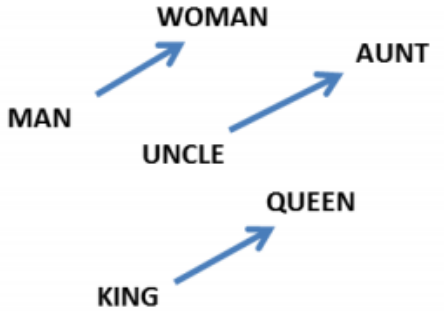
http://www.nature.com/nature/journal/v532/n7600/full/nature17637.html

词嵌入

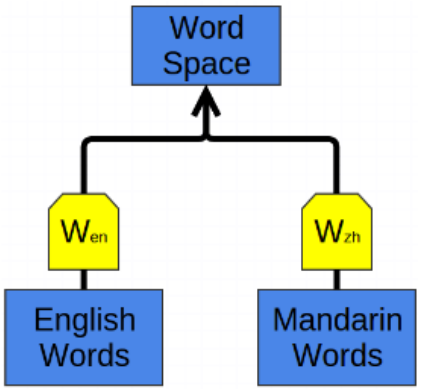
W("woman")−W("man") ≈ W("aunt")−W("uncle")

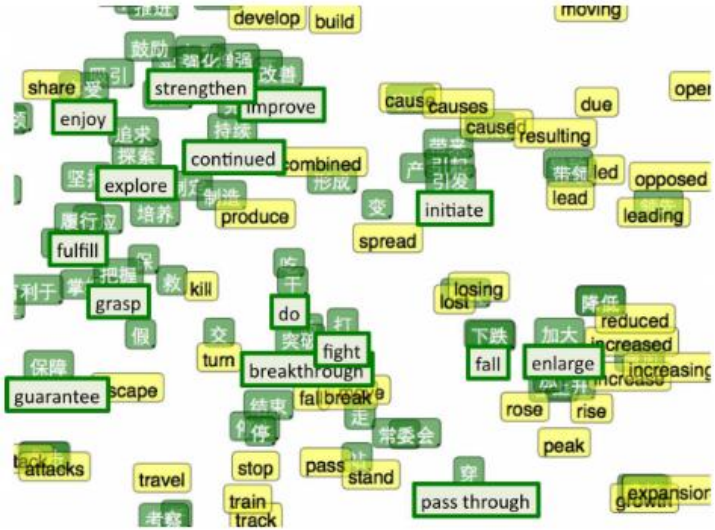
W("woman")−W("man") ≈ W("queen")−W("king")

W("中国")−W("北京") ≈ W("英国")−W("伦敦")



From Mikolov et al. (2013)

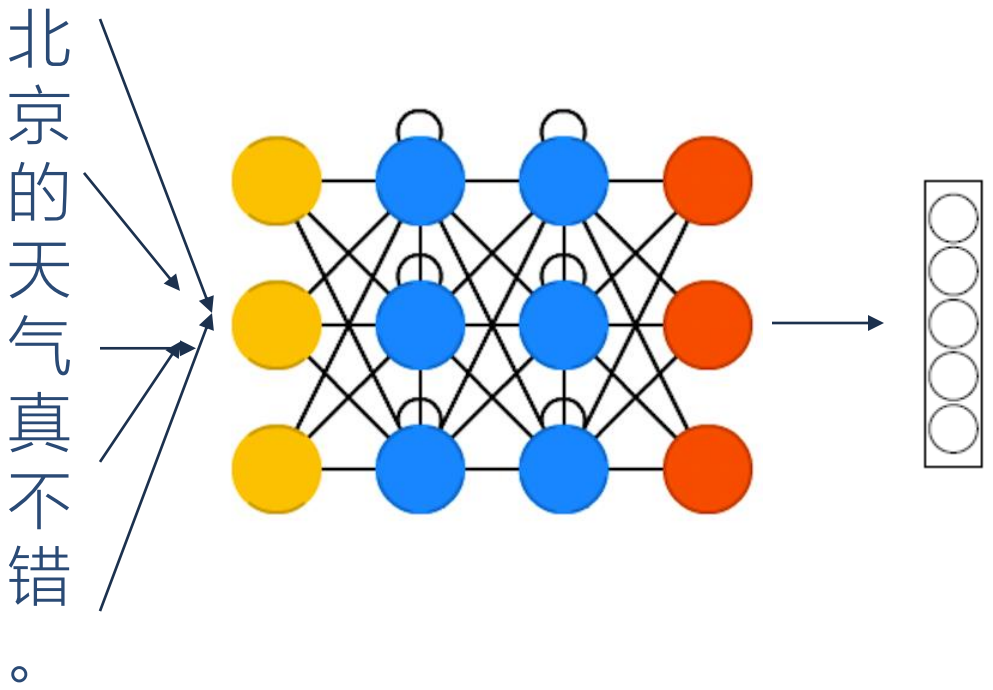




Socher et al. (2013)

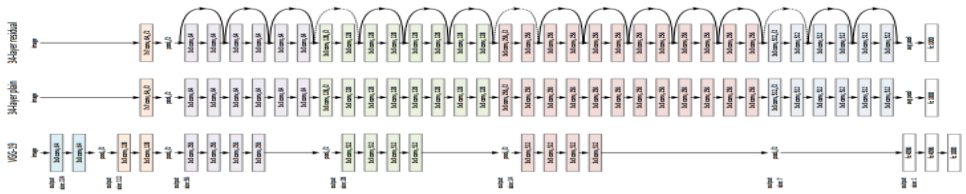
语言表示学习

- 词
- 短语
  - 组合语义模型
- 句子
  - 连续词袋模型
  - 序列模型
  - 递归组合模型
  - 卷积模型
- 篇章
  - 层次模型

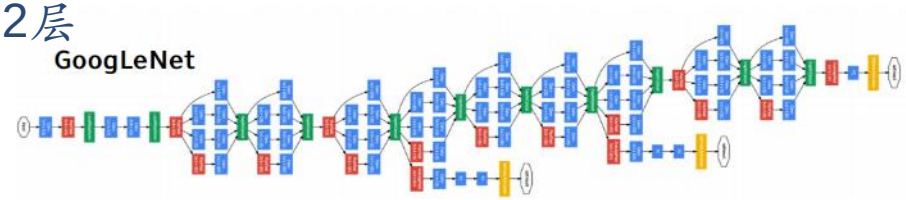


为什么语言表示学习更难?

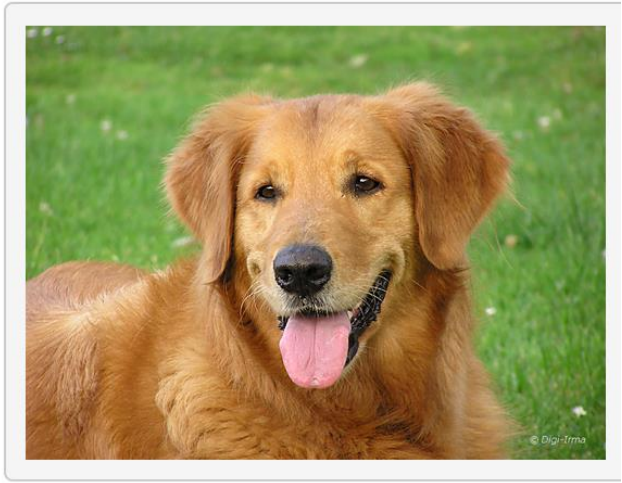
152 层



22层



计算机视觉中的深层网络模型



Results:

- golden retriever: 0.97293
- Tibetan mastiff: 0.01576
- Irish setter: 0.00364
- redbone: 0.00152
- standard poodle: 0.00127

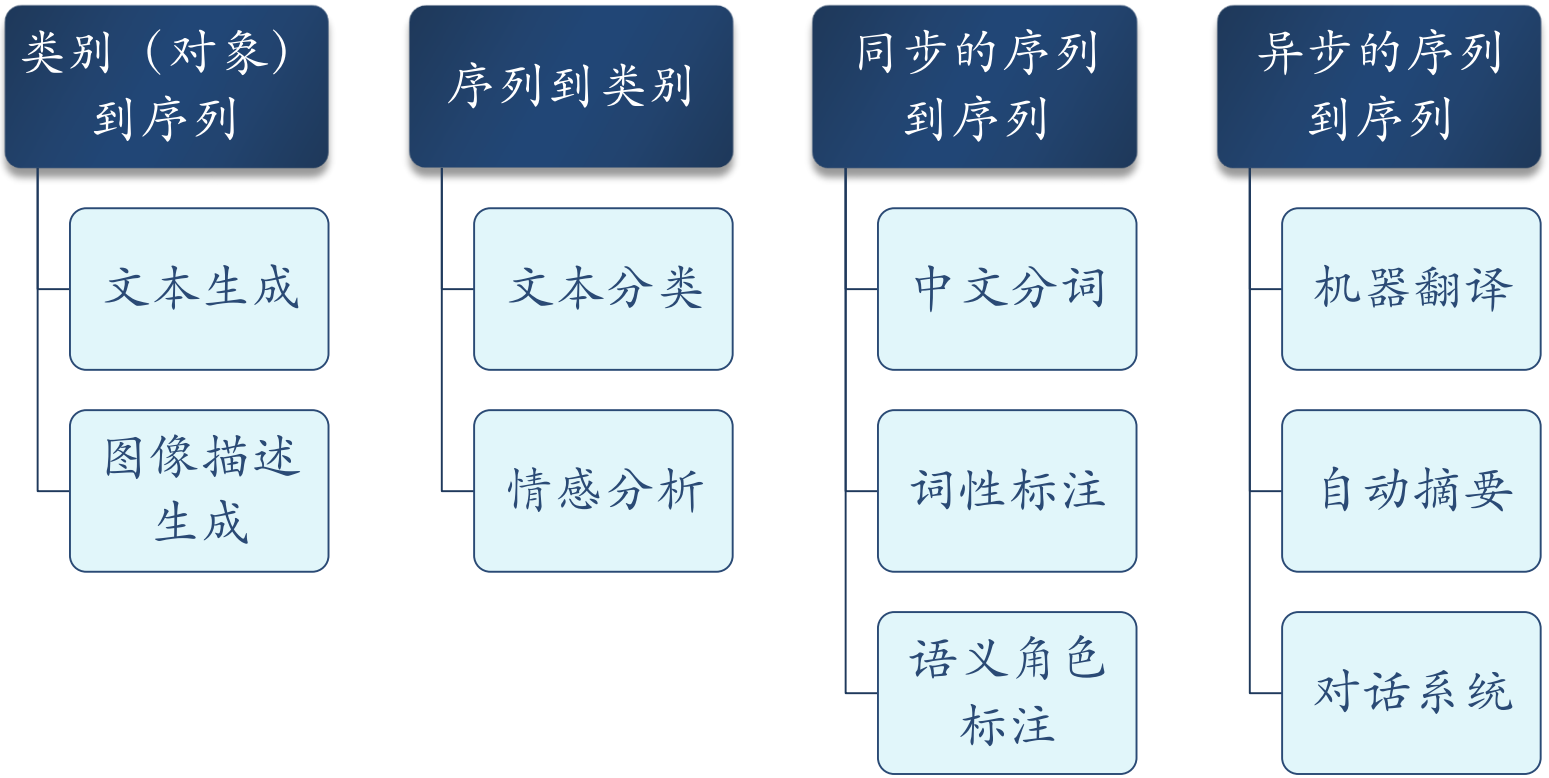
对应NLP的最底层：词汇

文本与图像信息的差异

	输入量	信息量	关系	底层特征
图像	二维像素集 200X200	黑白：128-256 彩色：3（128-256）	欧氏空间	纹理，形状 彩色
文本	一维离散词符号序列 几千-几万个词	共250K（英文词类） 一般用几千个词	语法关系 句法关系 语义关系	句子长度，句子在段落中的位置，段落在文章中的位置，..

自然语言处理的新范式

- 在得到字、句子表示之后，自然语言处理任务类型划分为

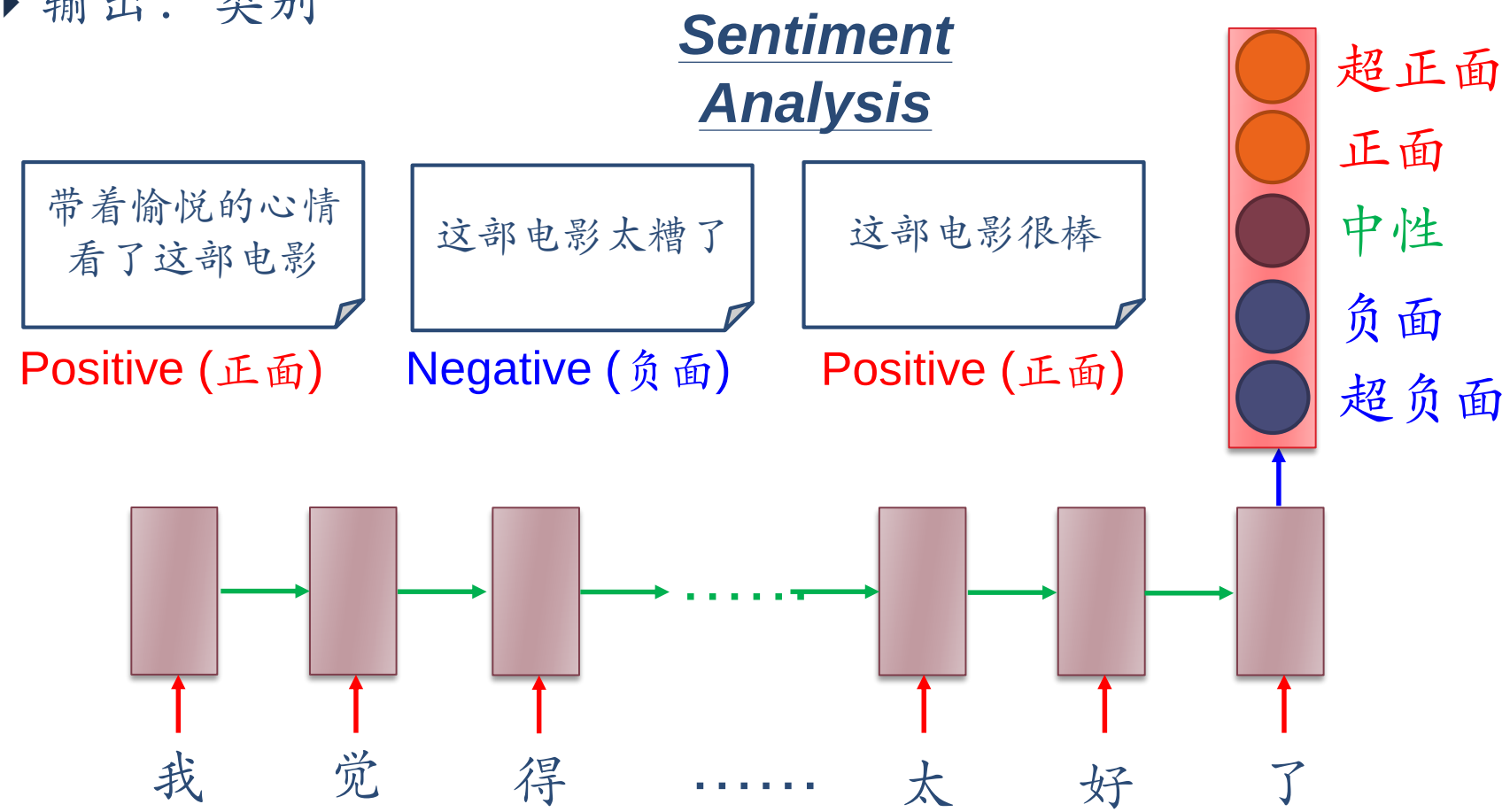


减轻了对特征工程的依赖！

序列到类别

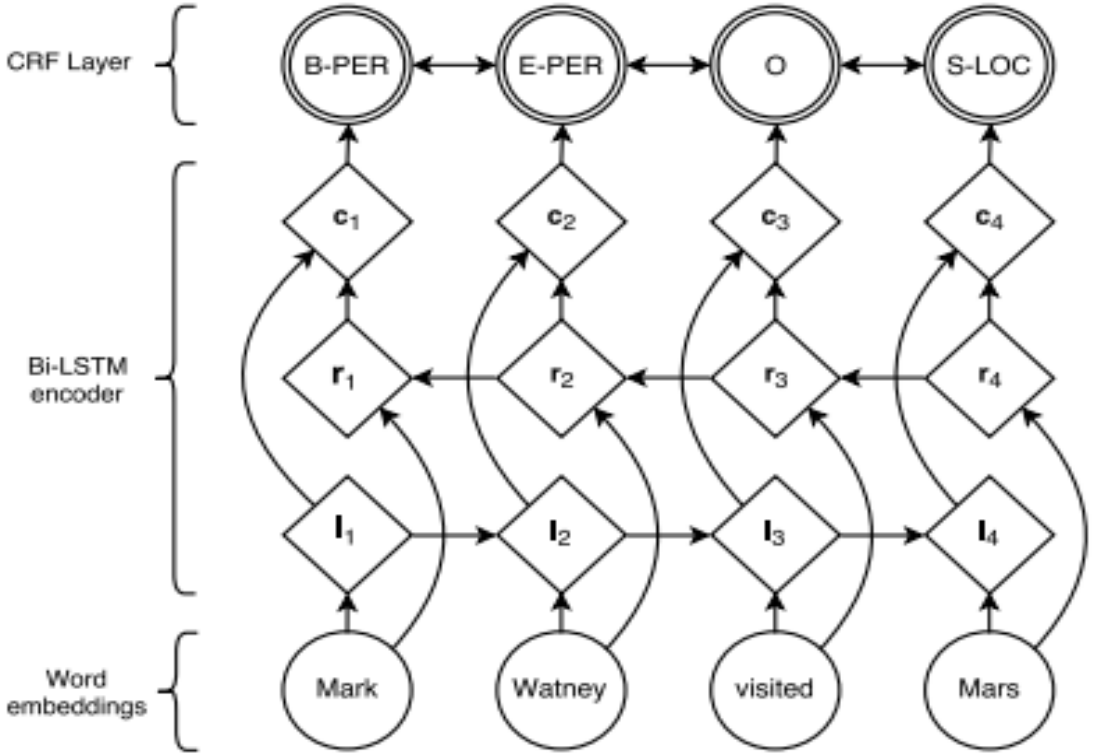
来源：李宏毅《1天搞懂深度学习》

- 输入：序列
- 输出：类别



同步的序列到序列模式

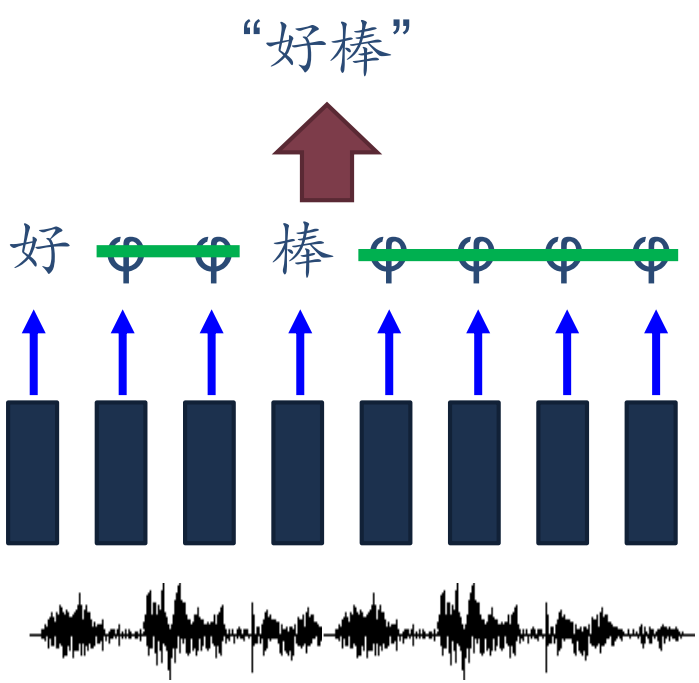
- 序列标注

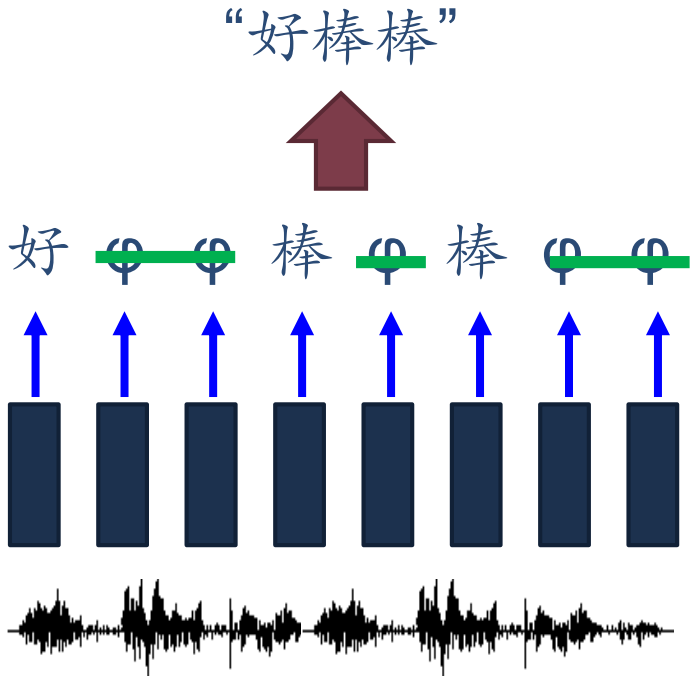


语音识别

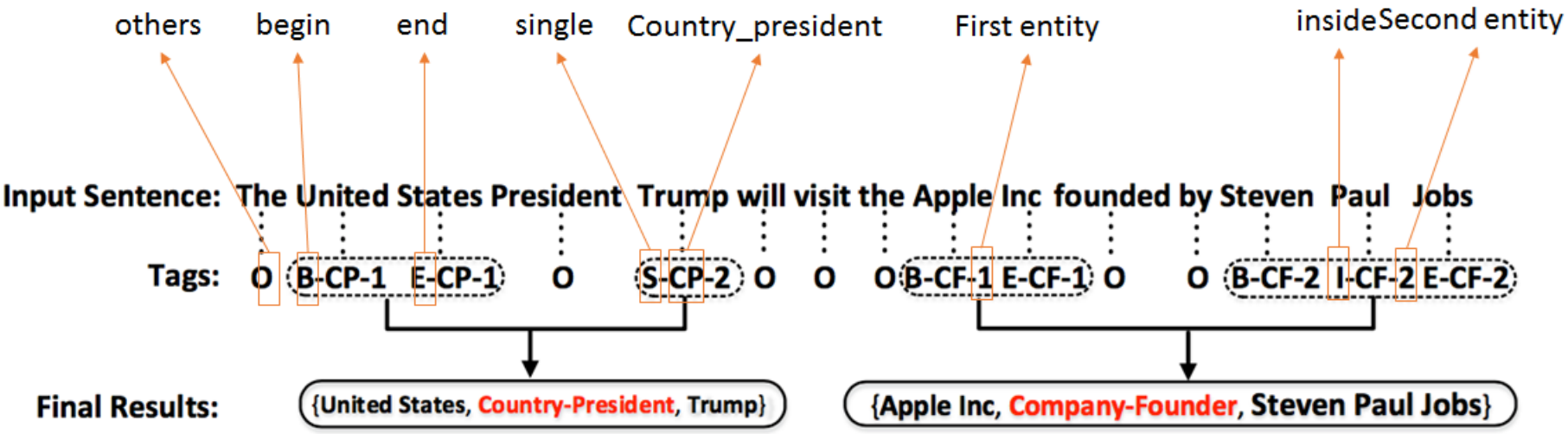
来源：李宏毅《1天搞懂深度学习》

- Connectionist Temporal Classification (CTC) [Alex Graves, ICML’06][Alex Graves, ICML’14][Haşim Sak, Interspeech’15][Jie Li, Interspeech’15][Andrew Senior, ASRU’15]





