

A Brief Introduction to Me and My Research

石雨凌

上海财经大学

2021.7.26

目录

个人简介

研究经历

偏微分方程的有限元方法研究

预训练模型的可解释性研究

Kaggle Question Answering 项目

Kaggle Sentence Classification 项目

金融知识图谱自动化构建项目

总结

个人简介

- ▶ 上海财经大学, 数学与应用数学专业, 年级排名 13/104, 本学期成绩排名 1/104. TOEFL 102 分.
- ▶ 主要课程: 数学分析 (3.3), 高等代数 (4.0), 概率论 (4.0), 文本挖掘 (4.0), 深度学习 (4.0), 人工智能 (4.0).
- ▶ 完成多项自然语言处理相关研究项目, 且有一篇计算数学论文已发表在 SCI 一区杂志.
- ▶ 曾获高中全国物理竞赛省一等奖 (实验部分全省第 8 名), 自学大学与研究生阶段物理课程, 数理基础扎实.

目录

个人简介

研究经历

偏微分方程的有限元方法研究

预训练模型的可解释性研究

Kaggle Question Answering 项目

Kaggle Sentence Classification 项目

金融知识图谱自动化构建项目

总结

有限元方法简介

Finite element space and grids:

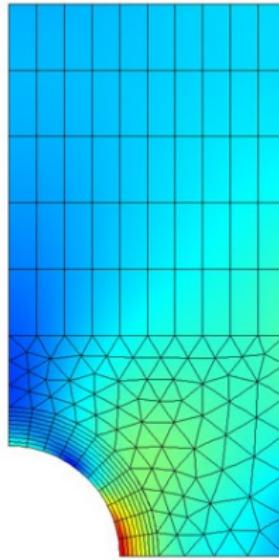
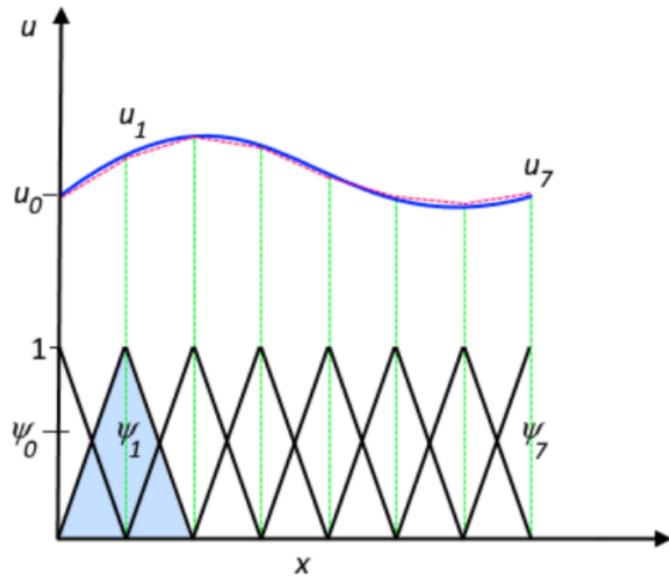


图: Linear basis in 1D; Grids

主要内容

Equation:

$$\begin{cases} \varepsilon^2 \Delta^2 u - \Delta u = f & \text{in } \Omega, \\ u = \partial_n u = 0 & \text{on } \partial\Omega, \end{cases}$$

- ▶ Original ways to solve:
 - ▶ Conforming elements: computational expensive
 - ▶ Non-conforming elements: isn't convergent

主要内容

Our work:

- ▶ Modified the right hand side via projection:

$$(\nabla w_h, \nabla \chi_h) = (f, \chi_h) \quad \forall \chi_h \in W_h$$

$$\varepsilon^2 a_h(u_{h0}, v_h) + b_h(u_{h0}, v_h) = (\nabla w_h, \nabla_h v_h) \quad \forall v_h \in V_{h0}$$

