Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

对抗式多任务学习

Adversarial Multi-task Learning

李筱雅　硕士

2019年7月26日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

背景简介

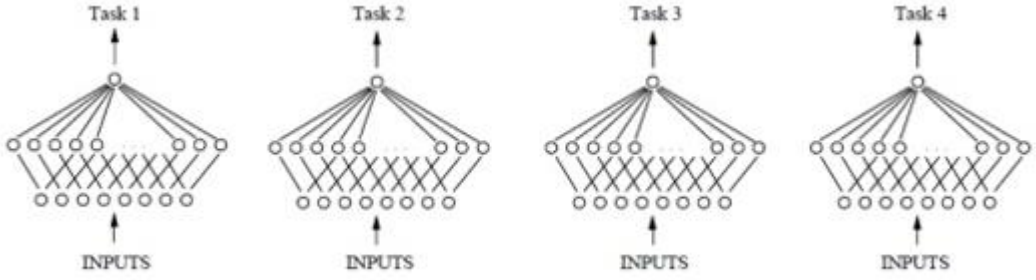
- 预期收获
  - 1. 理解单任务与多任务学习的定义及其应用场景
  - 2.了解多任务模型的基本模型框架
  - 3. 了解对抗式多任务学习所解决的问题
  - 4. 理解对抗式多任务学习的算法原理

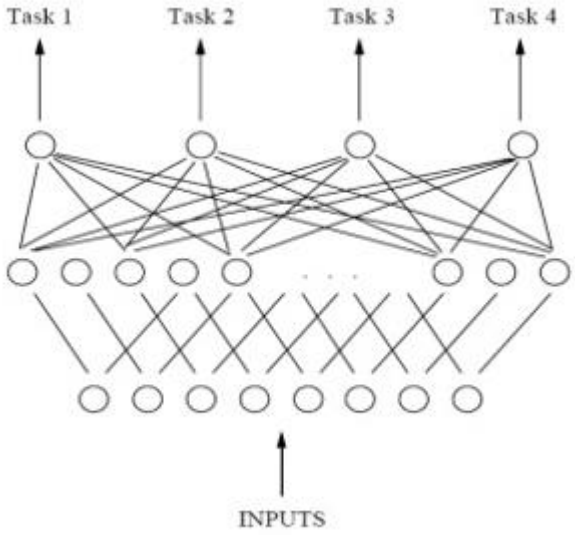
背景简介

- 多任务学习提出背景
  - 由Argyriou et al. 于2008年提出，用于预测多个学校的学生成绩。研究发现，在这个任务上使用多任务学习的的预测效果优于Ridge Regression和Lasso。
  - 对于139个中学的15362个学生，每个中学都可以看作一个预测任务
  - 单任务学习忽视了多个任务之间的关系，多任务学习既考虑任务之间的关联，又考虑任务之间的差别。

基本概念

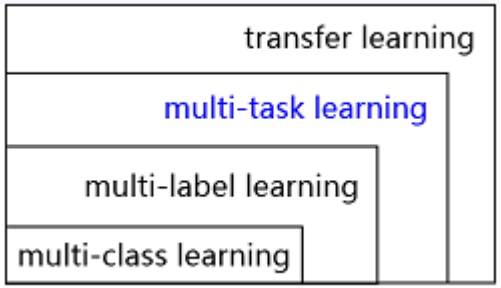
- 基本概念
  - 单任务学习：一次只学习一个任务，大部分的机器学习任务均属于单任务
  - 多任务学习：基于共享表示，把多个相关任务放在一起并行学习的机器学习方法





基本概念

- 基本概念
  - 迁移学习（Transfer Learning）：利用源任务（source task）学习到的知识来帮助学习另一个目标任务（target task）
  - 多标签学习（Multi-Label Learning）：给每一个样本添加一系列的目标标签
  - 多类别学习（Multi-Class Learning）:超过2个分类的分类任务