信息抽取



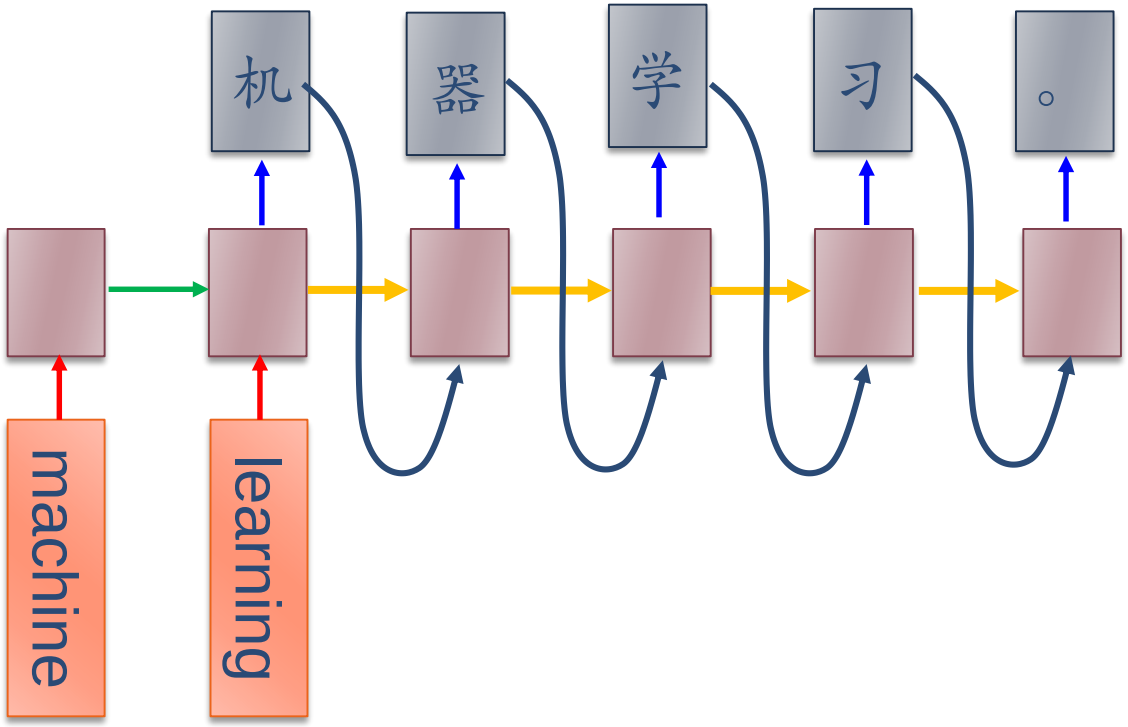
Number of tags: $2 * 4 * |R| + 1$

$|R|$ is the number of relation, 4 means begin, end, single, inside

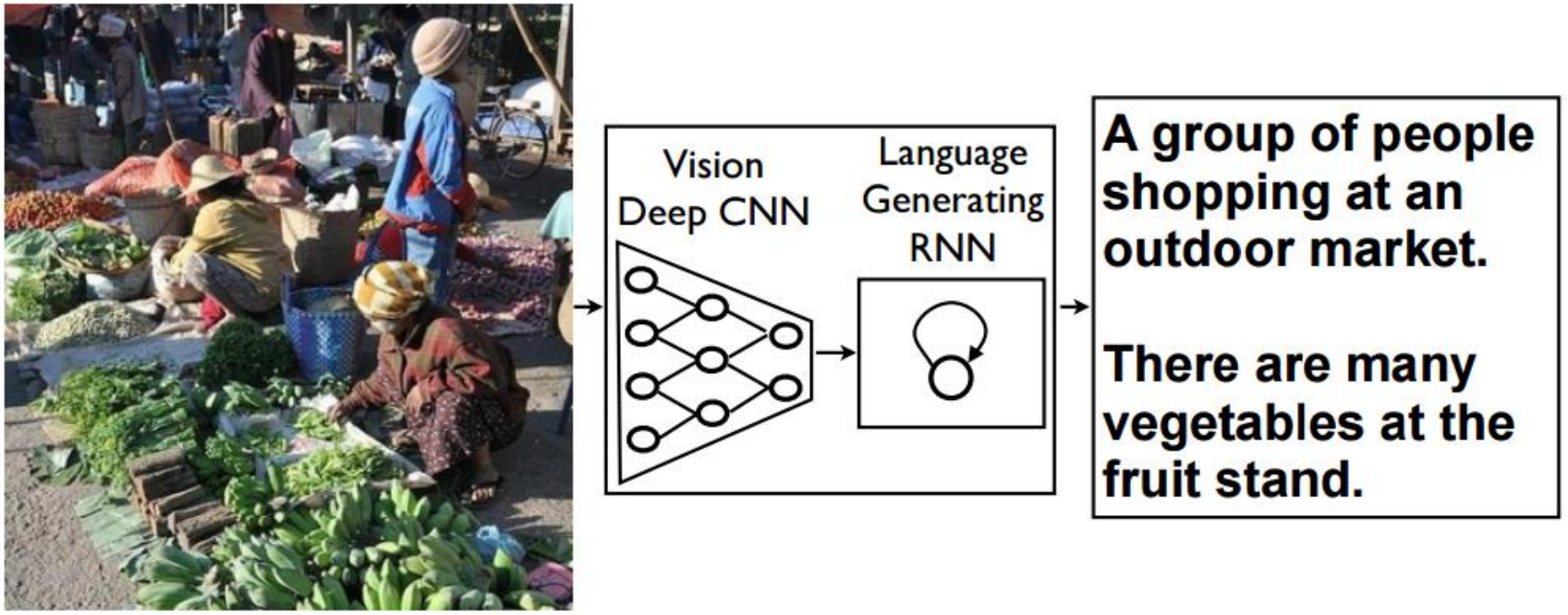
异步的序列到序列模式

来源：李宏毅《1天搞懂深度学习》

- 机器翻译

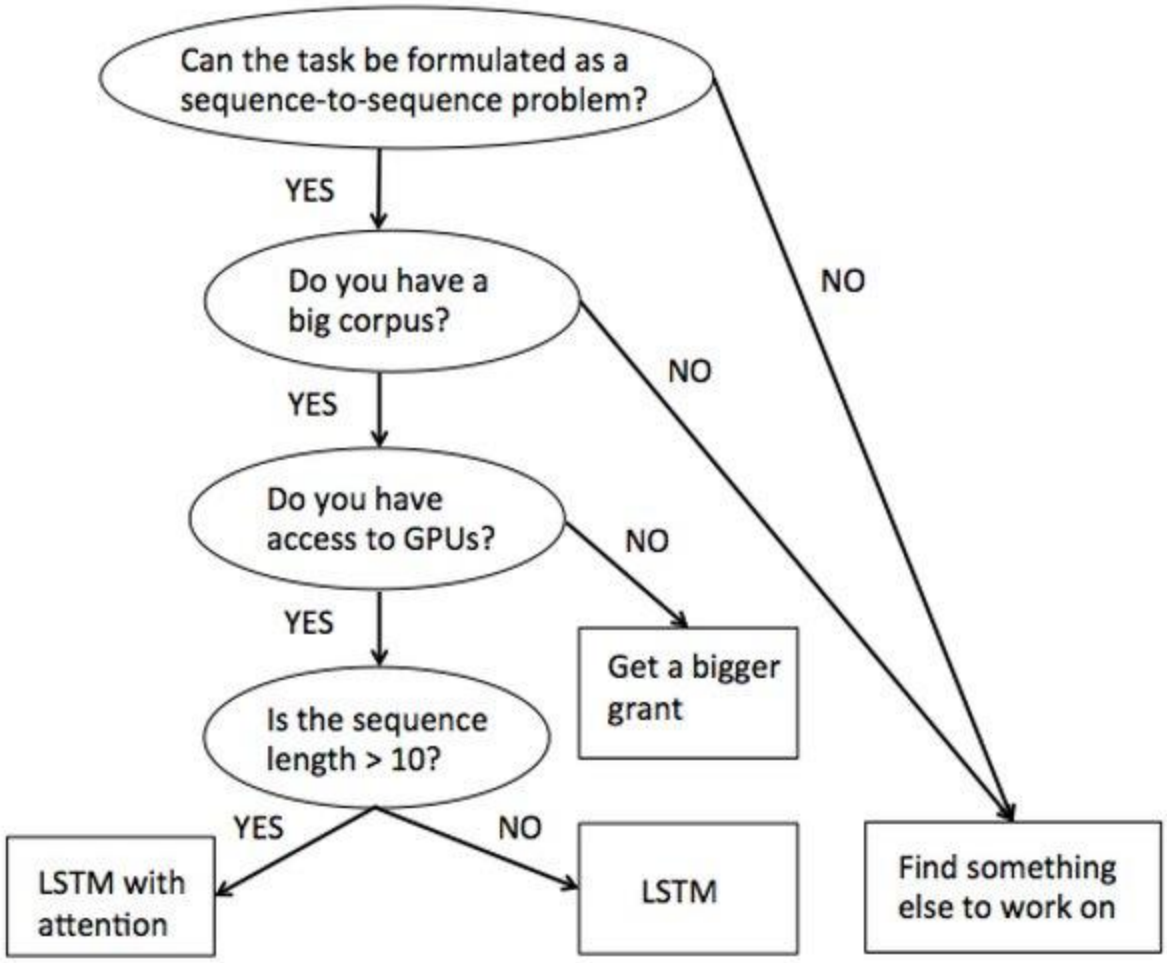


看图说话


Vision
Deep CNN
Language
Generating
RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

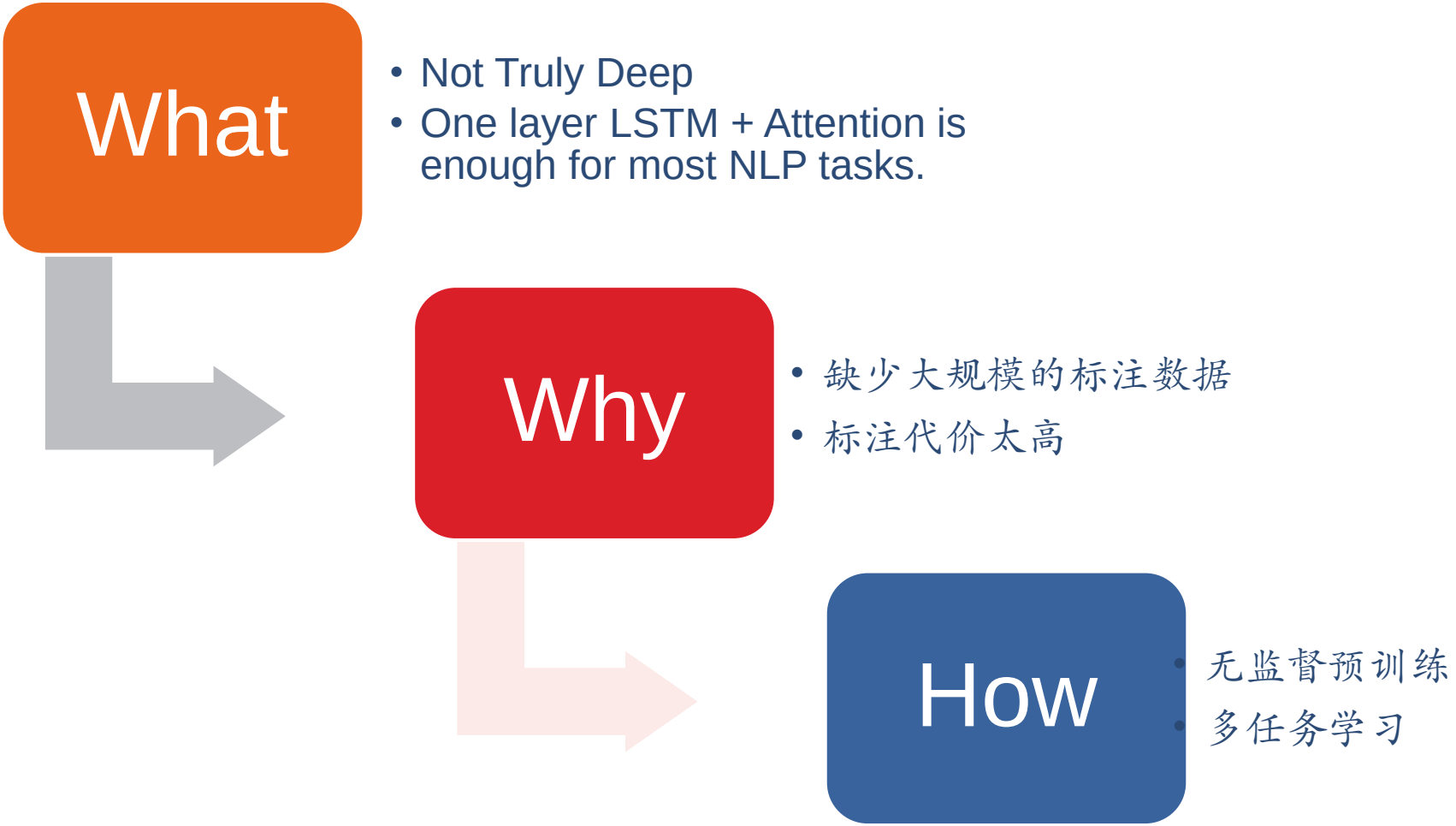
One Size fits ALL!

One Size fits ALL!



深度学习在自然语言处理中的困境

深度学习在自然语言处理中的“困境”



无监督预训练

无监督预训练

- 词级别
  - 语言模型
  - Word2Vec （CBOW and Skip-Gram）
  - GLOVE
  - FastText
  - ELMo: Embeddings from Language Models

- 句子级别
  - Skip-Thought
  - Paragraph Vector
  - …

NAACL 2018 Best Paper

Deep contextualized word representations

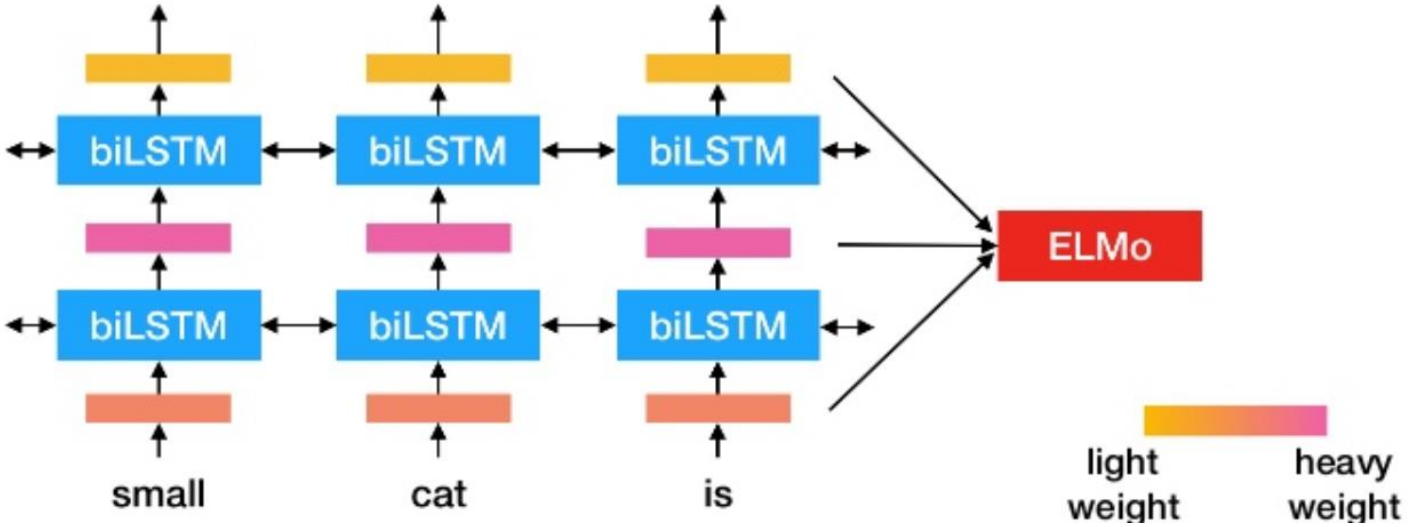
Matthew E. Peters[†], **Mark Neumann**[†], **Mohit Iyyer**[†], **Matt Gardner**[†],
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*], **Kenton Lee**[*], **Luke Zettlemoyer**[†*]
{csquared,kentonl,lsz}@cs.washington.edu

[†]Allen Institute for Artificial Intelligence
[*]Paul G. Allen School of Computer Science & Engineering, University of Washington

ELMo: Embeddings from Language Models

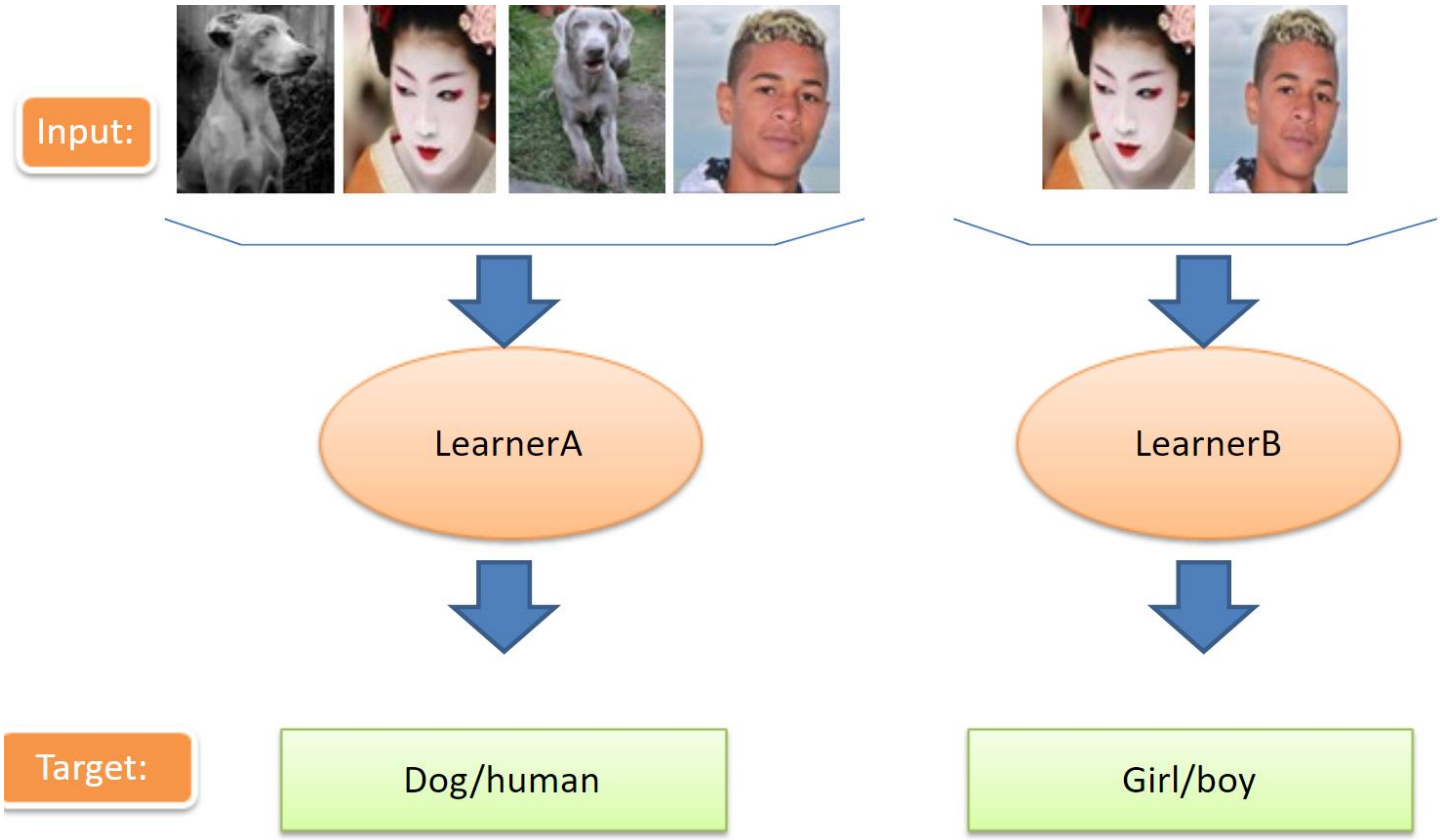
$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}.$$