- ▶ Decoupled the left hand side into four simple equations:

$$(\operatorname{curl}_h z_h, \operatorname{curl}_h v_h) = (\nabla w_h, \nabla_h v_h) \quad \forall v_h \in V_{h0}$$

$$(\phi_h, \psi_h) + \varepsilon^2 (\nabla_h \phi_h, \nabla_h \psi_h) + (\operatorname{div}_h \psi_h, p_h) = (\operatorname{curl}_h z_h, \psi_h) \quad \forall \psi_h \in V_{h0}^{CR}$$

$$(\operatorname{div}_h \phi_h, q_h) = 0 \quad \forall q_h \in Q_h$$

$$(\operatorname{curl}_h u_{h0}, \operatorname{curl}_h \chi_h) = (\phi_h, \operatorname{curl}_h \chi_h) \quad \forall \chi_h \in V_{h0}$$

实验结果

- ▶ Equations solved efficiently with the simplest elements.
- ▶ Final paper published in *Journal of Scientific Computing*.

h	#dofs	Eq.(5.1)	Eq.(5.7a)	Eq.(5.7b)-(5.7c)	Eq.(5.7d)
		steps	steps	steps	steps
2^{-1}	24	1	1	16	1
2^{-2}	112	1	4	27	3
2^{-3}	480	4	5	34	5
2^{-4}	1984	6	7	34	7
2^{-5}	8064	6	9	41	9
2^{-6}	32512	7	11	43	11
2^{-7}	130560	7	14	44	14
2^{-8}	523264	9	17	46	17
2^{-9}	2095104	9	20	50	21
2^{-10}	8384512	12	27	55	27

图: Robust iteration steps when solving

目录

个人简介

研究经历

偏微分方程的有限元方法研究

预训练模型的可解释性研究

Kaggle Question Answering 项目

Kaggle Sentence Classification 项目

金融知识图谱自动化构建项目

总结

研究背景

Existing methods

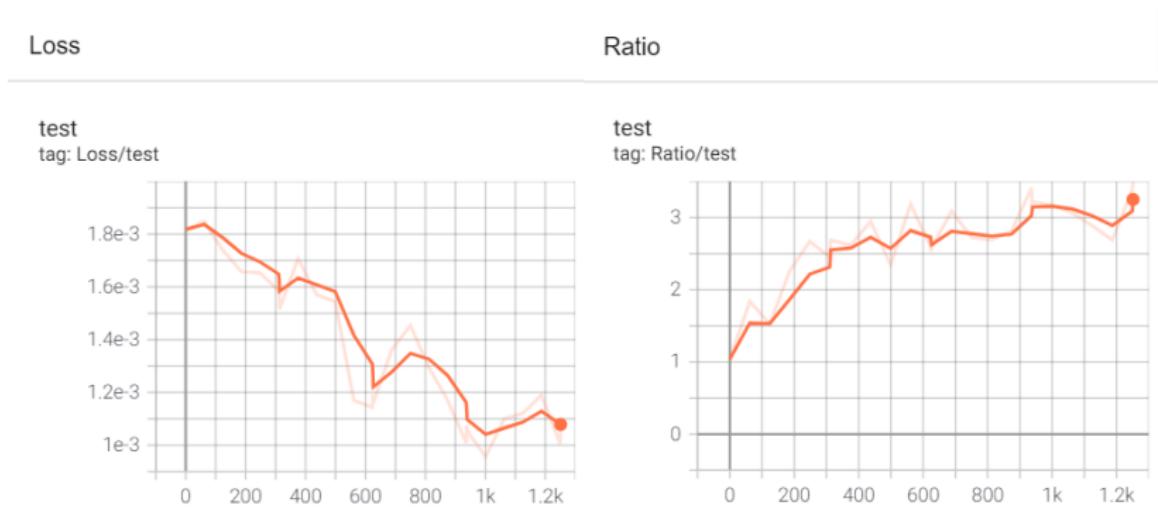
- ▶ Gradient based methods: gradient, dot product with embeddings, integrated gradient ...
- ▶ Perturbation based methods: input reduction, adversarial perturbations ...