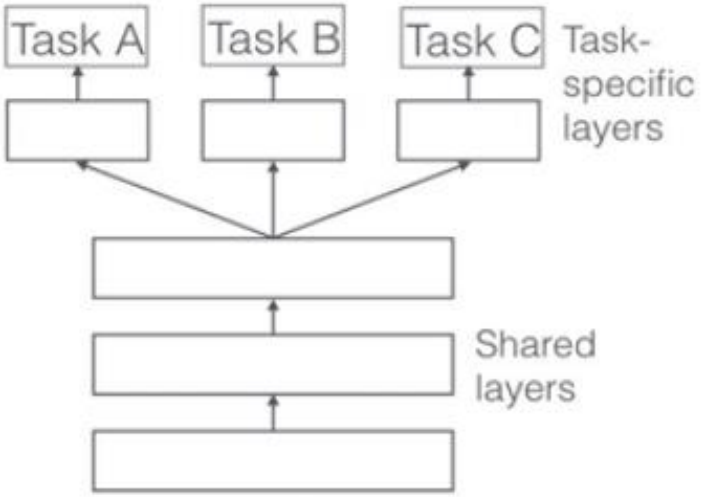


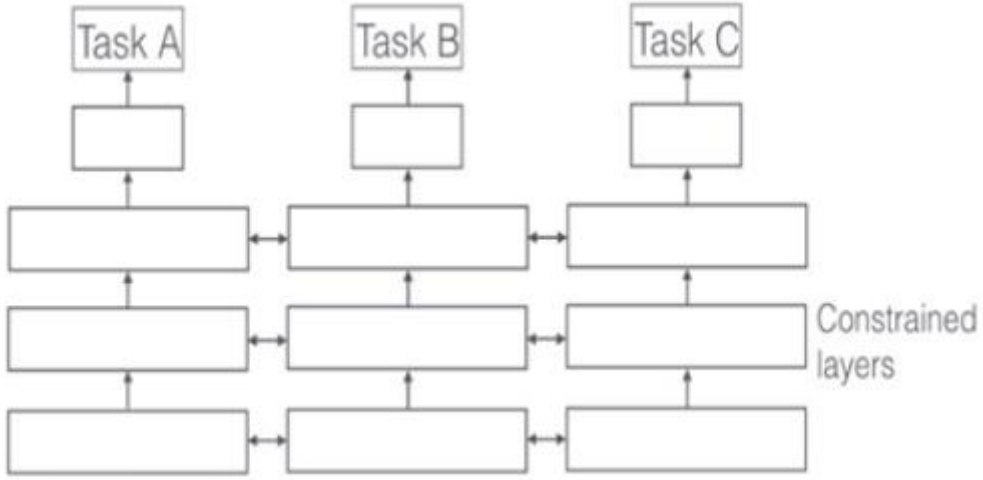
基本概念

- 基本概念
  - 相关任务：
    - LearnigAlg（Main Task,Related Task）
      >LearningAlg(MainTask)
    - LearningAlg(Main Task | Related task)
      > LearningAlg(Main Task)
  - 共享表示
    - 基于参数的共享：如多任务学习，高斯处理过程
    - 基于约束的共享：如均值，联合特征学习

基本概念

- 参数共享
  - 硬共享机制：所有任务之间共享隐藏层，同时保留几个任务的输出层来实现
  - 软共享机制：每个任务都有自己的模型，自己的参数





算法原理

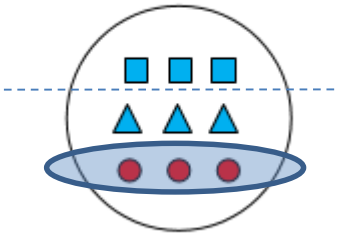
T	对抗式多任务学习
I	多个相关任务，或一个可拆分为多个相关子任务的主任务
P	特征共享：Shared-Private Model 训练：(1)对抗式损失 (2)正交约束
O	高性能分类器

P	尽可能提高特征提取的准确性
C	具有多个相关任务
D	如何良好地利用多个任务之间的相关关系
L	IJCAI2016，ACL2017，ACL2018

算法原理

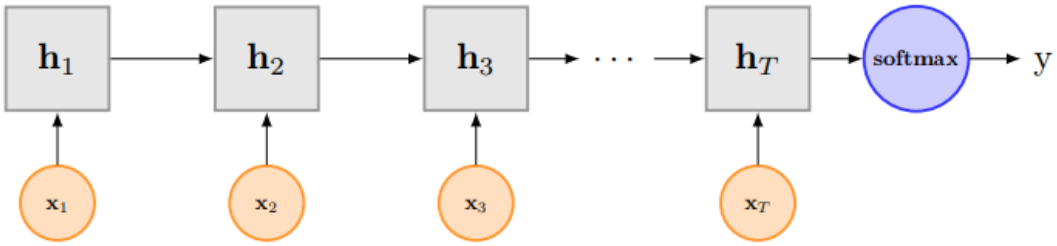
- 多任务学习
  - 数据：N个评论类数据集，如对书籍，DVD，电子产品等的评论
  - 目标：对N个数据集中的评论分类，判断是消极评论还是积极的评论，每个数据集都看作一个任务