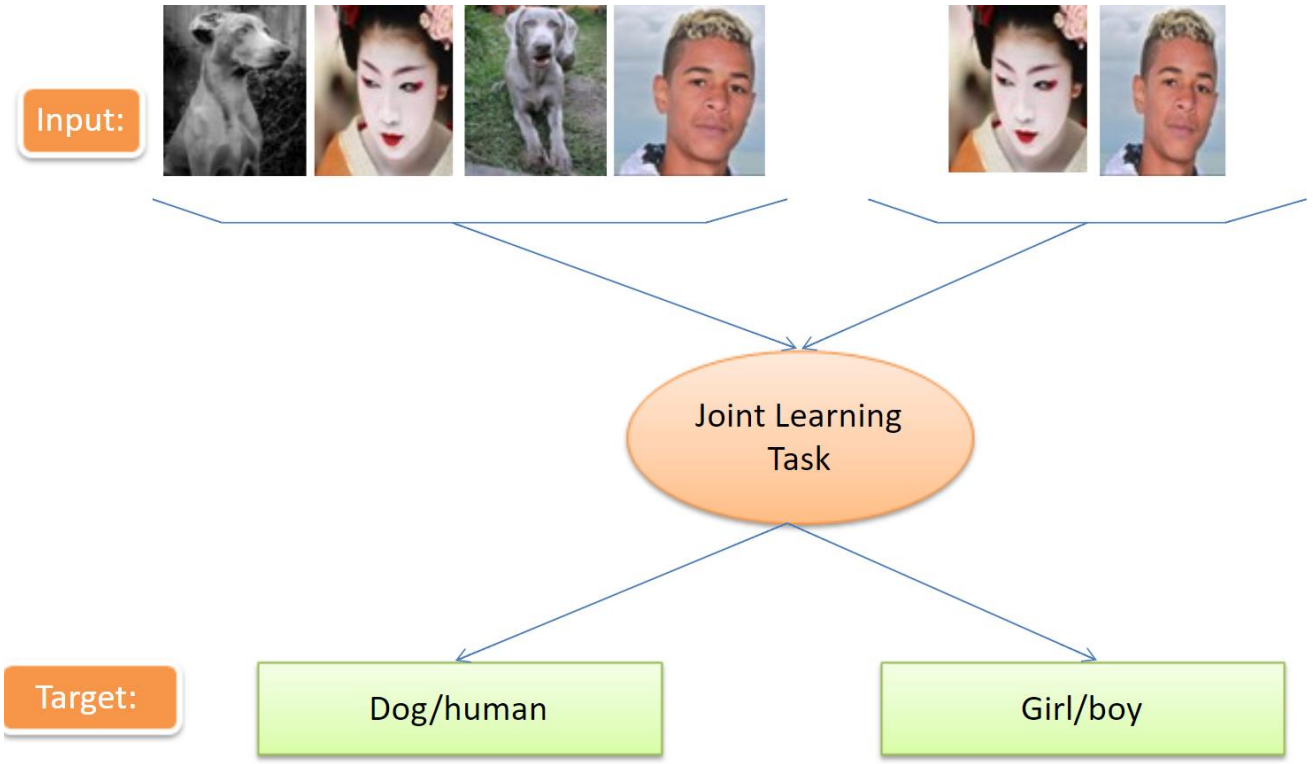
https://arxiv.org/pdf/1802.05365.pdf

多任务学习

单任务学习



多任务学习 Multitask Learning



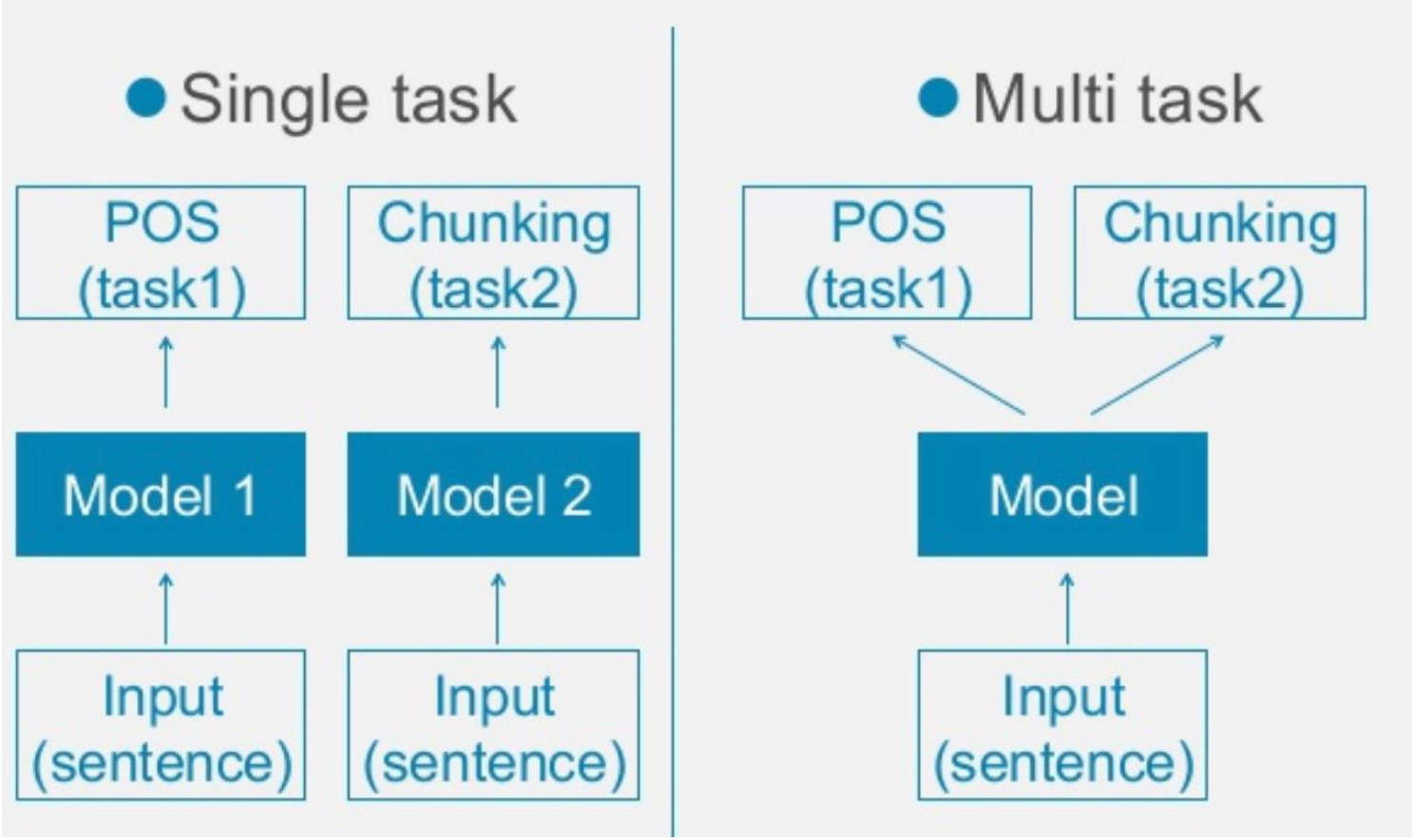
Machine Learning 1997

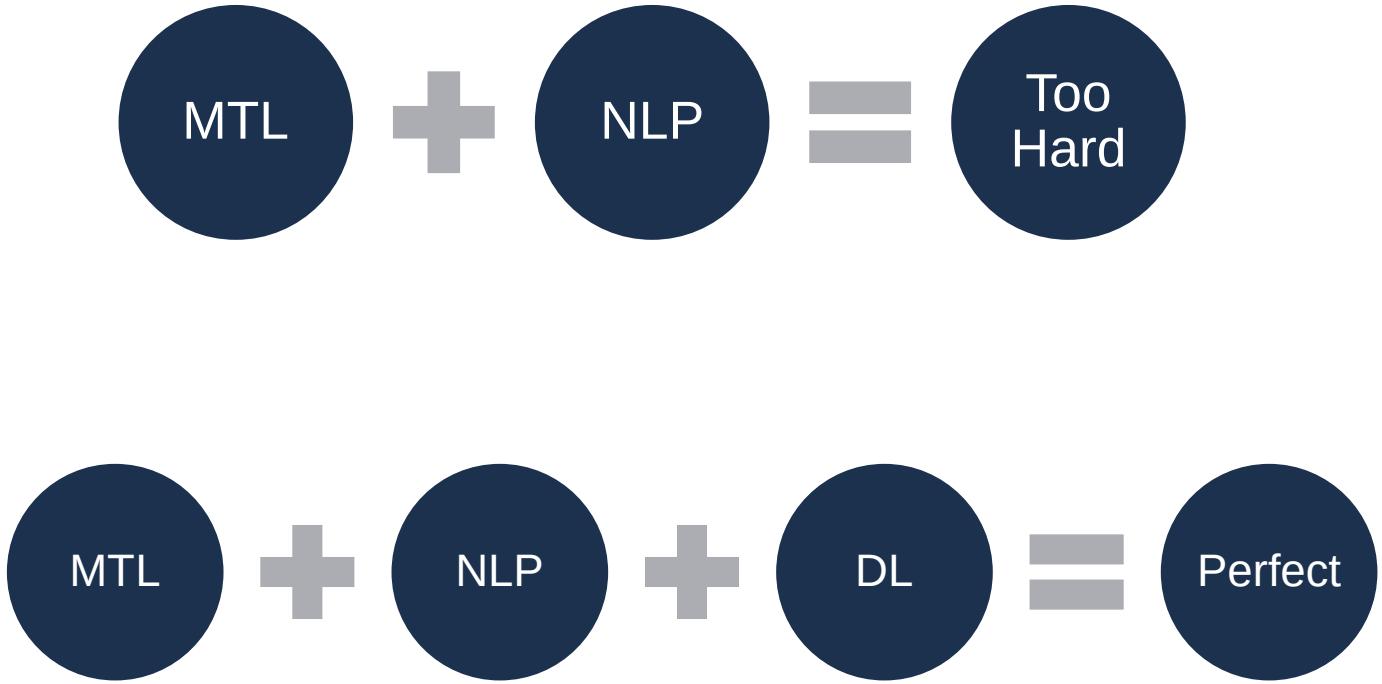
Multitask Learning*

RICH CARUANA caruana@cs.cmu.edu
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

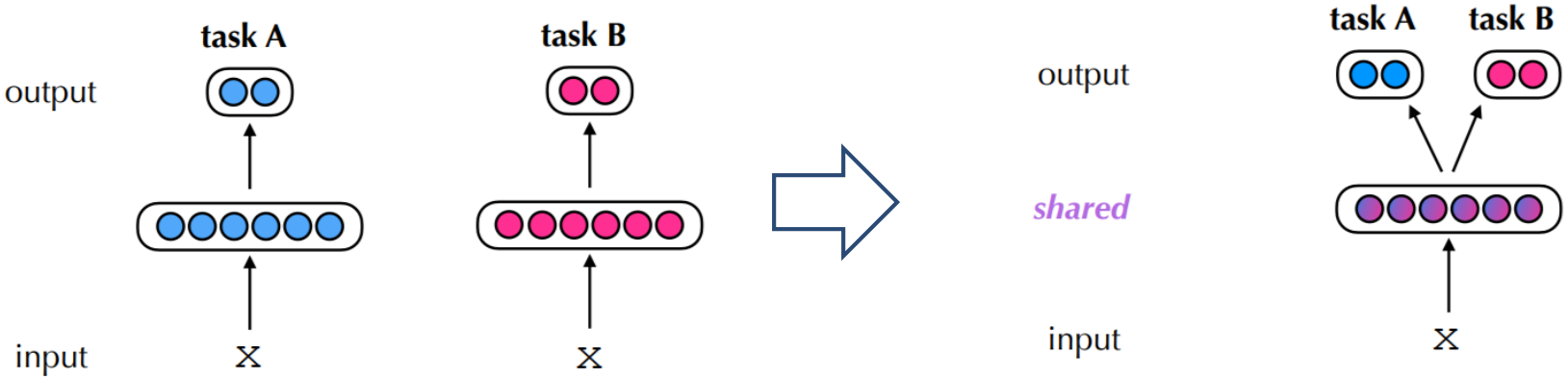
Multitask Learning is an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias. It does this by learning tasks in parallel while using a shared representation; what is learned for each task can help other tasks be learned better.

A NLP Example



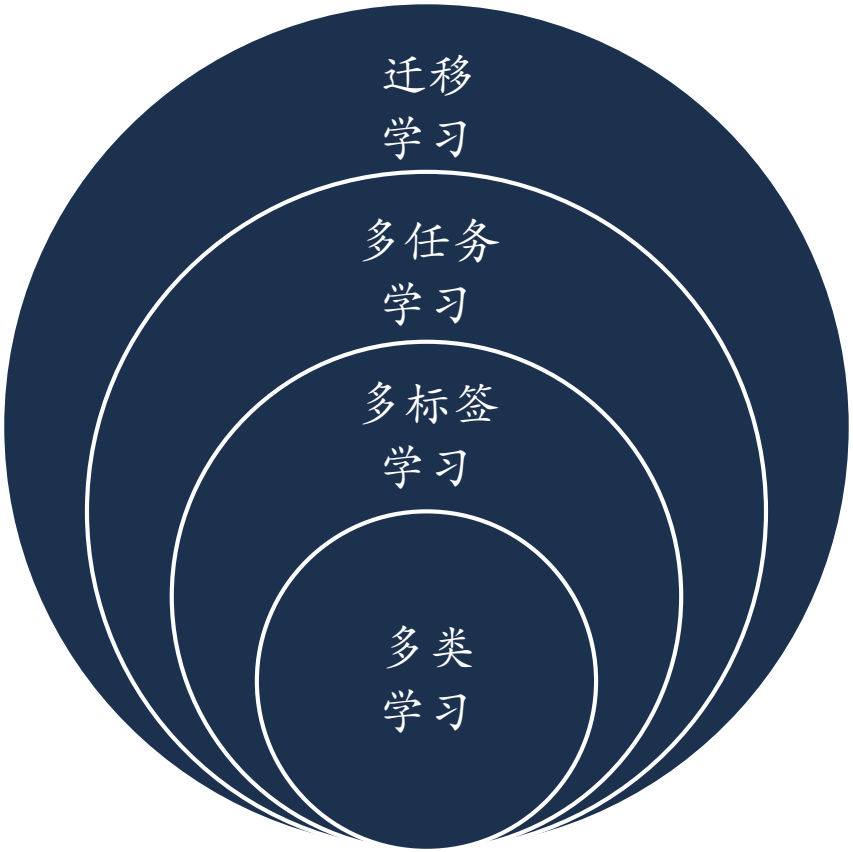

MTL
NLP
Too Hard
MTL
NLP
DL
Perfect

多任务学习+深度学习



Neural network based approaches make MTL particulary attractive/easy

不同学习范式之间的关系



- 迁移学习 Transfer Learning
  - 在源领域上学习模型
  - 泛化到目标领域上
- 多任务学习 Multi-Task Learning
  - 同时建模多个相关任务
  - 不同任务有不同的数据和标签
- 多标签学习 Multi-Label Learning
  - 一个样本可以有多个标签
  - 建模标签之间的关系
- 多类学习 Multi-Class Learning
  - 一个样本只能属于一个标签

损失函数

- 损失函数

$$-\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k}\left(y_j^{(i)}\log\hat{y}_j^{(i)}+\left(1-y_j^{(i)}\right)\log\left(1-\hat{y}_j^{(i)}\right)\right)$$

Joint Losses?

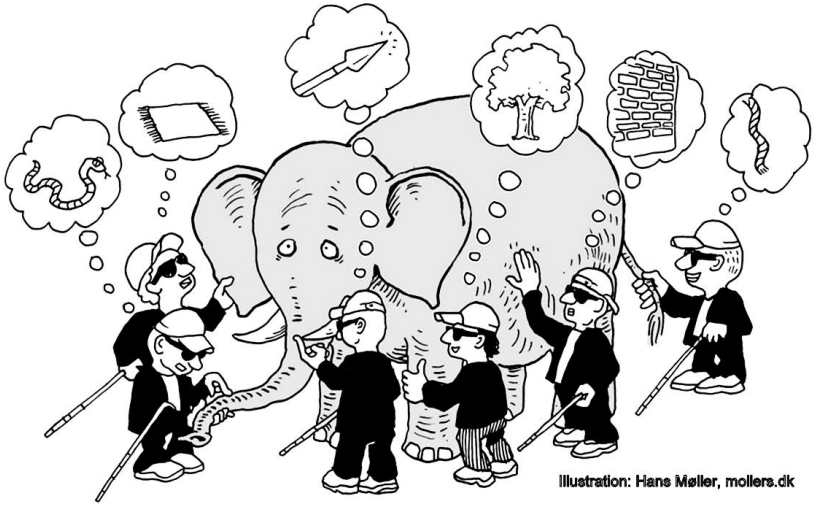
训练方式

- Joint Training: The training is achieved in a stochastic manner by looping over the tasks:
  - 1. Select a random task.
  - 2. Select a random training example from this task.
  - 3. Update the parameters for this task by taking a gradient step with respect to this example.
  - 4. Go to 1.

- Fine Tuning: After the joint learning phase, we can use a fine tuning strategy to further optimize the performance for each task.

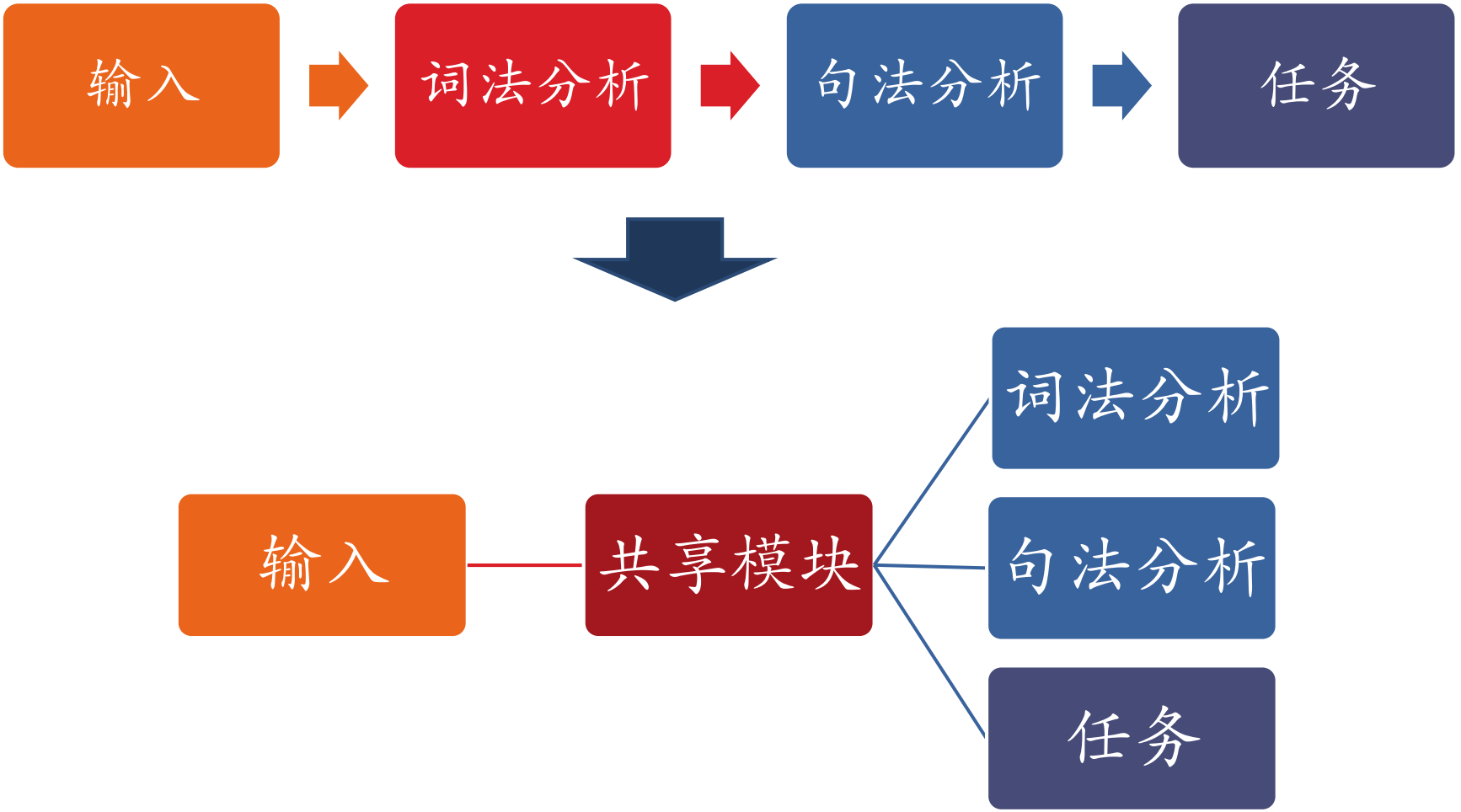
Why does MTL work?

