



# Machine Learning

## Lab2

Fall 2022

Instructor: Xiaodong Gu





# 任务

---

## 三选一

- 实现基于GPT2的程序生成系统
- 实现基于卷积神经网络的图像分割系统
- 阅读AI论文, 写文献阅读报告(文献综述+创新点+思路)

## 团队任务：

- 1-3名组员
- 其中一位组员提交
- 文件名SID1\_NAME1\_SID2\_NAME2\_SID3\_NAME3.zip

## 截止日期：

2022年12月24日

有问题请联系助教石雨凌。yuling.shi@sjtu.edu.cn



# 基于GPT2的程序生成

---

- 任务：

参考课堂讲解的 GPT2 finetuning 实现程序自动生成系统。使用 Canvas 上传的数据集，调试代码并进行调参实验。最后对生成的程序样例进行展示。



# AI编程

---

注意事项：

1. 需要配置开发环境, 如：PyCharm+Anaconda, python=3.7, torch=1.10.1
2. 如果使用 gpt2 训练较慢, 可以适当减小 seq\_len、使用 distilgpt2 (<https://huggingface.co/distilgpt2>)
3. 读取示例数据的代码已经给出, 可以参考此代码构建自己的 Dataset 类
4. DataLoader 中可以设置如 num\_workers=4 提高计算效率
5. 在架构较新的GPU上 (18 年后发布的NIVDIA GPU), 可以使用混合精度训练提高效率
6. 为了减轻工作量, 可以参考开源代码 (报告中注明来源), 但要有自己的发挥。



# AI编程

---

- 提交：

- 代码及运行说明
- 实验报告。包括但不限于系统设计、训练过程（如loss曲线）、调参实验及结果（不同参数下的perplexity等指标）、样例展示等。
  - **训练结果指标仅作为一项参考，不是主要的评价标准！**

- 系统设计

- 模型设计

- 训练方法

- 实验结果

- 训练过程（如loss曲线）

- 调参实验及结果（如模型在不同超参数下的精确度）

- 生成代码展示



# AI编程

- 评分：综合评价功能、质量和工作量

## 功能：

代码无法运行



能完成功能、鼓励举一反三、尝试新方法



## 质量：

生成内容无意义、报告质量低



生成可读程序、报告完整思路清晰



## 工作量：

直接提交示例代码或  
完全照搬开源代码



显示出对代码有理解、重构、或改进



# CNN图像分割

- 任务：

参考 <https://www.kaggle.com/code/gokulkarthik/image-segmentation-with-unet-pytorch>

在所提供数据集上完成基于CNN的图像分割任务。

- 数据集简介：

- 共有 715 张图片 (data/images) 和对应的分割标注 (data/masks), 标注共有 9 个类别





# CNN图像分割

---

注意事项及建议：

1. 需要配置开发环境, 如：PyCharm+Anaconda, python=3.7, torch=1.10.1
2. 读取示例数据的代码已经给出,可以参考此代码构建自己的 Dataset 类
3. DataLoader 中可以设置如 num\_workers=4 提高计算效率
4. 在架构较新的GPU上 (18 年后发布的NIVDIA GPU), 可以使用混合精度训练提高效率
5. 为了减轻工作量，可以参考开源代码（报告中注明来源），但要有自己的发挥。





# CNN图像分割

- 提交：

- 代码及运行说明
- 运行结果：所能调试出的最高 IOU 及对应的 epoch 数
  - 准确率仅作为一项参考，不是主要的评价标准！
- 实验报告。包括但不限于以下：

- 数据处理

- 图片预处理 (可以弱化)

- 系统设计

- 模型设计

- 训练方法

- 实验结果

- 预测准确率

- 训练过程 (如loss曲线, tensor board)

- 调参实验及结果分析 (如模型在不同训练参数下的准确率及分析)