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

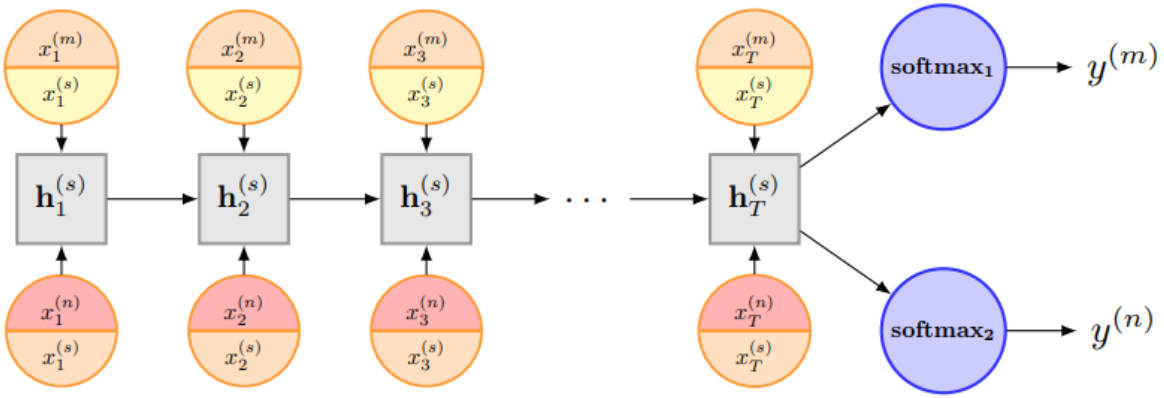


算法原理

传统模型



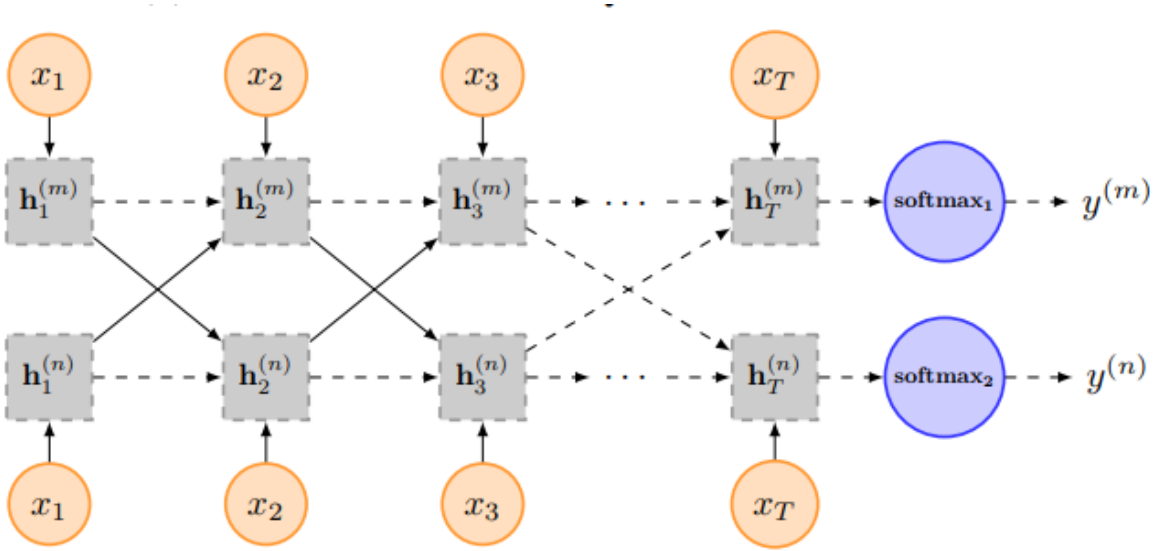
多任务学习模型1



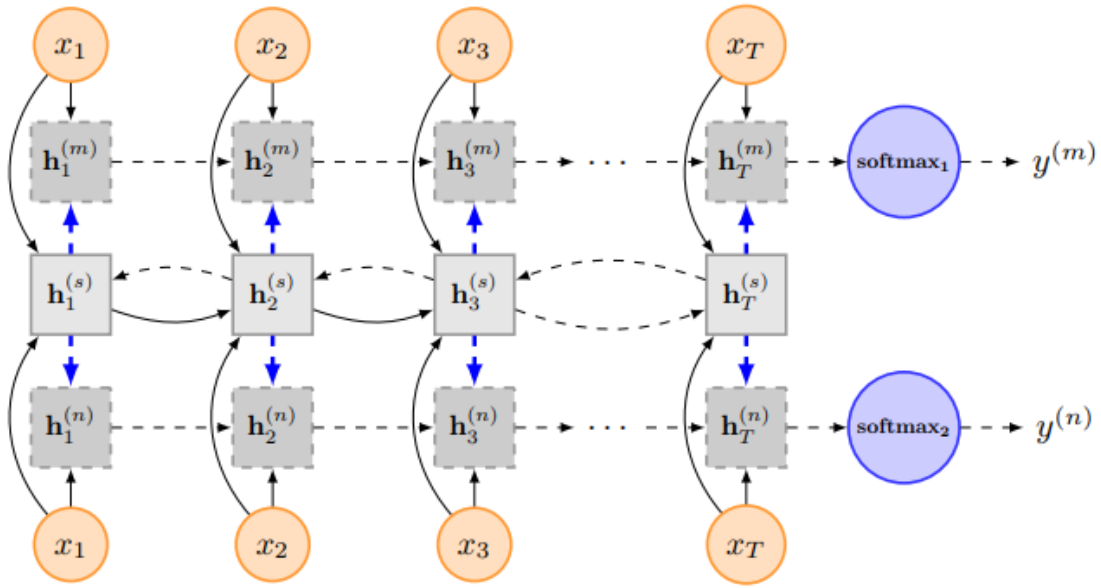
$$\hat{\mathbf{x}}_t^{(m)} = \mathbf{x}_t^{(m)} \oplus \mathbf{x}_t^{(s)}$$