- 隐式的数据增强
- 更好的表示学习
  - 一个好的表示需要能够提高多个任务的性能。
- 正则化
  - 共享参数在一定程度上弱化了网络能力，防止过拟合
- Eavesdropping（窃听）



自然语言处理中的多任务学习

思维方式的改变



自然语言中的多任务学习

多领域（Multi-Doman）任务

- Multi-Domain Text Classification
- Multi-Domain Sentiment Analysis

多级（Multi-Level）任务

- part-of-speech (POS)
- tagging, named entity recognition (NER)
- semantic role labeling (SRL)

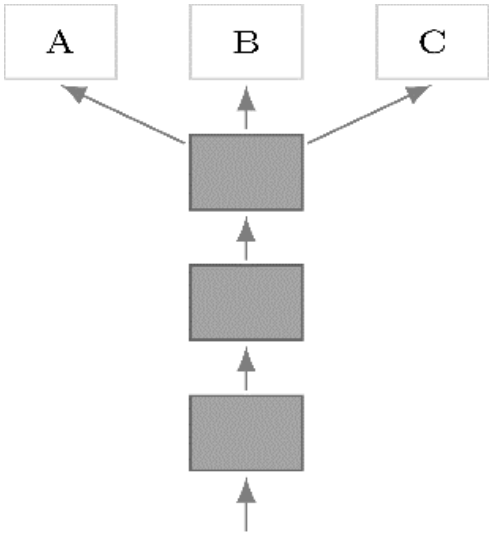
多语言（Multi-Linguistic）任务

- Machine translation
- Multi-lingual parsing

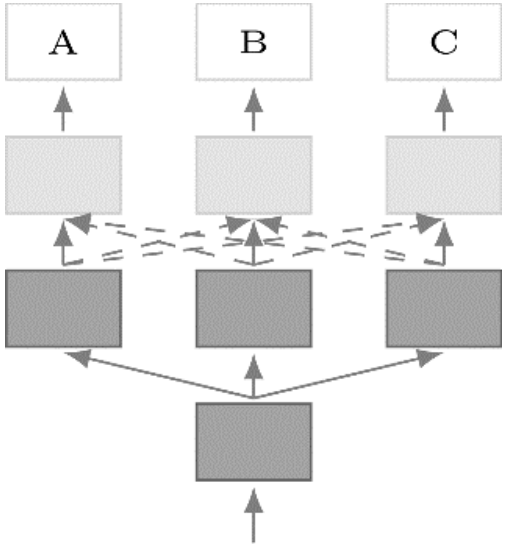
多模态（Multi-Modality）任务

- Visual QA
- Image Caption

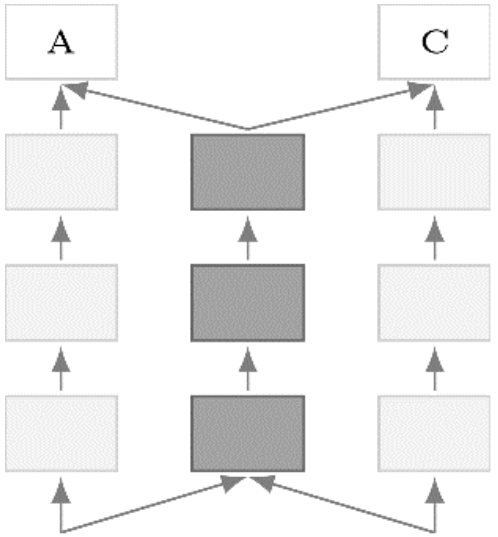
深度学习+多任务学习



(a) 硬共享模式



(b) 软共享模式



(c) 共享-私有模式

共享模式

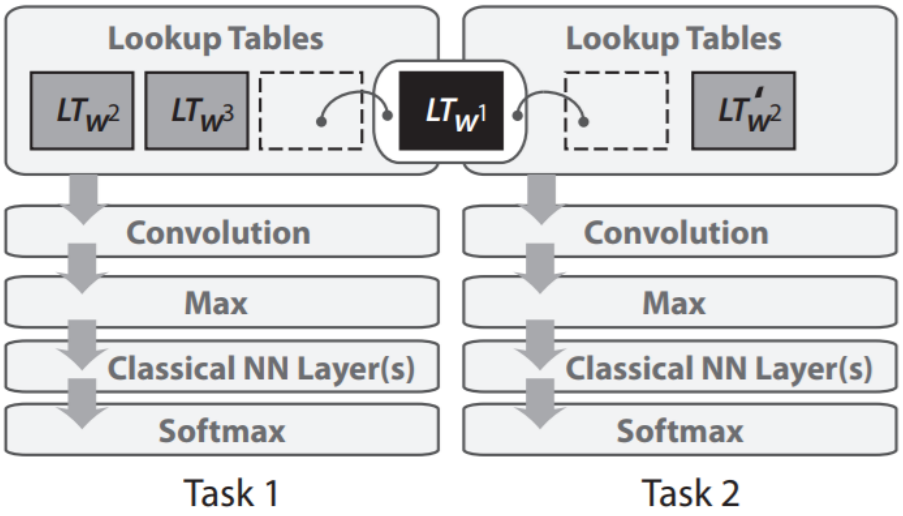
- 硬共享模式
- 软共享模式
- 共享私有模式
- 多级共享模式
- 函数共享模式
- 主辅共享模式
- …

硬共享模式

ICML 2008

A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning

Ronan Collobert COLLOBER@NEC-LABS.COM
Jason Weston JASONW@NEC-LABS.COM
NEC Labs America, 4 Independence Way, Princeton, NJ 08540 USA

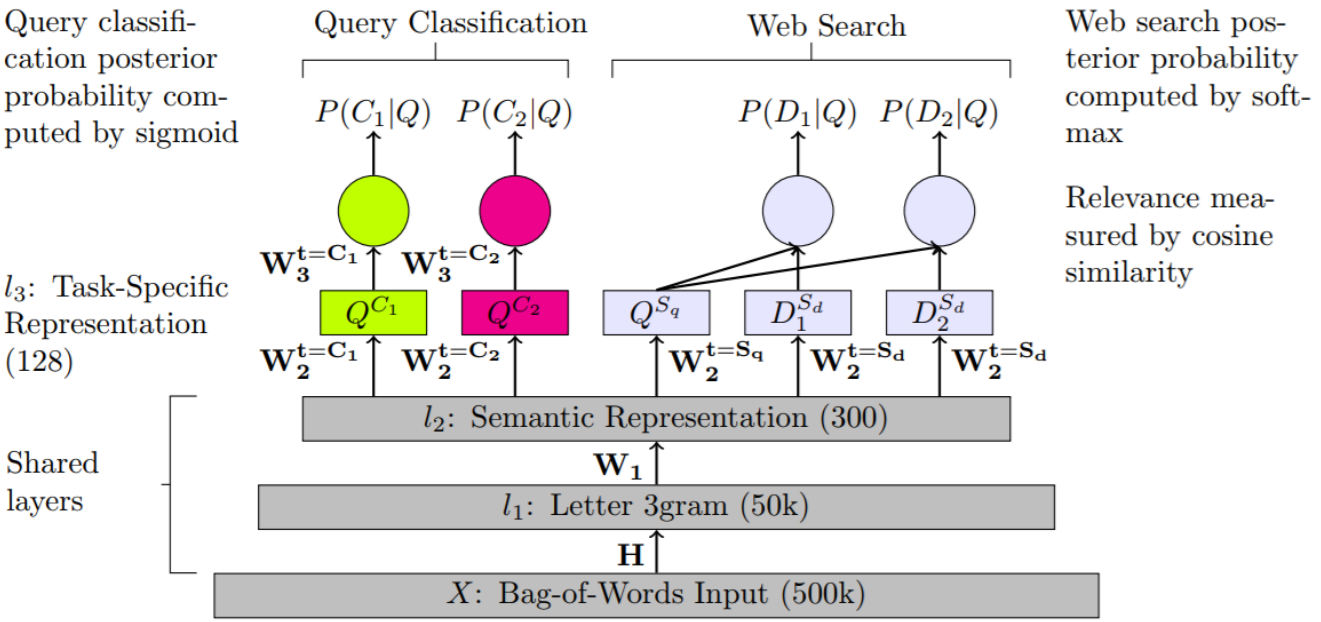


	wsz=15	wsz=50	wsz=100
SRL	16.54	17.33	18.40
SRL + POS	15.99	16.57	16.53
SRL + Chunking	16.42	16.39	16.48
SRL + NER	16.67	17.29	17.21
SRL + Synonyms	15.46	15.17	15.17
SRL + Language model	14.42	14.30	14.46
SRL + POS + Chunking	16.46	15.95	16.41
SRL + POS + NER	16.45	16.89	16.29
SRL + POS + Chunking +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

Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval

NAACL 2015

Xiaodong Liu[†*], **Jianfeng Gao**[‡], **Xiaodong He**[‡], **Li Deng**[‡], **Kevin Duh**[†] **and Ye-yi Wang**[‡]
[†]Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan
[‡]Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

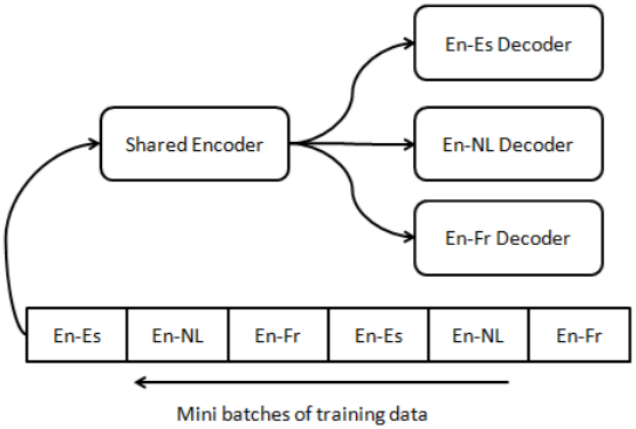


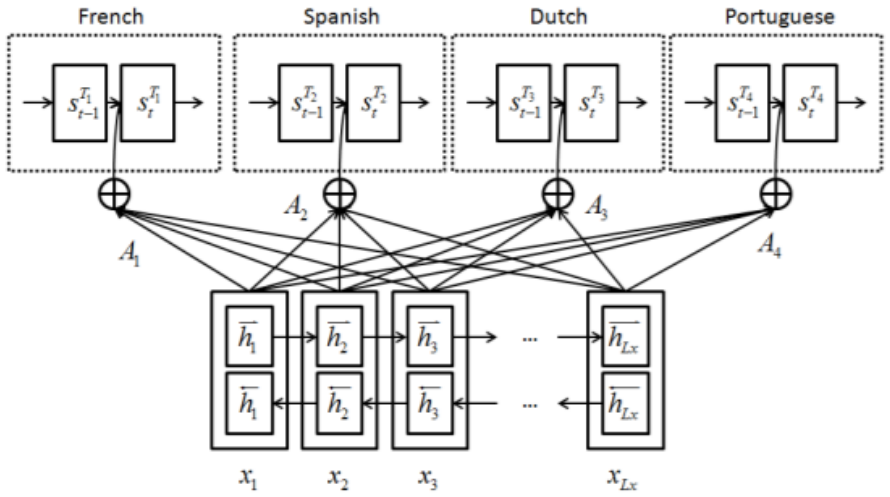
http://www.aclweb.org/anthology/N15-1092

Multi-Task Learning for Multiple Language Translation

ACL 2015

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu and Haifeng Wang
Baidu Inc, Beijing, China



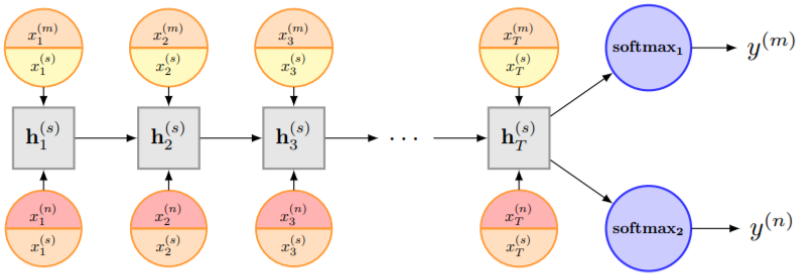


end to multi-end model

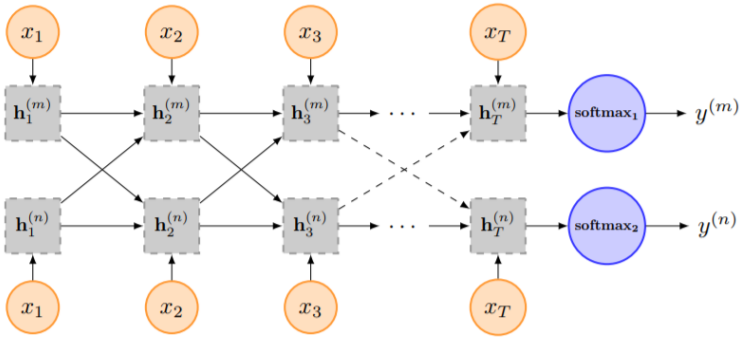
IJCAI 2016

Recurrent Neural Network for Text Classification with Multi-Task Learning

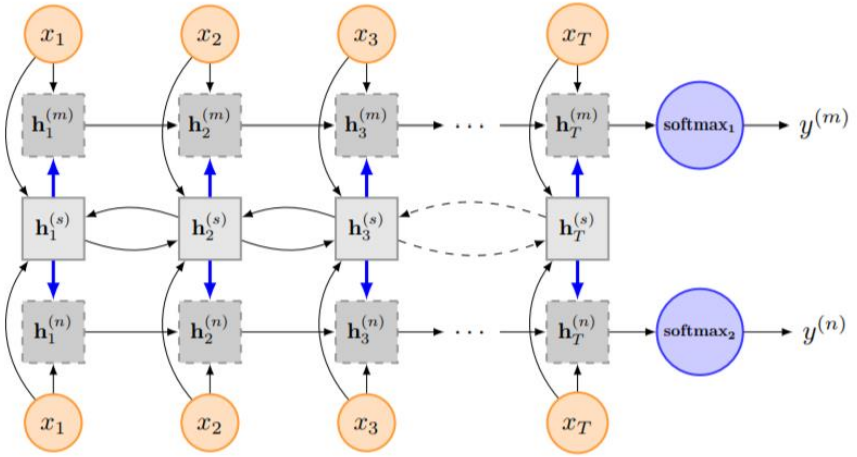
Pengfei Liu **Xipeng Qiu*** **Xuanjing Huang**
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University



(a) Model-I: Uniform-Layer Architecture



(b) Model-II: Coupled-Layer Architecture



(c) Model-III: Shared-Layer Architecture

https://arxiv.org/pdf/1605.05101.pdf

IJCAI 2018

Same Representation, Different Attentions: Shareable Sentence Representation Learning from Multiple Tasks

Renjie Zheng, Junkun Chen, Xipeng Qiu*
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University

The infantile cart is easy to use,

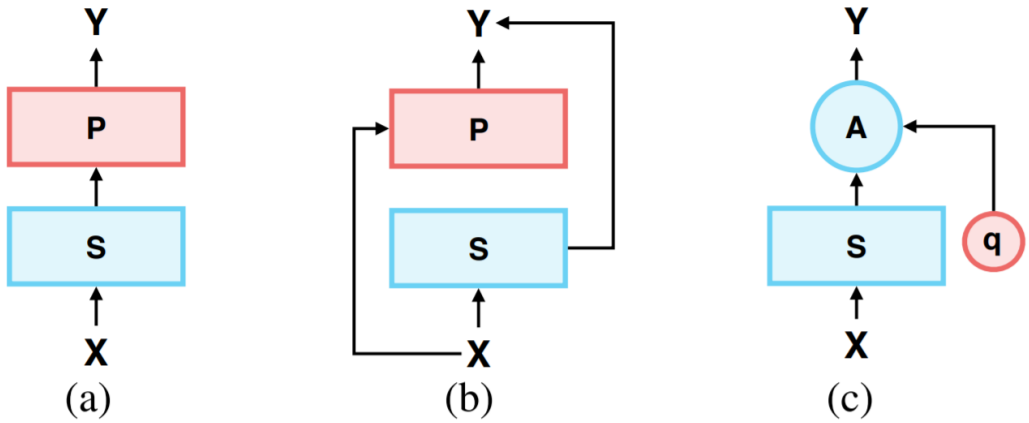


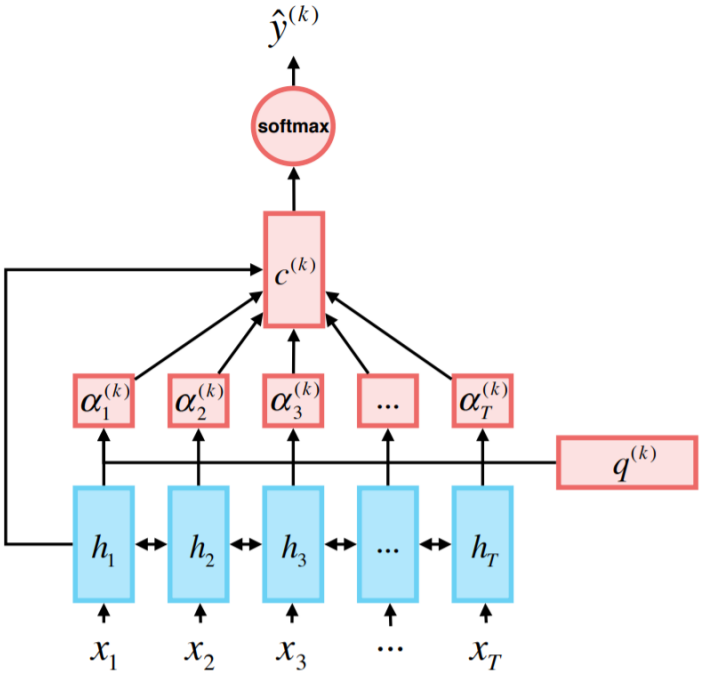
Figure 1: Three schemes of information sharing in multi-task leaning. (a) stacked shared-private scheme, (b) parallel shared-private scheme, (c) our proposed attentive sharing scheme.

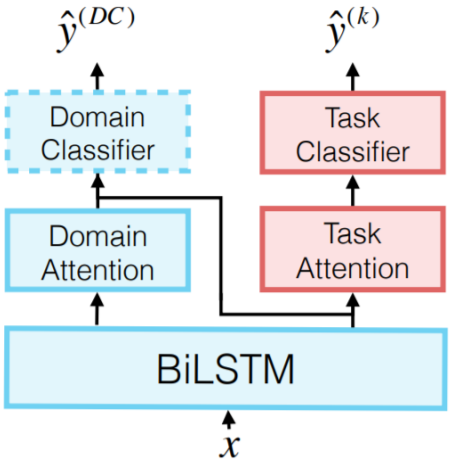
https://arxiv.org/pdf/1804.08139.pdf

IJCAI 2018

Same Representation, Different Attentions: Shareable Sentence Representation Learning from Multiple Tasks

Renjie Zheng, Junkun Chen, Xipeng Qiu*
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University





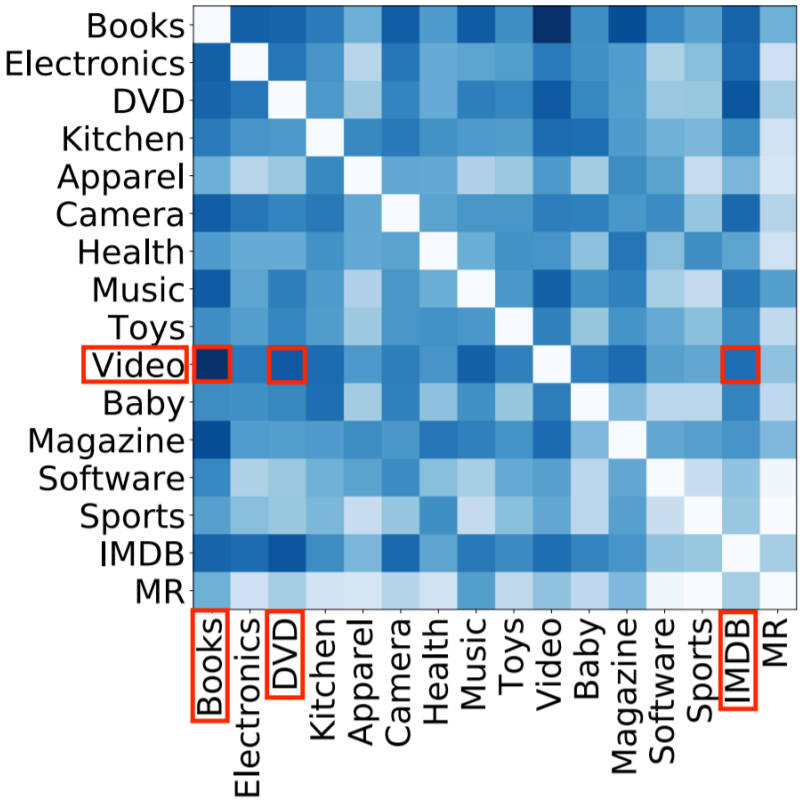
共享表示，不同注意力选择

https://arxiv.org/pdf/1804.08139.pdf

IJCAI 2018

Same Representation, Different Attentions: Shareable Sentence Representation Learning from Multiple Tasks

Renjie Zheng, Junkun Chen, Xipeng Qiu*
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University

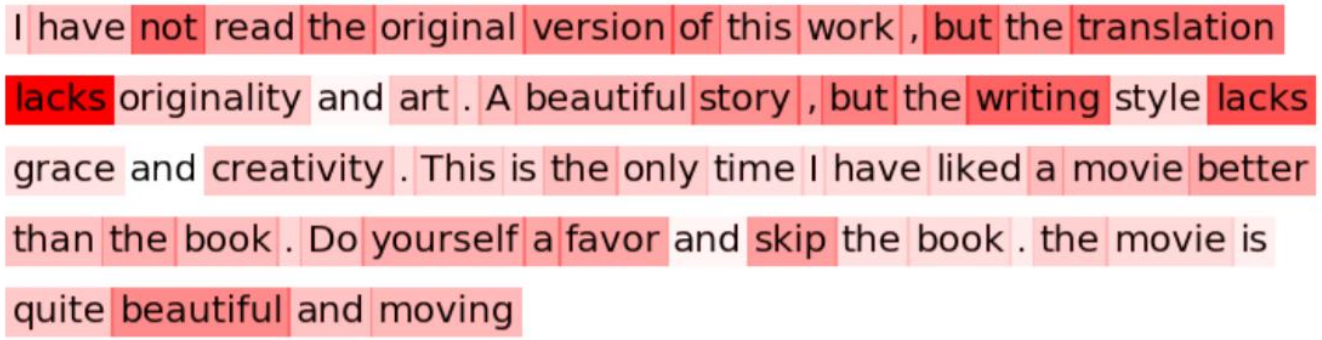


Similarity Matrix of Different Task's query vector

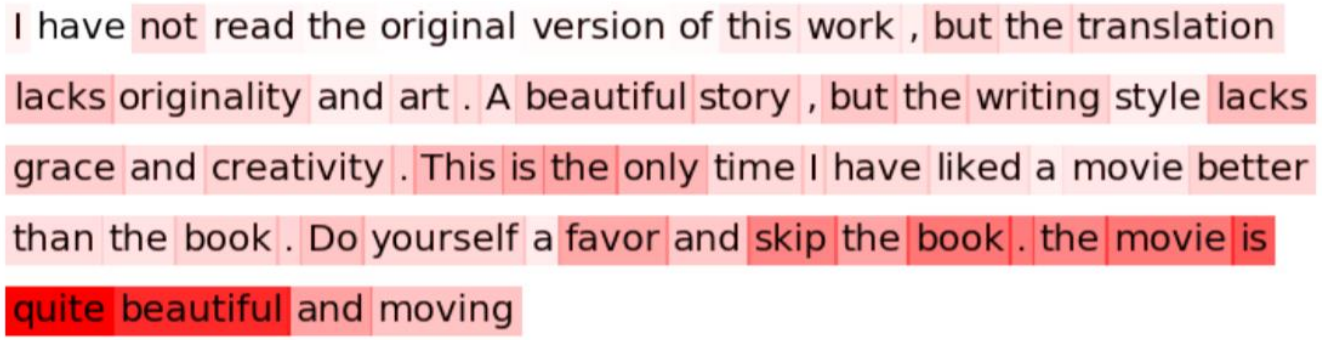
IJCAI 2018

Same Representation, Different Attentions: Shareable Sentence Representation Learning from Multiple Tasks