Trying to explain the **relations between words** that model is learning during the training process.

$$\begin{aligned} \text{Loss}(x_1, \dots, x_d) = & f(a_1, \dots, a_d) + \sum_{j=1}^d \frac{\partial f(a_1, \dots, a_d)}{\partial x_j} (x_j - a_j) \\ & + \frac{1}{2!} \sum_{j=1}^d \sum_{k=1}^d \frac{\partial^2 f(a_1, \dots, a_d)}{\partial x_j \partial x_k} (x_j - a_j) (x_k - a_k) \\ & + \frac{1}{3!} \sum_{j=1}^d \sum_{k=1}^d \sum_{l=1}^d \frac{\partial^3 f(a_1, \dots, a_d)}{\partial x_j \partial x_k \partial x_l} (x_j - a_j) (x_k - a_k) (x_l - a_l) + \dots \end{aligned}$$

实验结果

- ▶ Model learning relations between "essential" words?
- ▶ $\frac{\partial Loss}{\partial E_i E_j}$ during training BERT-Base:



实验结果

► Sentences:

1. Two young boys of opposing teams play football₁, while wearing full protection uniforms and helmets.
2. Boys play football₂.

	word	most relavent	score
0	[SEP]	[SEP]	0.044089
2	[CLS]	[SEP]	0.025706
4	[CLS]	[SEP]	0.024833
6	[SEP]	.	0.014351
8	football11	[SEP]	0.014056
10	.	[SEP]	0.013824
12	[SEP]	football12	0.011118
14	[SEP]	football11	0.010987
16	football12	[SEP]	0.010924
18	[SEP]	helmets	0.010910

图: At the beginning of training

实验结果

▶ Sentences:

1. Two young boys of opposing teams play football₁, while wearing full protection uniforms and helmets.
2. Boys play football₂.

	word	most relavent	score
0	football11	football12	0.066224
2	and	football12	0.053644
4	football12	protection	0.034782
6	boys	football12	0.025395
8	football12	play	0.021575
10	uniforms	football12	0.018982
12	[CLS]	football12	0.018318
14	[SEP]	football12	0.013509
16	helmets	football12	0.012761
18	football11	protection	0.012189

图: Trained for 200 batches

目录

个人简介

研究经历

偏微分方程的有限元方法研究

预训练模型的可解释性研究

Kaggle Question Answering 项目

Kaggle Sentence Classification 项目

金融知识图谱自动化构建项目

总结

项目介绍

数据集介绍

- ▶ Used the Natural Questions (NQ) (**Kwiatkowski et al., 2019**) dataset from Google AI.
- ▶ Each example is comprised of a google query and a corresponding Wikipedia page.

主要方案

- ▶ Fine-tuning on SQuAD 2.0
- ▶ Mixed Precision Training
- ▶ Hard Negative Sampling
- ▶ Sifting candidates

实验结果

Model	Public F1	Private F1
Kaggle Best	0.713	0.717
BERT Base baseline	0.516	0.482
BERT Base (Hard Negative Sampling)	0.579	0.574
BERT Sifted* → ALBERT xlarge	0.640	0.659
Sifted → BERT Base+ALBERT _(ensemble)	0.665	0.666
Sifted → BERT Large+ALBERT _(ensemble)	0.738	0.718

* Sifted here stands for first using BERT base to sift candidates

目录

个人简介

研究经历

偏微分方程的有限元方法研究

预训练模型的可解释性研究

Kaggle Question Answering 项目

Kaggle Sentence Classification 项目

金融知识图谱自动化构建项目

总结

主要内容

数据集介绍

- ▶ Complaints from citizens. The task is to predict label (200 classes in total).

主要方案

- ▶ Pre-trained models on similar dataset THUCNews.
- ▶ Adopted **focal loss** for the long-tailed distributed labels.
- ▶ Designed a **auxiliary sentence pair task**.
- ▶ Also tried adversarial training, data augmentation, pesudo labels, adding other layers after BERT, using RoBERTa-wwm, ERNIE, etc.