算法原理

多任务
学习模型2

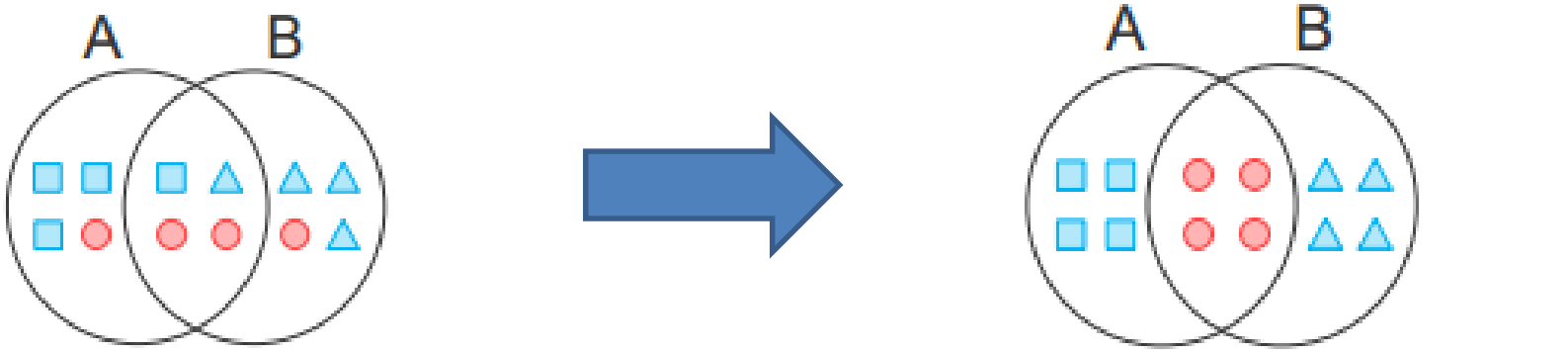


多任务
学习模型3



算法原理

- 对抗式多任务学习
  - 研究问题：基于神经网络的方法在共享层提取共享特征时，被某些特定任务的特定特征所污染。
  - 如情感分析任务中：
    - The infantile cart is simple and easy to use （中立）
    - The kind of humor is infantile and boring（消极）