Renjie Zheng, Junkun Chen, Xipeng Qiu*
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University



(a) Attention of task "Books" in SA-MTL, Output: Negative

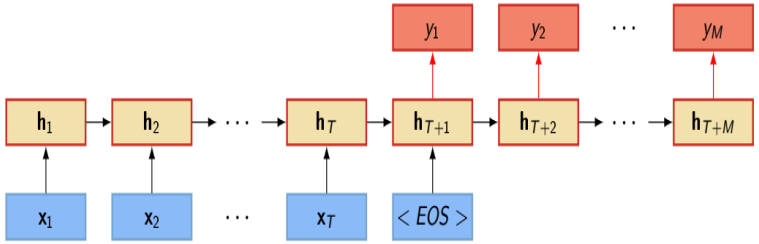


(b) Attention of task "DVD" in SA-MTL, Output: Positive

ICLR 2018

LEARNING GENERAL PURPOSE DISTRIBUTED SENTENCE REPRESENTATIONS VIA LARGE SCALE MULTI-TASK LEARNING

Sandeep Subramanian[1,2,3]*, **Adam Trischler**[3], **Yoshua Bengio**[1,2,4] **& Christopher J Pal**[1,5]
[1] Montréal Institute for Learning Algorithms (MILA)
[2] Université de Montréal
[3] Microsoft Research Montreal
[4] CIFAR Senior Fellow
[5] École Polytechnique de Montréal



MULTI-TASK SEQUENCE-TO-SEQUENCE LEARNING

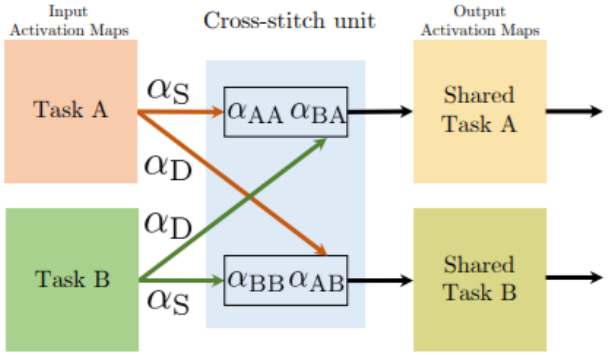
Task	Sentence Pairs
En-Fr (WMT14)	40M
En-De (WMT15)	5M
Skipthought (BookCorpus)	74M
AllNLI (SNLI + MultiNLI)	1M
Parsing (PTB + 1-billion word)	4M
Total	124M

软共享模式

CVPR 2016 (Spotlight)

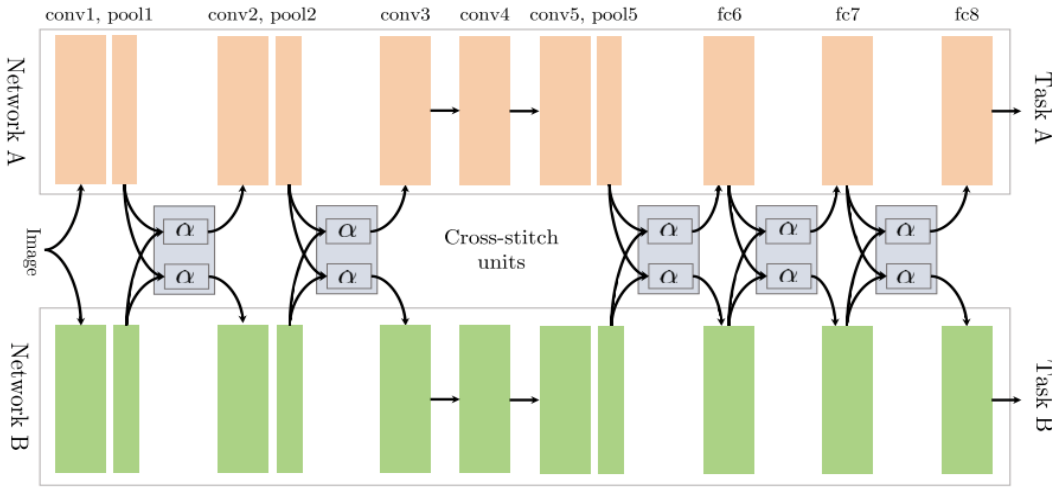
Cross-stitch Networks for Multi-task Learning

Ishan Misra* Abhinav Shrivastava* Abhinav Gupta Martial Hebert
The Robotics Institute, Carnegie Mellon University



cross-stitch

$$\begin{bmatrix} \tilde{x}_{\mathrm{A}}^{ij} \\ \tilde{x}_{\mathrm{B}}^{ij} \end{bmatrix} = \begin{bmatrix} \alpha_{\mathrm{AA}} & \alpha_{\mathrm{AB}} \\ \alpha_{BA} & \alpha_{\mathrm{BB}} \end{bmatrix} \begin{bmatrix} x_{\mathrm{A}}^{ij} \\ x_{\mathrm{B}}^{ij} \end{bmatrix}$$

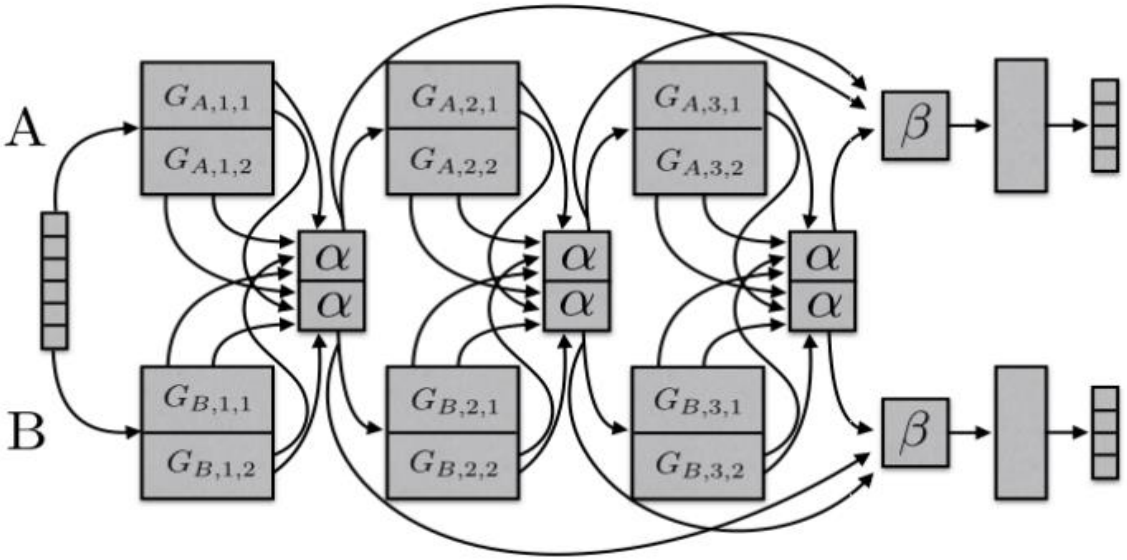


https://arxiv.org/pdf/1604.03539.pdf

arXiv preprint arXiv:1705.08142

Learning what to share between loosely related tasks

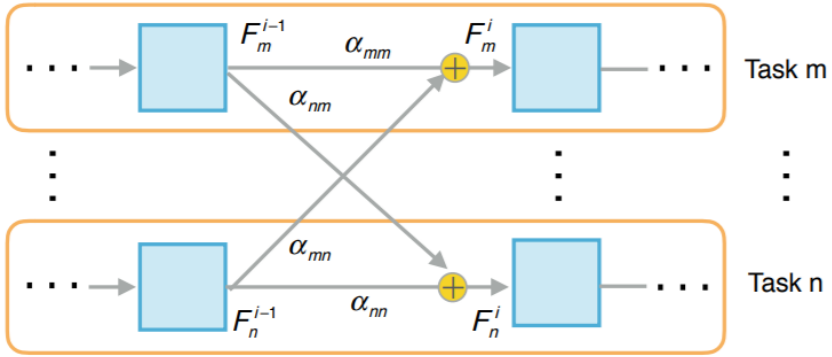
Sebastian Ruder[12]*, Joachim Bingel[3], Isabelle Augenstein[4]*, Anders Søgaard[3]
[1]Insight Research Centre, National University of Ireland, Galway
[2]Aylien Ltd., Dublin, Ireland
[3]Department of Computer Science, University of Copenhagen, Denmark
[4]Department of Computer Science, UCL, UK



Sluice networks

Learning What to Share: Leaky Multi-Task Network for Text Classification

Liqiang Xiao[1,2], **Honglun Zhang**[1,2], **Wenqing Chen**[1,2], **Yongkun Wang**[3], **Yaohui Jin**[1,2]
[1] State Key Lab of Advanced Optical Communication System and Network,
Shanghai Jiao Tong University



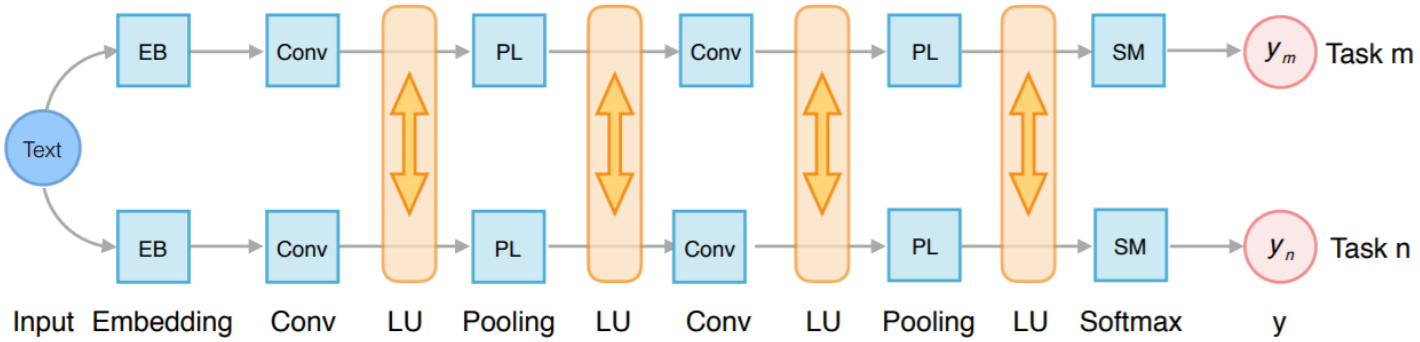
$$\begin{bmatrix} \mathbf{F}_1^i \\ \vdots \\ \mathbf{F}_M^i \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1M} \\ \vdots & \ddots & \vdots \\ \alpha_{M1} & \cdots & \alpha_{MM} \end{bmatrix} \begin{bmatrix} \mathbf{F}_1^{i-1} \\ \vdots \\ \mathbf{F}_M^{i-1} \end{bmatrix}$$

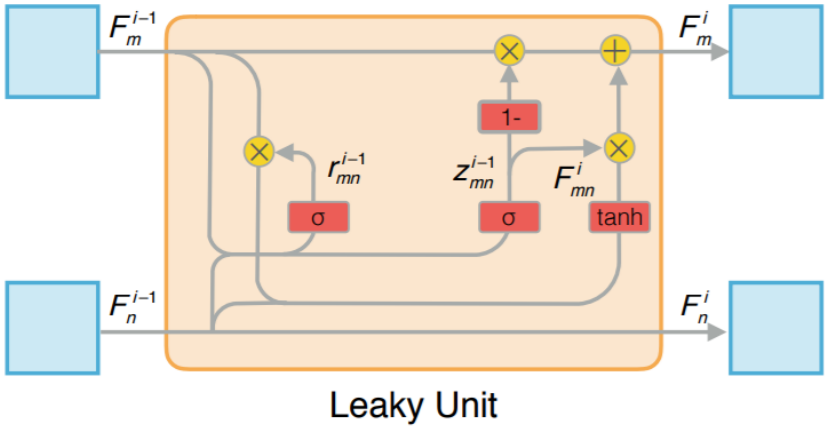
α_{i-1} is updated by back-propagation, which reflects the strength of association between tasks but has no selection for the features.

http://www.aclweb.org/anthology/C18-1175

Learning What to Share: Leaky Multi-Task Network for Text Classification

Liqiang Xiao[1,2], **Honglun Zhang**[1,2], **Wenqing Chen**[1,2], **Yongkun Wang**[3], **Yaohui Jin**[1,2]
[1] State Key Lab of Advanced Optical Communication System and Network,
Shanghai Jiao Tong University





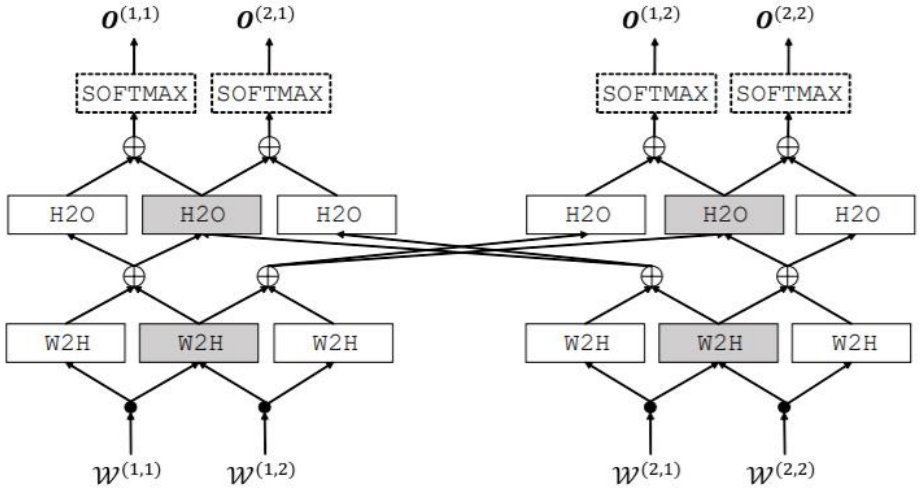
Leaky Unit

COLING 2018

Multi-task and Multi-lingual Joint Learning of Neural Lexical Utterance Classification based on Partially-shared Modeling

Ryo Masumura, Tomohiro Tanaka, Ryuichiro Higashinaka, Hirokazu Masataki and Yushi Aono
NTT Media Intelligence Laboratories, NTT Corporation, Japan

Task	Utterance	Label
DA	Hello, how are you today?	GREETING
	I am so sorry to hear of your son's accident.	SYMPATHY/AGREE
	Lets go to school an hour early today.	PROPOSAL
ENE	What is the highest mountain in the world?	MOUNTAIN
	Who is president of the united states?	PERSON
	What is the name of the most recent Star Wars movie?	MOVIE
QT	Do you like egg salad?	TRUE/FALSE
	How do you correct a hook in a golf swing?	EXPLANATION:METHOD
	Why is blood red?	EXPLANATION:CAUSE



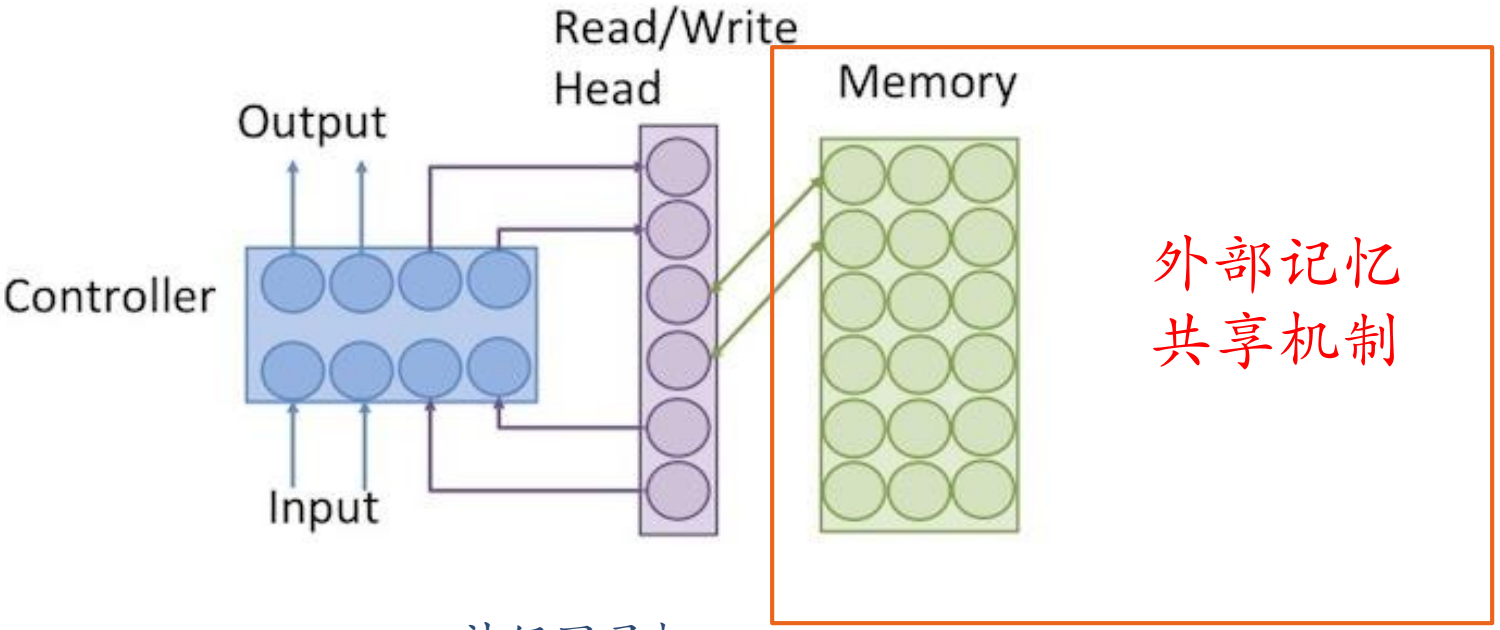
Multi-task and multi-lingual partially-shared modeling

http://www.aclweb.org/anthology/C18-1304

共享-私有模式

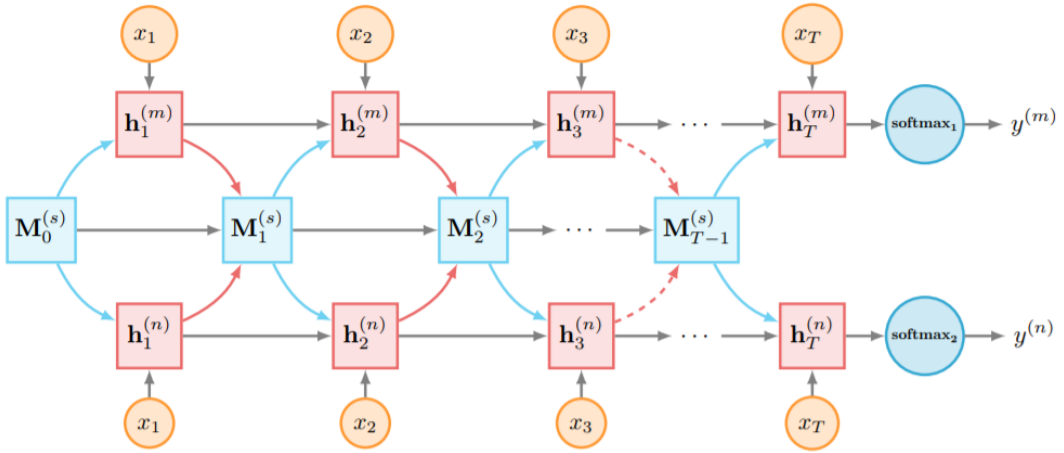
Deep Multi-Task Learning with Shared Memory

Pengfei Liu Xipeng Qiu* Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University

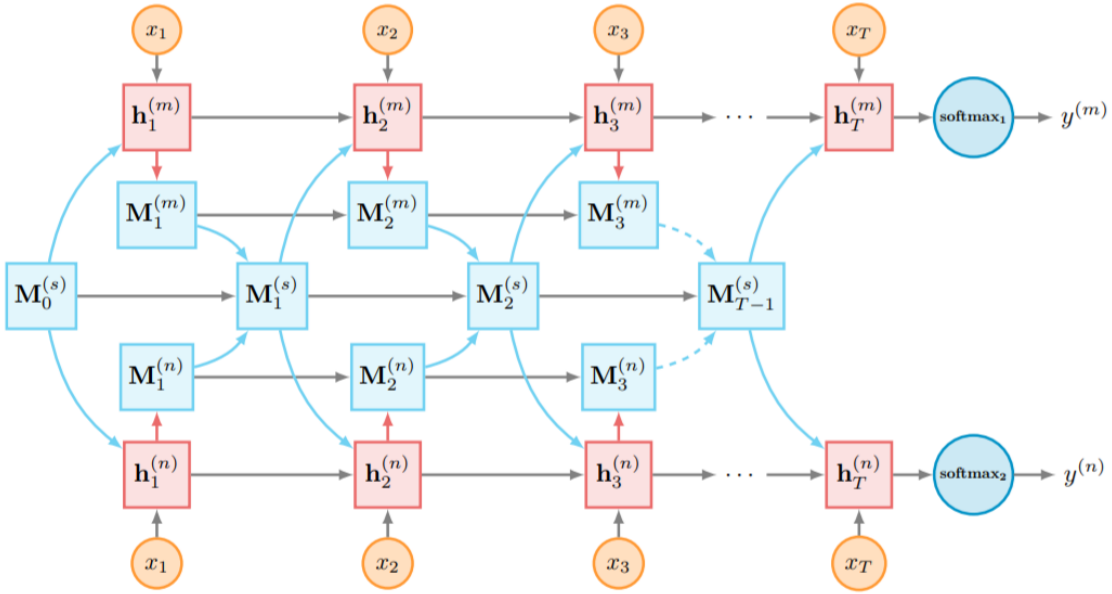


神经图灵机

https://aclweb.org/anthology/D16-1012



(a) Global Memory Architecture



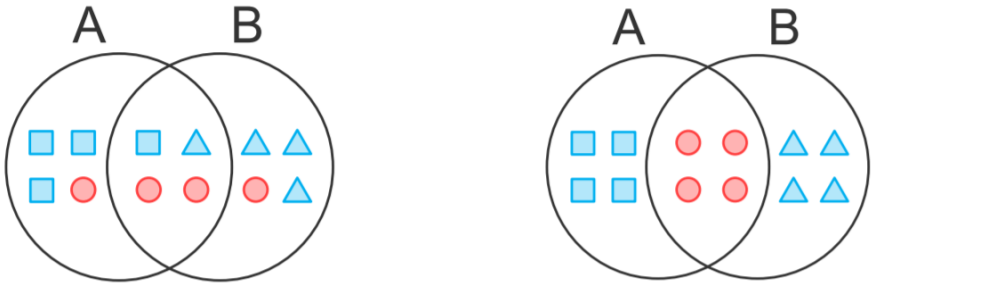
(b) Local-Global Hybrid Memory Architecture

ACL 2017

Adversarial Multi-task Learning for Text Classification

Pengfei Liu Xipeng Qiu Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University

The infantile cart is simple and easy to use.
This kind of humour is infantile and boring.