辅助任务设计

- ▶ The original classification task failed to utilize information in the labels.
- ▶ An auxiliary sentence pair task is then designed.

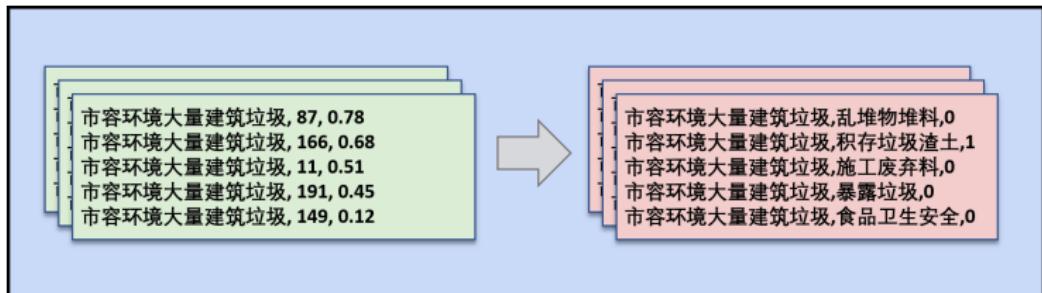


图: Generating pair data

实验结果

- ▶ Highest score in class (led by 0.4%).

表: Selected Experiment Results

Model	Public score	Private score
ERNIE ¹	0.7981	0.8000
ERNIE ²	0.7995	0.8030
ERNIE ³	0.8049	0.8010

¹Original Task: text classification

²Focal Loss

³Auxiliary Task: sentence pair classification

目录

个人简介

研究经历

偏微分方程的有限元方法研究

预训练模型的可解释性研究

Kaggle Question Answering 项目

Kaggle Sentence Classification 项目

金融知识图谱自动化构建项目

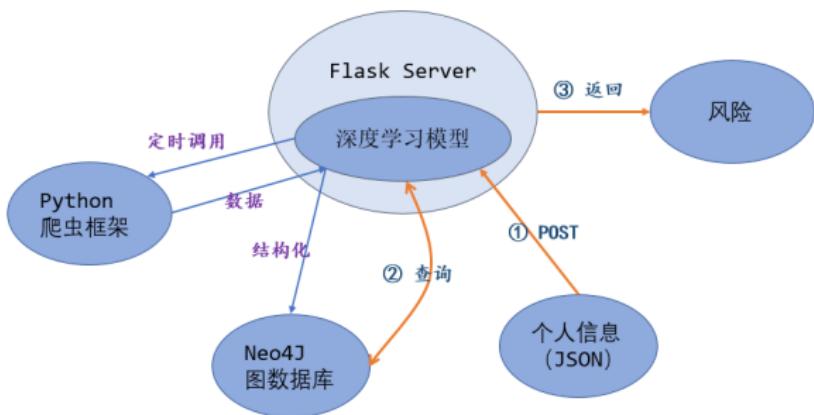
总结

项目简介

- ▶ 与平安公司导师合作的应用型研究项目，旨在帮助信贷面审人员快速识别申请人相关的潜在风险。
- ▶ 作为项目负责人，每周组织组会学习分享实体识别，关系抽取等知识图谱构建的关键技术。
- ▶ 使用 BERT 抽取最新资讯中涉及的风险事件，并将风险标签与资讯文本共同输入 BERT-BiGRU，进一步提升了实体识别任务的准确率。

项目简介

- ▶ 完成了从数据的自动化爬取, 构建知识图谱到 Neo4j 图数据库存储和前端应用的完整流程.
- ▶ 被选为校“大学生创新创业计划”优秀项目, 并在学术论坛登台展示 (2%), 目前正进一步参加上海市相关竞赛.



总结

- ▶ 具有良好的数学基础和丰富 NLP 项目经验.
- ▶ 积极乐观热爱研究, 自学能力强能主动探索.
- ▶ 希望有机会能在中文信息处理实验室继续提升自己, 为中文文本的研究贡献力量.



Thanks for your attention!