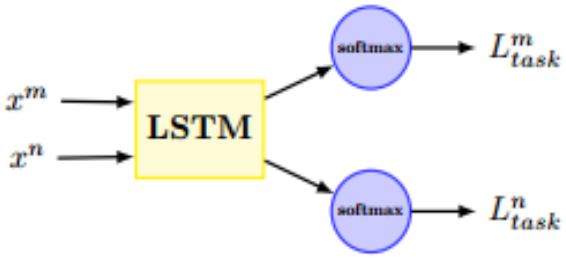
(a) Shared-Private Model

(b) Adversarial Shared-Private Model

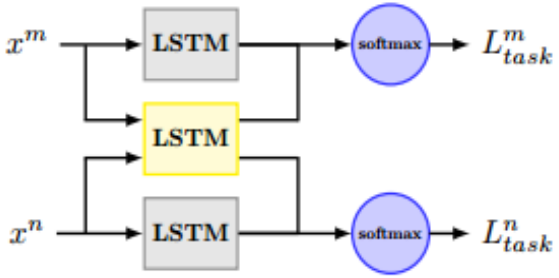
算法原理

- 两种共享方式
  - Fully-Shared Model：同一个特征空间，特征完全共享
  - Shared-Private Model：两个特征空间
- 特定任务的输出层

$$L_{Task} = \sum_{k=1}^{K} \alpha_k L(y^{(k)}, y^{(k)})$$



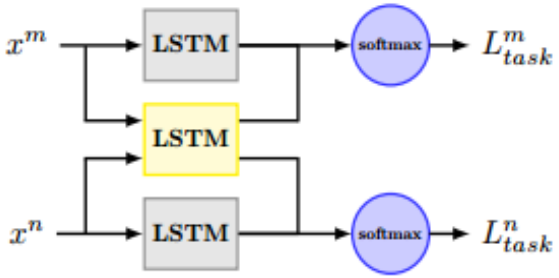
(a) Fully Shared Model (FS-MTL)



(b) Shared-Private Model (SP-MTL)

算法原理

- 训练过程
  - 对抗式损失：用于将私有特征从共享特征中剔除
  - 沿用生成式对抗网络的思想
    - 使用生成器生成数据去拟合真实数据，直到使判别器无法判别数据来自真实数据还是生成数据
    - 《2017.11.5-GAN-周妍汶》



(b) Shared-Private Model (SP-MTL)

生成器

$$\begin{cases} s_t^k = LSTM(x_t, s_{t-1}^k, \theta_s) \\ h_t^k = LSTM(x_t, \mathrm{h}_{t-1}^m, \theta_s) \end{cases}$$

判别器 $\rightarrow D(s_T^k, \theta_D) = \mathrm{softmax}(\mathrm{b} + \mathrm{U}s_T^k)$

对抗式损失 $\rightarrow L_{Adv} = \min_{\theta_s}(\lambda \max_{\theta_D}(\sum_{k=1}^{K}\sum_{i=1}^{N_k} d_i^k \log[\mathrm{D}(\mathrm{E}(\mathrm{x}^k))]))$