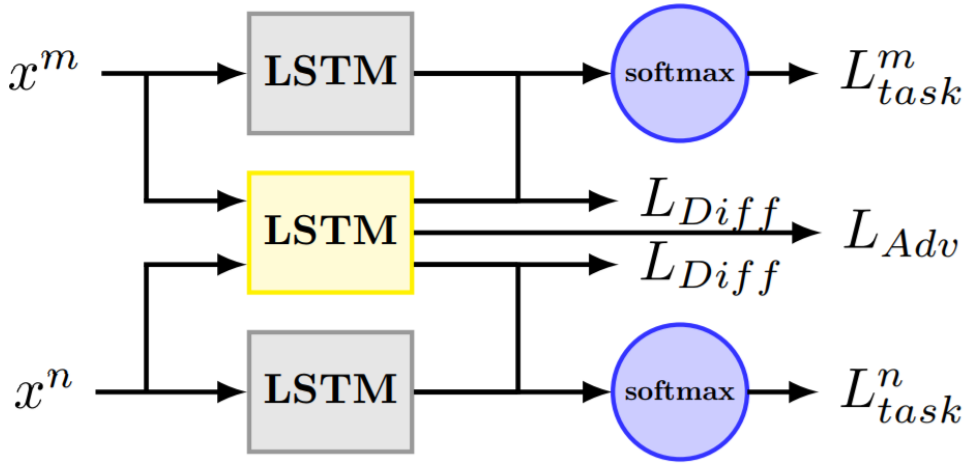


(a) Shared-Private Model (b) Adversarial Shared-Private Model

Two sharing schemes for task A and task B.
The overlap between two black circles denotes shared space.
The blue triangles and boxes represent the task-specific features
The red circles denote the features which can be shared.

http://aclweb.org/anthology/P/P17/P17-1001.pdf

对抗学习



$$L = L_{Task} + \lambda L_{Adv} + \gamma L_{Diff}$$

$$L_{Adv} = \min_{\theta_s} \left(\lambda \max_{\theta_D} \left(\sum_{k=1}^{K} \sum_{i=1}^{N_k} d_i^k \log[D(E(\mathbf{x}^k))] \right) \right)$$

$$L_{\text{diff}} = \sum_{k=1}^{K} \left\| \mathbf{S}^{k\top} \mathbf{H}^k \right\|_F^2$$

多任务学习数据集

Dataset	Train	Dev.	Test	Unlab.	Avg. L	Vocab.
Books	1400	200	400	2000	159	62K
Elec.	1398	200	400	2000	101	30K
DVD	1400	200	400	2000	173	69K
Kitchen	1400	200	400	2000	89	28K
Apparel	1400	200	400	2000	57	21K
Camera	1397	200	400	2000	130	26K
Health	1400	200	400	2000	81	26K
Music	1400	200	400	2000	136	60K
Toys	1400	200	400	2000	90	28K
Video	1400	200	400	2000	156	57K
Baby	1300	200	400	2000	104	26K
Mag.	1370	200	400	2000	117	30K
Soft.	1315	200	400	475	129	26K
Sports	1400	200	400	2000	94	30K
IMDB	1400	200	400	2000	269	44K
MR	1400	200	400	2000	21	12K

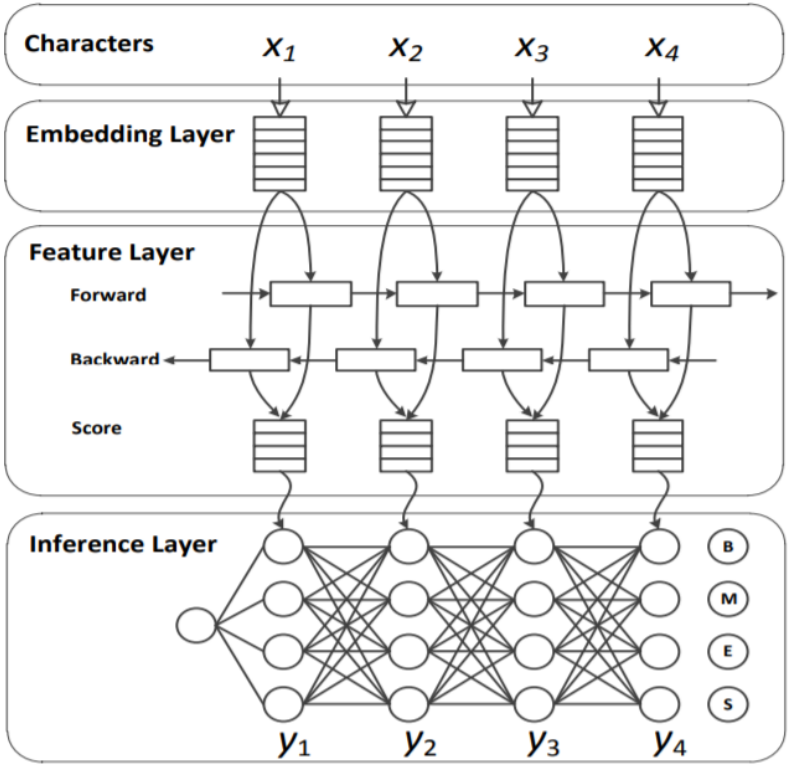
结果

Task	Single Task				Multiple Tasks				
	LSTM	BiLSTM	sLSTM	Avg.	MT-DNN	MT-CNN	FS-MTL	SP-MTL	ASP-MTL
Books	20.5	19.0	18.0	19.2	$17.8_{(-1.4)}$	$15.5_{(-3.7)}$	$17.5_{(-1.7)}$	$18.8_{(-0.4)}$	$16.0_{(-3.2)}$
Electronics	19.5	21.5	23.3	21.4	$18.3_{(-3.1)}$	$16.8_{(-4.6)}$	$14.3_{(-7.1)}$	$15.3_{(-6.1)}$	$13.2_{(-8.2)}$
DVD	18.3	19.5	22.0	19.9	$15.8_{(-4.1)}$	$16.0_{(-3.9)}$	$16.5_{(-3.4)}$	$16.0_{(-3.9)}$	$14.5_{(-5.4)}$
Kitchen	22.0	18.8	19.5	20.1	$19.3_{(-0.8)}$	$16.8_{(-3.3)}$	$14.0_{(-6.1)}$	$14.8_{(-5.3)}$	$13.8_{(-6.3)}$
Apparel	16.8	14.0	16.3	15.7	$15.0_{(-0.7)}$	$16.3_{(+0.6)}$	$15.5_{(-0.2)}$	$13.5_{(-2.2)}$	$13.0_{(-2.7)}$
Camera	14.8	14.0	15.0	14.6	$13.8_{(-0.8)}$	$14.0_{(-0.6)}$	$13.5_{(-1.1)}$	$12.0_{(-2.6)}$	$10.8_{(-3.8)}$
Health	15.5	21.3	16.5	17.8	$14.3_{(-3.5)}$	$12.8_{(-5.0)}$	$12.0_{(-5.8)}$	$12.8_{(-5.0)}$	$11.8_{(-6.0)}$
Music	23.3	22.8	23.0	23.0	$15.3_{(-7.7)}$	$16.3_{(-6.7)}$	$18.8_{(-4.2)}$	$17.0_{(-6.0)}$	$17.5_{(-5.5)}$
Toys	16.8	15.3	16.8	16.3	$12.3_{(-4.0)}$	$10.8_{(-5.5)}$	$15.5_{(-0.8)}$	$14.8_{(-1.5)}$	$12.0_{(-4.3)}$
Video	18.5	16.3	16.3	17.0	$15.0_{(-2.0)}$	$18.5_{(+1.5)}$	$16.3_{(-0.7)}$	$16.8_{(-0.2)}$	$15.5_{(-1.5)}$
Baby	15.3	16.5	15.8	15.9	$12.0_{(-3.9)}$	$12.3_{(-3.6)}$	$12.0_{(-3.9)}$	$13.3_{(-2.6)}$	$11.8_{(-4.1)}$
Magazines	10.8	8.5	12.3	10.5	$10.5_{(+0.0)}$	$12.3_{(+1.8)}$	$7.5_{(-3.0)}$	$8.0_{(-2.5)}$	$7.8_{(-2.7)}$
Software	15.3	14.3	14.5	14.7	$14.3_{(-0.4)}$	$13.5_{(-1.2)}$	$13.8_{(-0.9)}$	$13.0_{(-1.7)}$	$12.8_{(-1.9)}$
Sports	18.3	16.0	17.5	17.3	$16.8_{(-0.5)}$	$16.0_{(-1.3)}$	$14.5_{(-2.8)}$	$12.8_{(-4.5)}$	$14.3_{(-3.0)}$
IMDB	18.3	15.0	18.5	17.3	$16.8_{(-0.5)}$	$13.8_{(-3.5)}$	$17.5_{(+0.2)}$	$15.3_{(-2.0)}$	$14.5_{(-2.8)}$
MR	27.3	25.3	28.0	26.9	$24.5_{(-2.4)}$	$25.5_{(-1.4)}$	$25.3_{(-1.6)}$	$24.0_{(-2.9)}$	$23.3_{(-3.6)}$
AVG	18.2	17.4	18.3	18.0	$15.7_{(-2.2)}$	$15.5_{(-2.5)}$	$15.3_{(-2.7)}$	$14.9_{(-3.1)}$	$13.9_{(-4.1)}$

ACL 2017 Outstanding Paper

Adversarial Multi-Criteria Learning for Chinese Word Segmentation

Xinchi Chen, Zhan Shi, Xipeng Qiu,* Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University



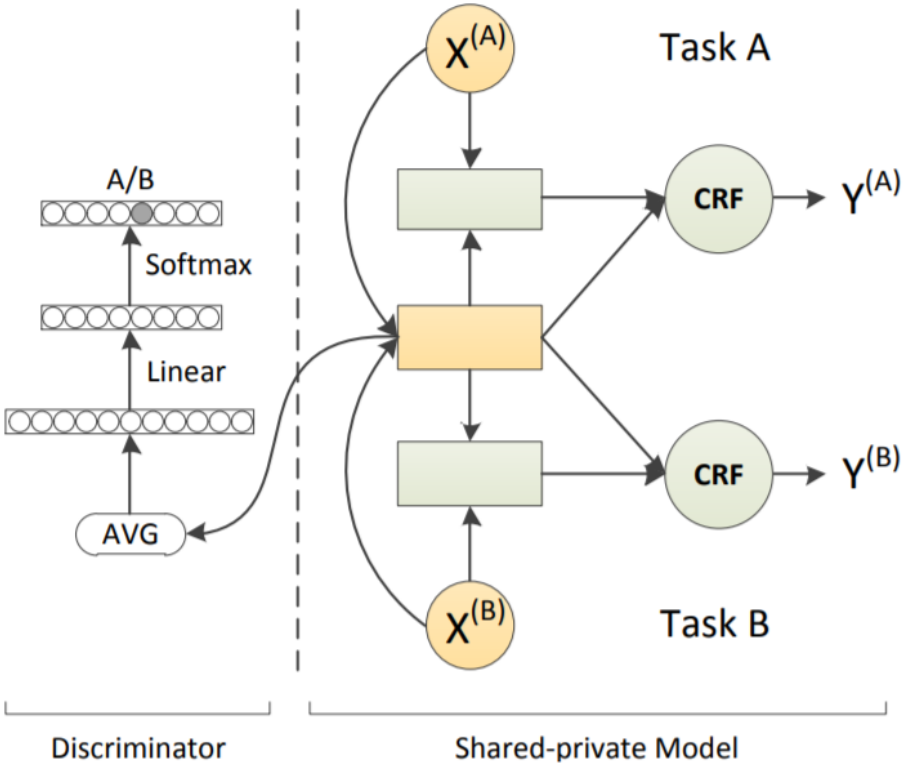
Corpora	Yao	Ming	reaches	the final	
CTB	姚明		进入	总决赛	
PKU	姚	明	进入	总	决赛

http://aclweb.org/anthology/P/P17/P17-1110.pdf

ACL 2017 Outstanding Paper

Adversarial Multi-Criteria Learning for Chinese Word Segmentation

Xinchi Chen, Zhan Shi, Xipeng Qiu,* Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University



http://aclweb.org/anthology/P/P17/P17-1110.pdf

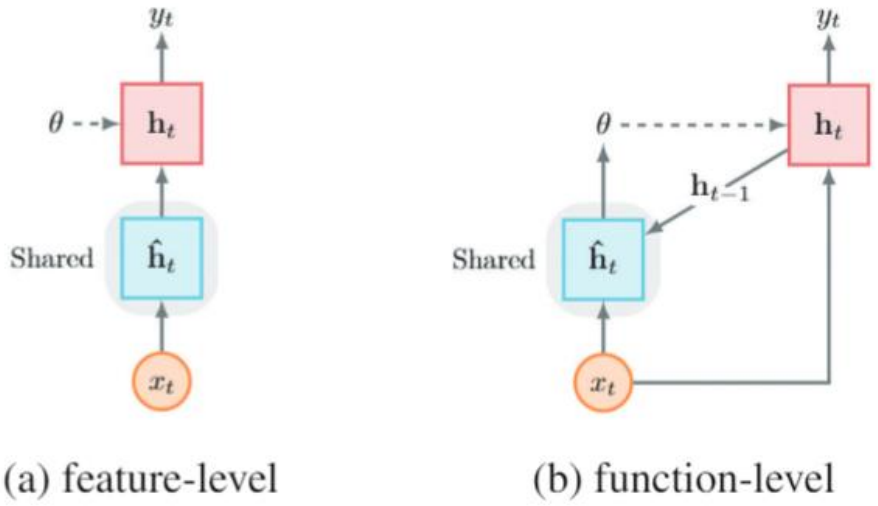
函数共享模式

AAAI 2018

Meta Multi-Task Learning for Sequence Modeling

Junkun Chen, Xipeng Qiu*, Pengfei Liu, Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, China