算法原理

- 训练过程
  - 正交约束：用于减轻私有和共享特征空间中的冗余特征

$$L_{diff} = \sum_{k=1}^{K} \| \mathrm{s}^{k^{\mathrm{T}}} H^{k} \|_F^2$$

  - 最终损失函数

$$L = L_{Tasks} + \lambda L_{Adv} + \gamma L_{diff}$$

算法原理

- 实验结果

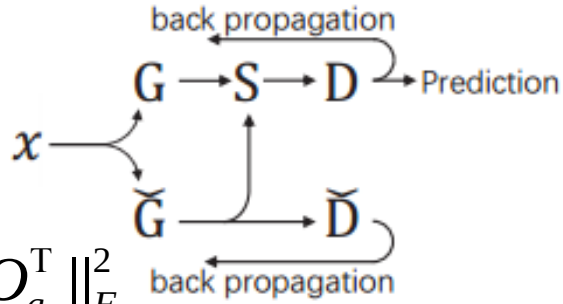
Task	Single Task				Multiple Tasks				
	LSTM	BiLSTM	sLSTM	Avg.	MT-DNN	MT-CNN	FS-MTL	SP-MTL	ASP-MTL
Books	20.5	19.0	18.0	19.2	$17.8_{(-1.4)}$	$15.5_{(-3.7)}$	$17.5_{(-1.7)}$	$18.8_{(-0.4)}$	$16.0_{(-3.2)}$
Electronics	19.5	21.5	23.3	21.4	$18.3_{(-3.1)}$	$16.8_{(-4.6)}$	$14.3_{(-7.1)}$	$15.3_{(-6.1)}$	$13.2_{(-8.2)}$
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	$15.7_{(-2.2)}$	$15.5_{(-2.5)}$	$15.3_{(-2.7)}$	$14.9_{(-3.1)}$	$13.9_{(-4.1)}$

算法原理

- 多任务学习有效的原因:
  - 隐层数据增加：有效增加了训练模型的样本大小
  - 注意力机制：任务嘈杂或数量有限且高维情况下，可以帮助模型将注意力集中在重要特征上
  - 窃听：其他任务难以学习的特征可以通过其他任务来学习
  - 表征偏置：多任务学习对任务的偏好造成模型偏差，有利于模型泛化到新任务上
  - 正则化：引入归纳偏置作为正则化项，降低了过拟合的风险

算法原理

- 在事件检测中的应用
  - 生成器：$o_t = LSTM(x_t;\theta)$
  - 判别器：$y = soft\max(W \cdot o_t + \hat{b})$
  - 自调节学习损失：$L_{diff}(O_g, O_{g'}) = \| O_g O_g^{\mathrm{T}} \|_F^2$
  - 训练：



$$\theta_g = \arg\min \mathrm{L}(y_g, \mathrm{y})$$

$$\theta_d = \arg\min(\mathrm{L}(y_g, \mathrm{y}) + \lambda \cdot \mathrm{L}_{diff})$$

$$\theta_{g'} = \arg\min \mathrm{L}(y_{g'}, \mathrm{y})$$

$$\theta_{d'} = \arg\max \mathrm{L}(y_{g'}, \mathrm{y})$$

算法原理

- 实验结果

Methods	P (%)	R (%)	F (%)
MSEP-EMD	70.4	65.0	67.6
Cross-Event	68.8	68.9	68.8
Cross-Entity	72.9	64.3	68.3
Joint (Local+Global)	73.7	62.3	67.5
CNN	71.8	66.4	69.0
DM-CNN	75.6	63.6	69.1
NC-CNN	-	-	71.3
FB-RNN (GRU)	66.8	68.0	67.4
Bi-RNN (GRU)	66.0	73.0	69.3
ANNs (ACE+FN)	77.6	65.2	70.7
DM-CNN*(ACE+Wiki)	75.7	66.0	70.5
ANN-S2 (ACE+FN)	76.8	67.5	71.9
Hybrid: Bi-LSTM+**CNN**	**84.6**	64.9	**73.4**
SELF: Bi-LSTM+**GAN**	71.3	**74.7**	**73.0**

Table 2: Detection performance (trigger identification plus multi-class classification)

优劣分析

- 优势
  - 对之前的通过共享参数粗略划分特征空间进行了改进，实现了共享特征和私有特征空间的精确划分
  - 将共享参数浓缩，更加容易迁移到新任务上
- 劣势

应用总结

- 算法的应用领域
  - 应用于多种场景的特征去噪（文本，数挖，网络安全）

参考文献

- IJCAI2016-Recurrent neural network for text classification with multi-task learning
- ACL2017-Adversarial Multi-task Learning for Text Classification
- ACL2018-Self-regulation: Employing a Generative Adversarial Network to Improve Event Detection
- https://zhuanlan.zhihu.com/p/27421983（多任务学习概述）

道德经

上善若水。水善利万物而不争，处众人之所恶，故几於道。居善地，心善渊与善仁，言善信，正善治，事善能，动善时。夫唯不争，故无尤。

谢谢！