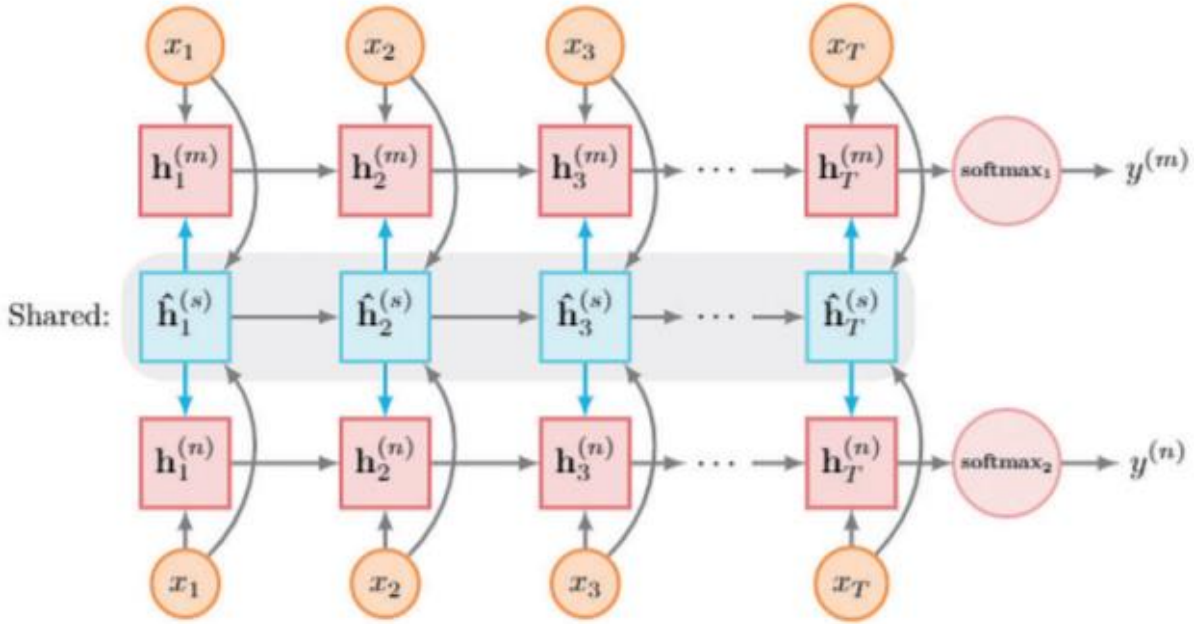
元学习：学习的学习



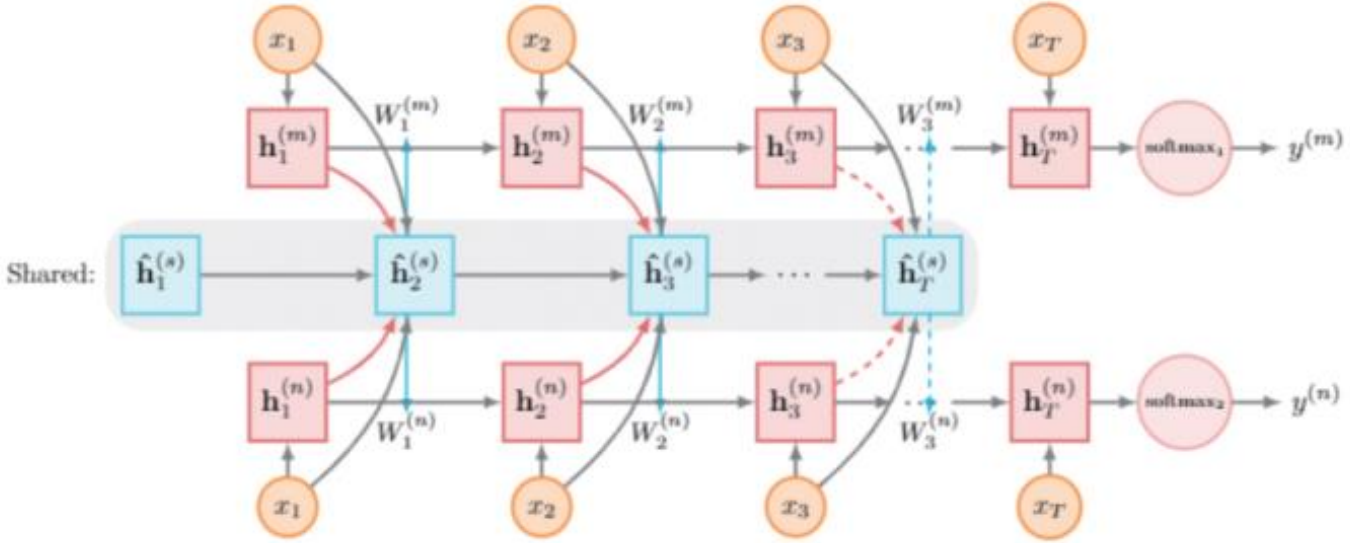
(a) feature-level (b) function-level

https://arxiv.org/pdf/1802.08969.pdf

共享特征



共享函数



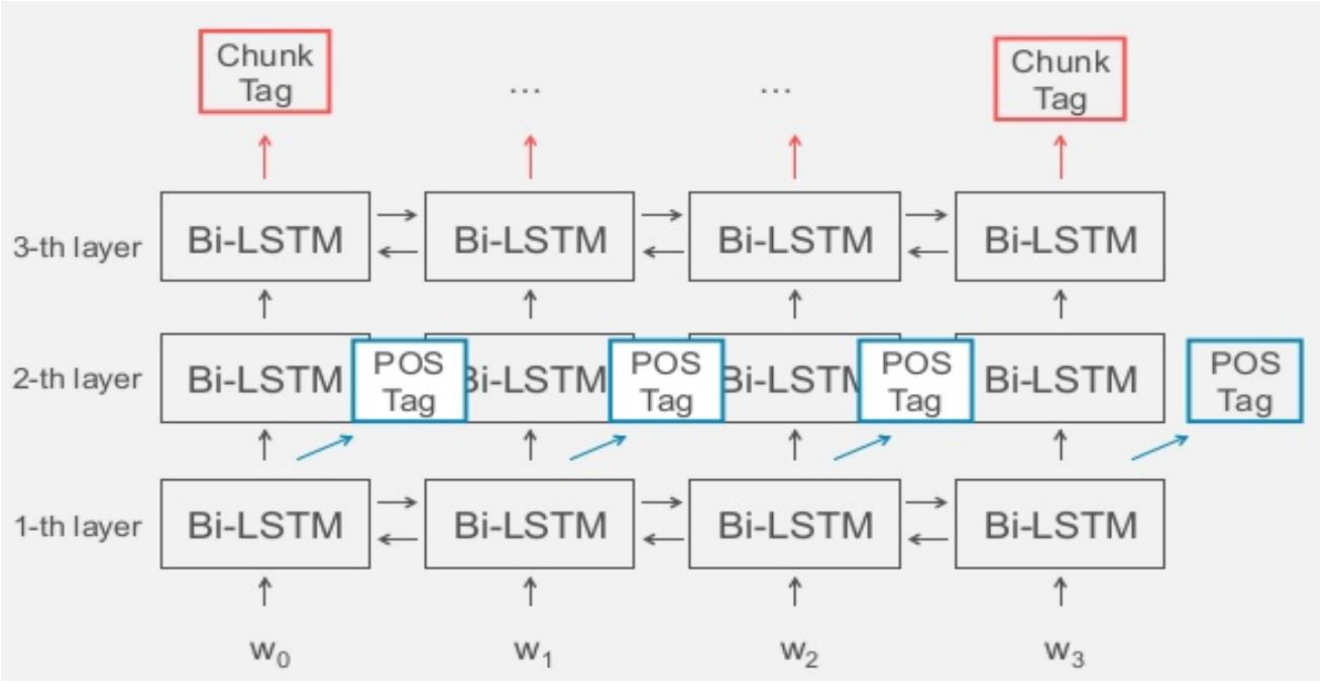
多级共享模式

ACL short 2016

Deep multi-task learning with low level tasks supervised at lower layers

Anders Søgaard
University of Copenhagen
soegaard@hum.ku.dk

Yoav Goldberg
Bar-Ilan University
yoav.goldberg@gmail.com

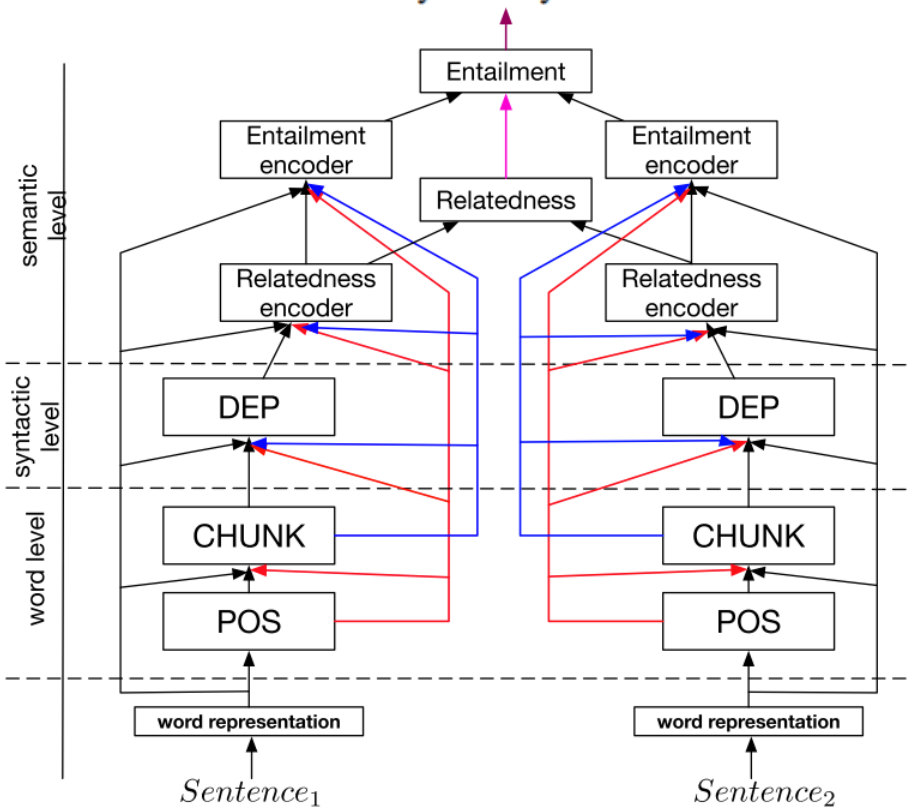


	LAYERS		DOMAINS			
	CHUNKS	POS	BROADCAST (6)	BC-NEWS (8)	MAGAZINES (1)	WEBLOGS (6)
BI-LSTM	3	-	88.98	91.84	90.09	90.36
	3	3	88.91	91.84	90.95	90.43
	3	1	**89.48**	**92.03**	**91.53**	**90.78**

http://anthology.aclweb.org/P16-2038

A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher
The University of Tokyo



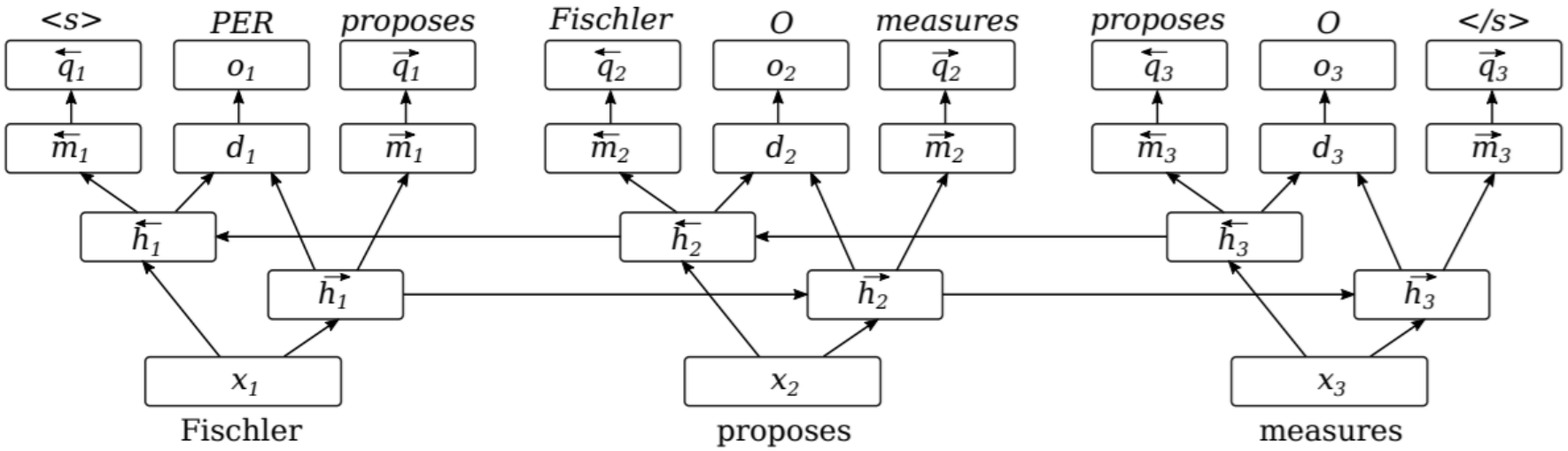
		Single	JMT_{all}	JMT_{AB}	JMT_{ABC}	JMT_{DE}	JMT_{CD}	JMT_{CE}
A ↑	POS	97.45	97.55	97.52	97.54	n/a	n/a	n/a
B ↑	Chunking	95.02	n/a	95.77	n/a	n/a	n/a	n/a
C ↑	Dependency UAS	93.35	94.67	n/a	94.71	n/a	93.53	93.57
	Dependency LAS	91.42	92.90	n/a	92.92	n/a	91.62	91.69
D ↓	Relatedness	0.247	0.233	n/a	n/a	0.238	0.251	n/a
E ↑	Entailment	81.8	86.2	n/a	n/a	86.8	n/a	82.4

主辅任务模式

ACL 2017

Semi-supervised Multitask Learning for Sequence Labeling

Marek Rei
The ALTA Institute
Computer Laboratory
University of Cambridge
United Kingdom



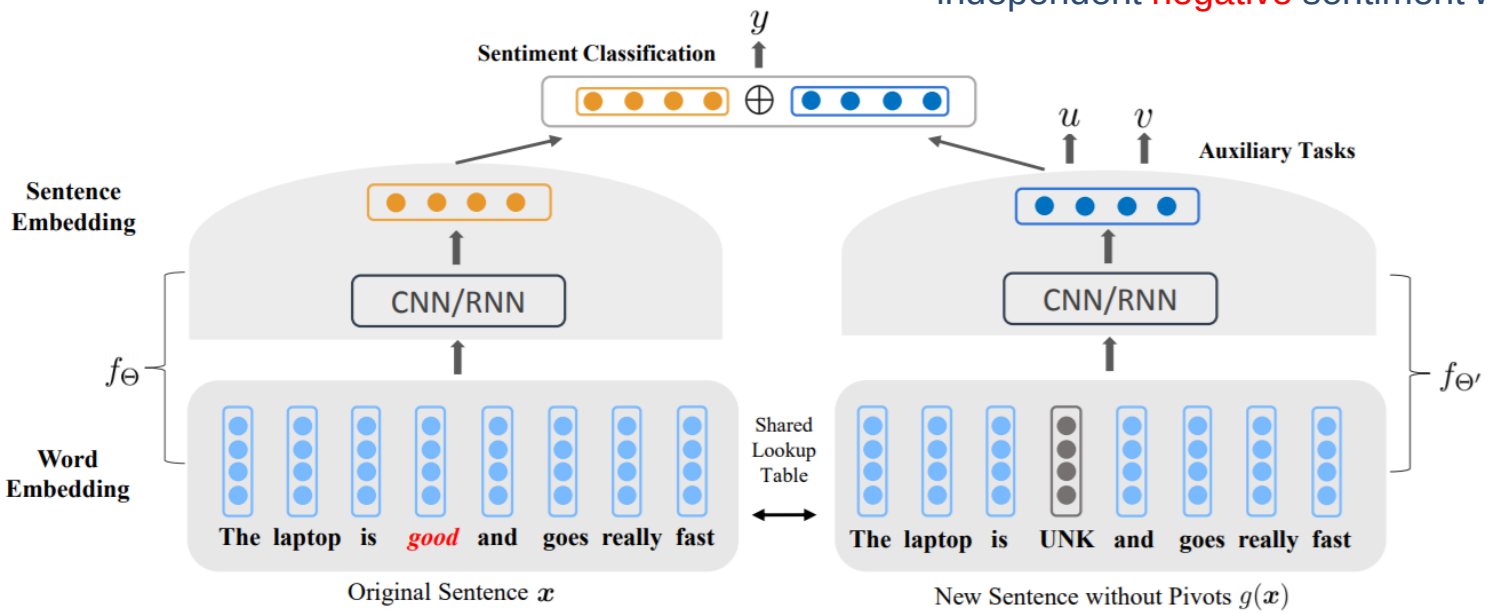
https://arxiv.org/pdf/1704.07156.pdf

Learning Sentence Embeddings with Auxiliary Tasks for Cross-Domain Sentiment Classification

Jianfei Yu
School of Information Systems
Singapore Management University
jfyu.2014@phdis.smu.edu.sg

Jing Jiang
School of Information Systems
Singapore Management University
jingjiang@smu.edu.sg

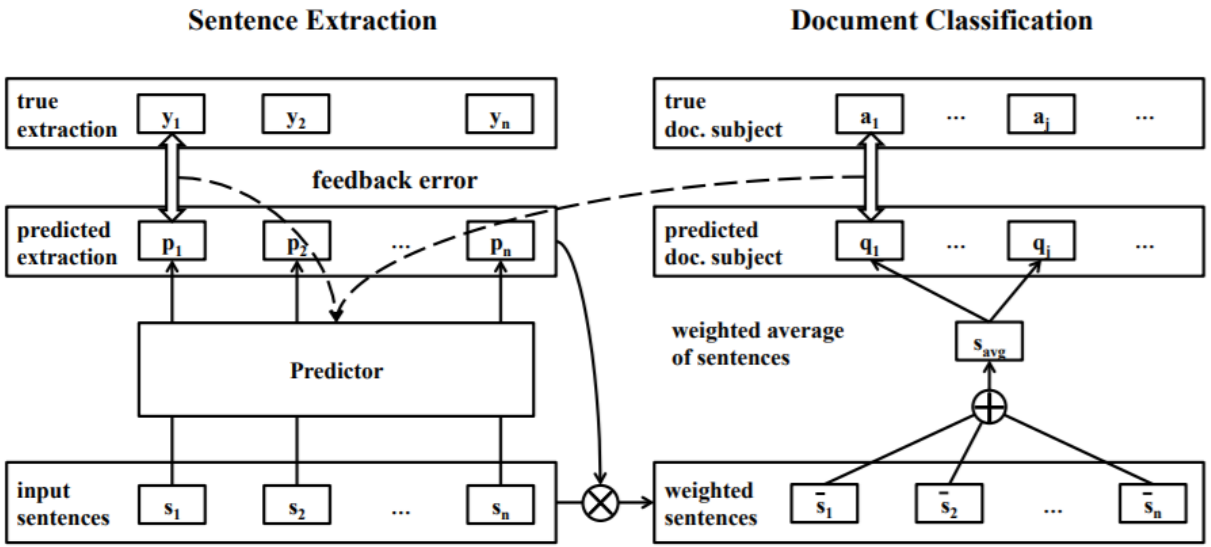
u indicates whether the original sentence x contains at least one domain independent positive sentiment word.
v indicates whether x contains at least one domain independent negative sentiment word.



http://www.aclweb.org/anthology/D/D16/D16-1023.pdf

Extractive Summarization Using Multi-Task Learning with Document Classification

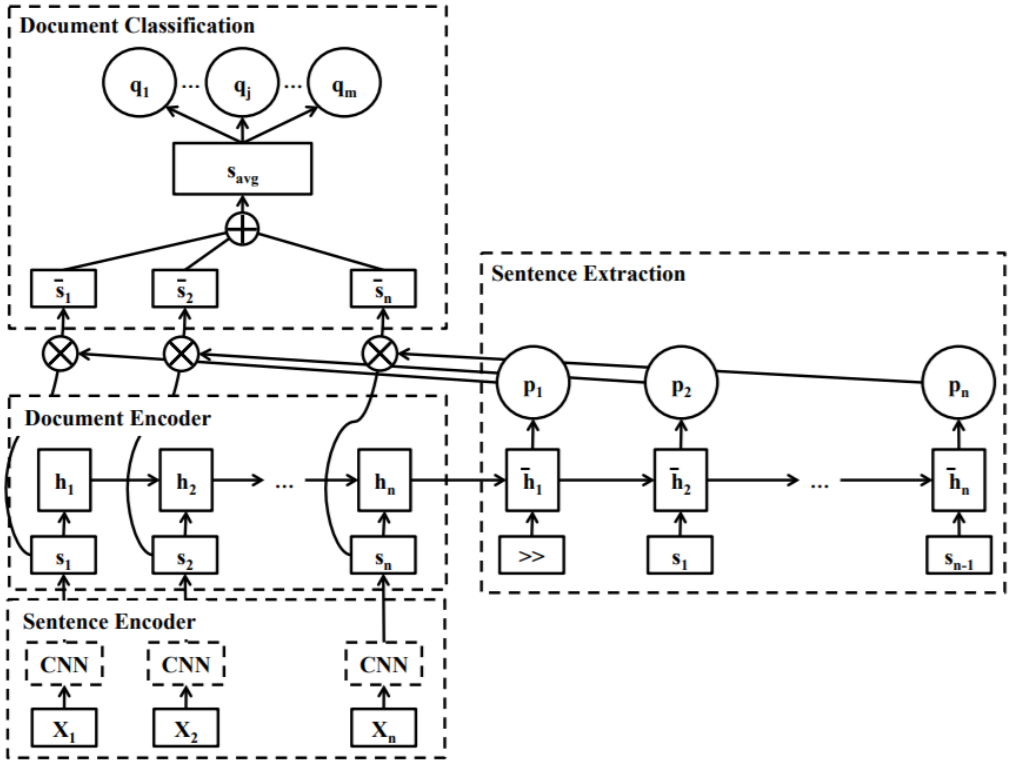
Masaru Isonuma,Toru Fujino, Junichiro Mori, Yutaka Matsuo and **Ichiro Sakata**
The University of Tokyo, Japan



http://www.aclweb.org/anthology/D17-1223

Extractive Summarization Using Multi-Task Learning with Document Classification

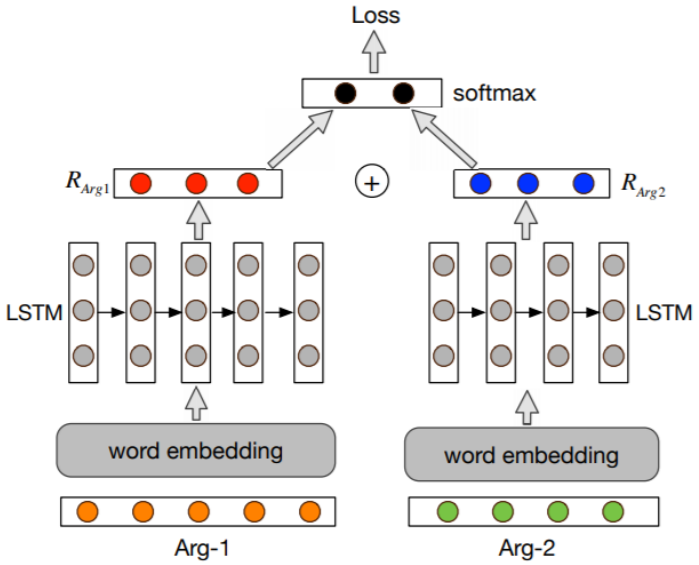
Masaru Isonuma,Toru Fujino, Junichiro Mori, Yutaka Matsuo and **Ichiro Sakata**
The University of Tokyo, Japan



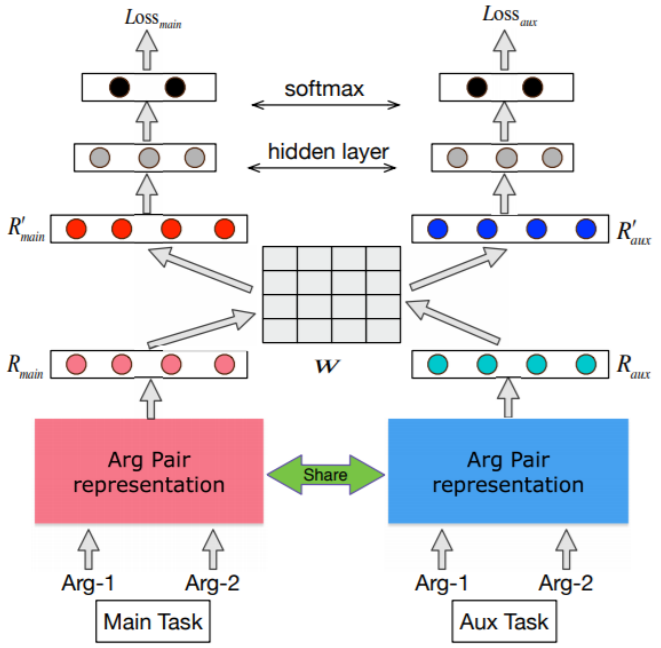
http://www.aclweb.org/anthology/D17-1223

Multi-task Attention-based Neural Networks for Implicit Discourse Relationship Representation and Identification

Man Lan[1,2], **Jianxiang Wang**[1,3*], **Yuanbin Wu**[1,2*], **Zheng-Yu Niu**[3*], **Haifeng Wang**[3]
[1] School of Computer Science and Software Engineering, East China Normal University
[2] Shanghai Key Laboratory of Multidimensional Information Processing, P.R.China
[3] Baidu Inc., Beijing, P.R.China



LSTM for discourse argument

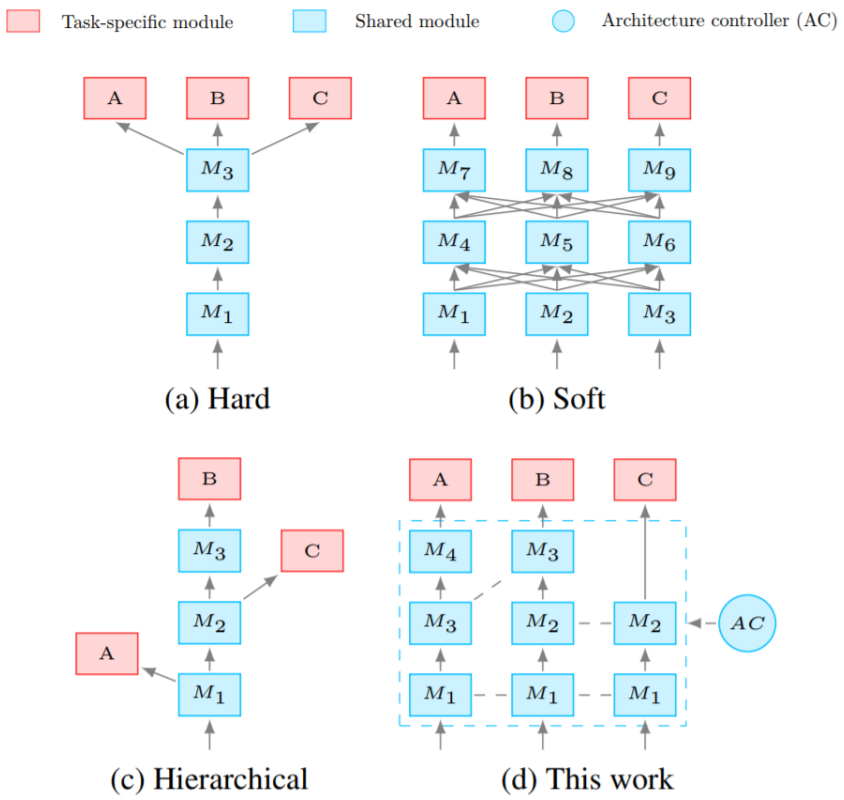


Multi-task Attention-based Neural Networks

共享模式搜索

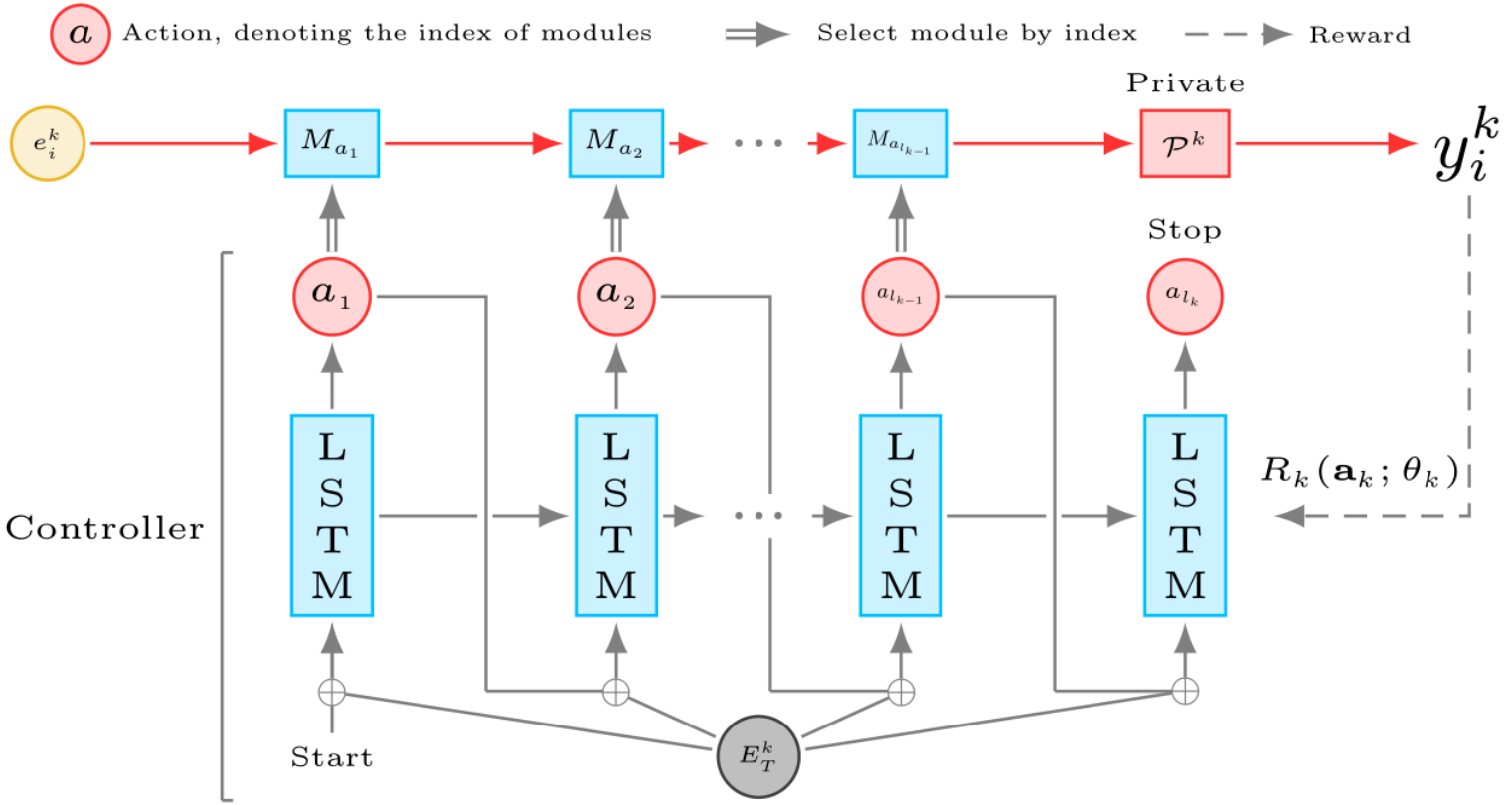
Exploring Shared Structures and Hierarchies for Multiple NLP Tasks

Junkun Chen*, **Kaiyu Chen***, **Xinchi Chen**, **Xipeng Qiu**†, **Xuanjing Huang**
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University

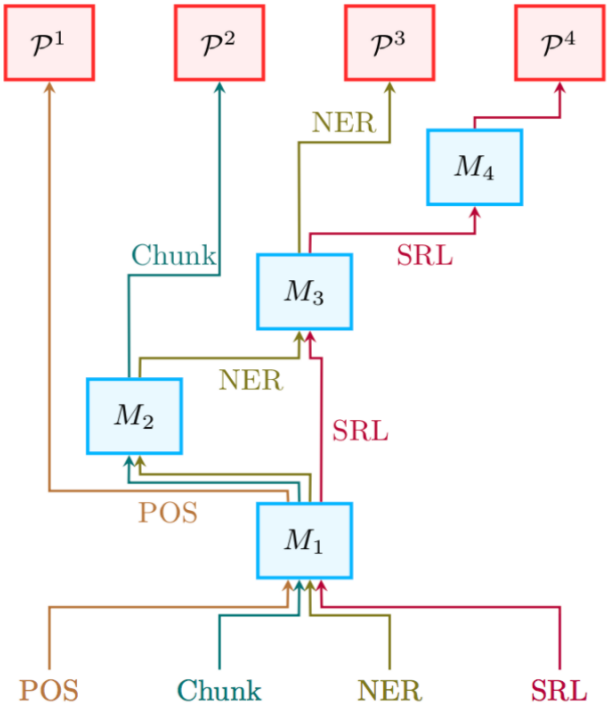


https://arxiv.org/pdf/1808.07658.pdf

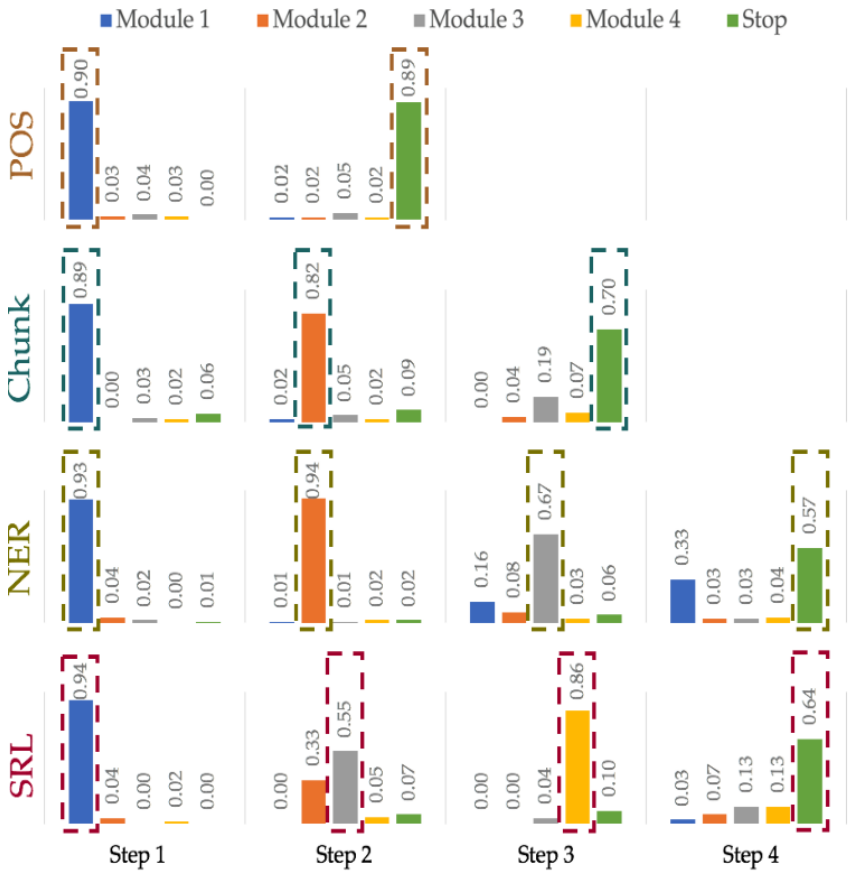
面向NLP的神经网络架构搜索



面向NLP的神经网络架构搜索



自动选择的共享模式

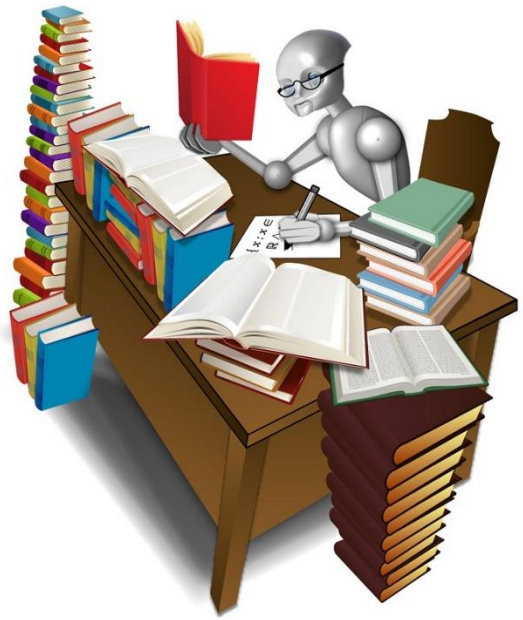


自动选择过程

新的多任务基准平台

机器阅读

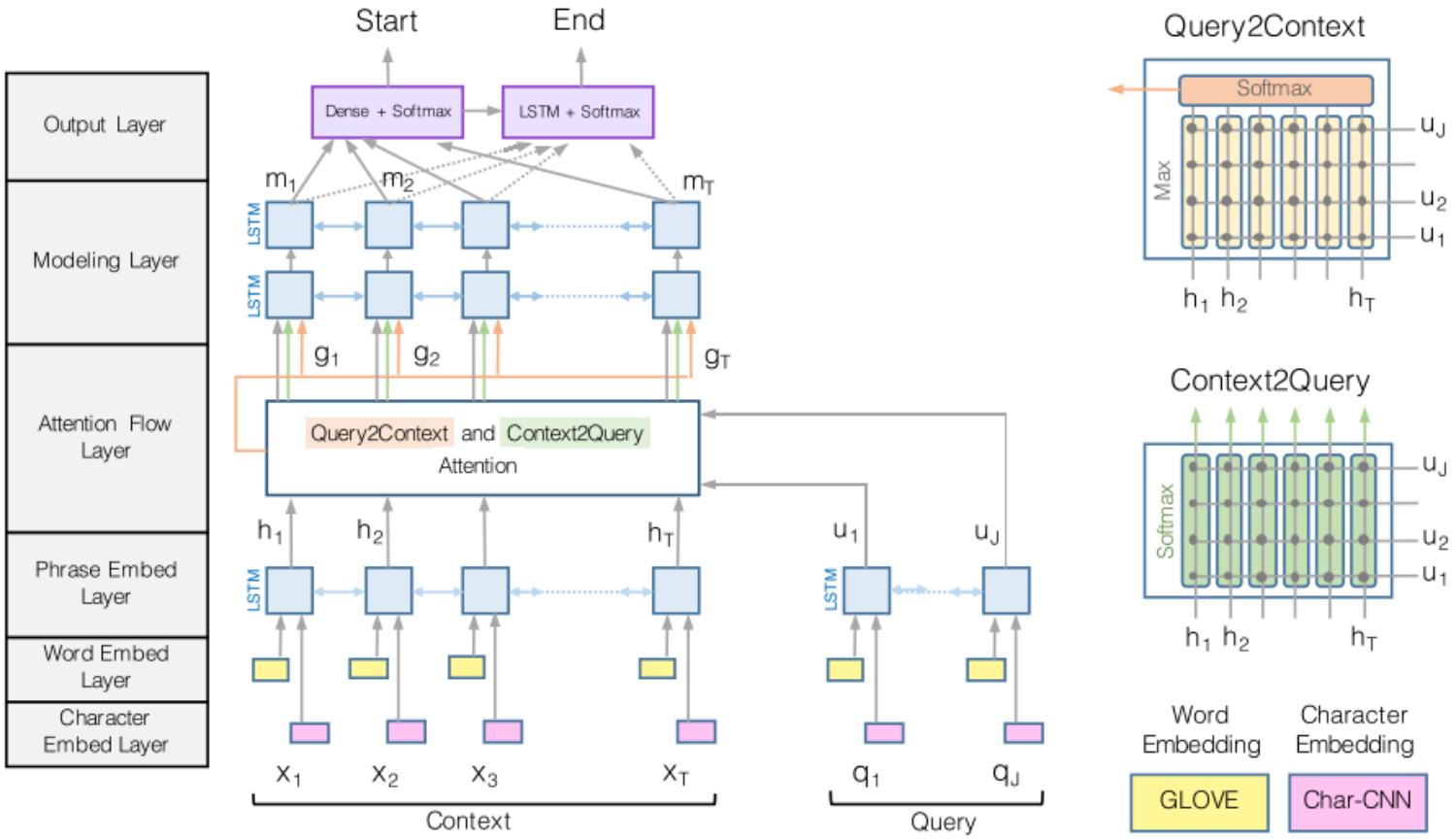
- 阅读一篇或多篇文档，并回答一些相关问题。



Document: What was supposed to be a fantasy sports car ride at Walt Disney World Speedway turned deadly when a Lamborghini crashed into a guardrail. The crash took place Sunday at the Exotic Driving Experience, which bills itself as a chance to drive your dream car on a racetrack. The Lamborghini's passenger, 36-year-old Gary Terry of Davenport, Florida, died at the scene, Florida Highway Patrol said. The driver of the Lamborghini, 24-year-old Tavon Watson of Kissimmee, Florida, lost control of the vehicle, the Highway Patrol said. (...)

Question:
Officials say the driver, 24-year-old Tavon Watson, lost control of a _____.

Bidirectional Attention (Seo et al.,2016)



The Natural Language Decathlon: Multitask Learning as Question Answering

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, Richard Socher
Salesforce Research

Examples

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

Question	Context	Answer
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan

http://decanlp.com

https://arxiv.org/pdf/1806.08730.pdf

The Natural Language Decathlon: Multitask Learning as Question Answering

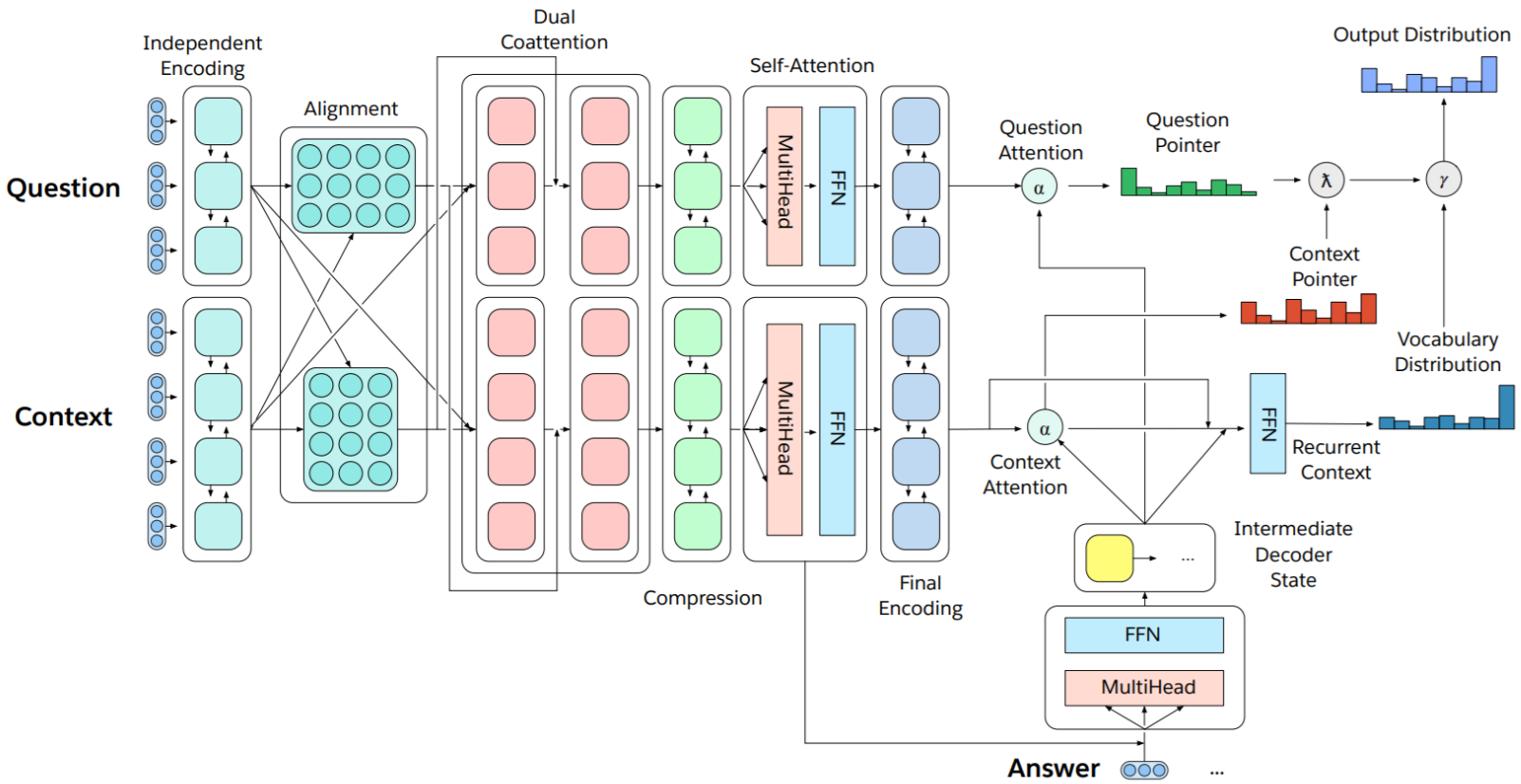
Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, Richard Socher
Salesforce Research

Task	Dataset	# Train	# Dev	# Test	Metric
Question Answering	SQuAD	87599	10570	9616	nF1
Machine Translation	IWSLT	196884	993	1305	BLEU
Summarization	CNN/DM	287227	13368	11490	ROUGE
Natural Language Inference	MNLI	392702	20000	20000	EM
Sentiment Analysis	SST	6920	872	1821	EM
Semantic Role Labeling	QA-SRL	6414	2183	2201	nF1
Zero-Shot Relation Extraction	QA-ZRE	840000	600	12000	cF1
Goal-Oriented Dialogue	WOZ	2536	830	1646	dsEM
Semantic Parsing	WikiSQL	56355	8421	15878	lfEM
Pronoun Resolution	MWSC	80	82	100	EM

https://arxiv.org/pdf/1806.08730.pdf

The Natural Language Decathlon: Multitask Learning as Question Answering

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, Richard Socher
Salesforce Research



https://arxiv.org/pdf/1806.08730.pdf

GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding

Alex Wang[1], Amanpreet Singh[1], Julian Michael[2], Felix Hill[3], Omer Levy[2], and Samuel R. Bowman[1]

[1]New York University, New York, NY
[2]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA
[3]DeepMind, London, UK

Corpus	|Train|	|Dev|	|Test|	Task	Metric	Domain
			Single-Sentence Tasks			
CoLA	10k	1k	1.1k	acceptability	Matthews	linguistics literature
SST-2	67k	872	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks			
MRPC	4k	N/A	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.5k	1.4k	sentence similarity	Pearson/Spearman	misc.
QQP	400k	N/A	391k	paraphrase	acc./F1	social QA Questions
			Inference Tasks			
MNLI	393k	20k	20k	NLI	acc. (match/mismatch)	misc.
QNLI	108k	11k	11k	QA/NLI	acc.	Wikipedia
RTE	2.7k	N/A	3k	NLI	acc.	misc.
WNLI	706	N/A	146	coreference/NLI	acc.	fiction books

https://arxiv.org/pdf/1804.07461.pdf

General Language Understanding Evaluation (GLUE) benchmark

Welcome to GLUE

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	Alec Radford	Singletask Pretrain Transformer		72.8	45.4	91.3	75.7/82.3	82.0/80.0	88.5/70.3	82.1	81.4	88.1	56.0	53.4	29.8
2	Samuel Bowman	BiLSTM+ELMo+Attn		70.5	36.0	90.4	77.9/84.9	75.1/73.3	84.7/64.8	76.4	76.1	79.9	56.8	65.1	26.5
3	GLUE Baselines	BiLSTM+ELMo+Attn		68.9	18.9	91.6	77.3/83.5	72.8/71.1	83.5/63.3	75.6	75.9	81.7	61.2	65.1	22.6
		GenSen		66.6	7.7	83.1	76.6/83.0	79.3/79.2	82.9/59.8	71.4	71.3	82.3	59.2	65.1	20.6
		Single Task BiLSTM+ELMo		66.2	35.0	90.2	69.0/80.8	64.0/60.2	85.7/65.6	72.9	73.4	69.4	50.1	65.1	19.5
		BiLSTM+Attn		65.7	0.0	85.0	75.1/83.7	73.9/71.8	84.3/63.6	72.2	72.1	82.1	61.7	63.7	24.6
		BiLSTM+ELMo		64.9	27.5	89.6	76.2/83.5	67.0/65.9	78.5/57.8	67.1	68.0	66.7	55.7	62.3	19.2
		Single Task BiLSTM+ELMo+Attn		64.8	35.0	90.2	68.8/80.2	55.5/52.5	86.5/66.1	76.9	76.7	61.1	50.3	65.1	27.9

https://gluebenchmark.com/

总结

- 自然语言处理简介
- 基于深度学习的自然语言处理
- 深度学习在自然语言处理中的困境
  - 无监督预训练
  - 多任务学习
- 自然语言处理中的多任务学习
  - 硬共享模式
  - 软共享模式
  - 共享-私有模式
  - 函数共享模式
  - 多级共享模式
  - 主辅任务模式
- 新的多任务基准平台

谢 谢