- 参数量



# CNN图像分割

- 评分：综合评价功能、质量和工作量

## 功能：

代码无法运行



完成分割功能，鼓励使用所学的知识综合设计网络



## 质量：

设计不合理、准确率低



模型合理、运行效率高、准确率高



## 工作量：

直接调试示例代码或完全照搬开源代码



有详细的参数调试及分析报告





# 文献阅读

---

精读自己感兴趣的AI论文，提交一个论文汇报(Research Statement)。

## 要求：

- 来自顶级会议(ICML, NeurIPS, ICLR, AAAI, IJCAI, EMNLP, ACL, ICCV, CVPR, ECCV, NAACL等)
- 发表于近3年 (2019及以后)
- 除了汇报的论文，可能还需要阅读若干相关文献。



# 文献阅读

---

提交：文献汇报ppt,

内容包括：

Background
Related Works
Approach
Implementation Details
Evaluation
New Thoughts (Limitations, Improvement, Applications, etc)
References

根据报告内容和质量评分，特别是New Thoughts



- **Background**

No strict requirement. You may consider:

- Investigate the technical trend of the same topic in the industry.
- What you have known about this topic.
- Your understanding about this topic.
- ...

- **Related Work**

Important technologies (papers) for this research topic.

List 2-3 papers and briefly describe the key ideas.

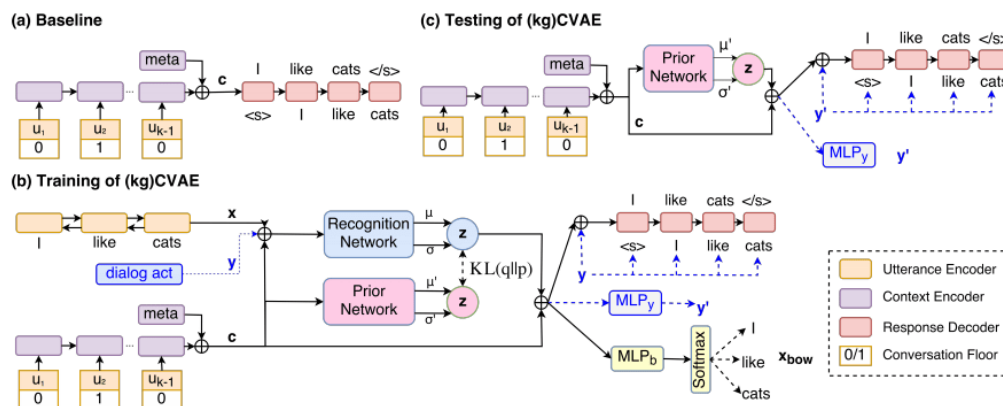
- **Motivation**

- what is the **main problem** of existing approaches?
- how do the authors address the problem?
- what is the **key idea** of the new approach?

## • Approach

Describe the approach using diagrams and descriptions (like how we introduced the Seq2Seq, Attention, Transformer, etc in the class).

For example:



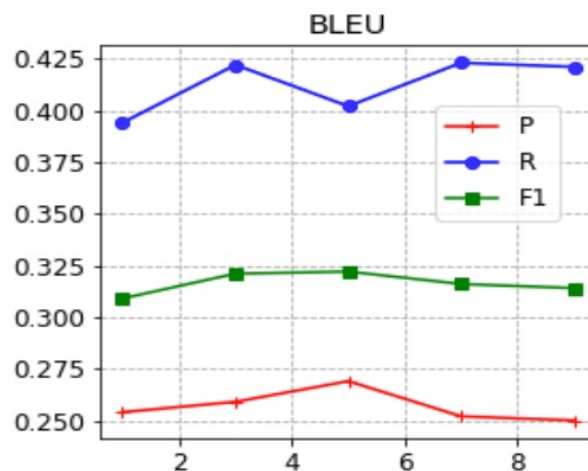
## • Implementation Details

key components and algorithms (e.g., encoder, decoder, etc).

- **Evaluation**

Show the experimental setup and results such as:

- data sets, baseline models, performance metrics, etc
- **Tables and curves** of results and the **comparison** of various models.
- brief **descriptions** about the results and comparisons.
- ...

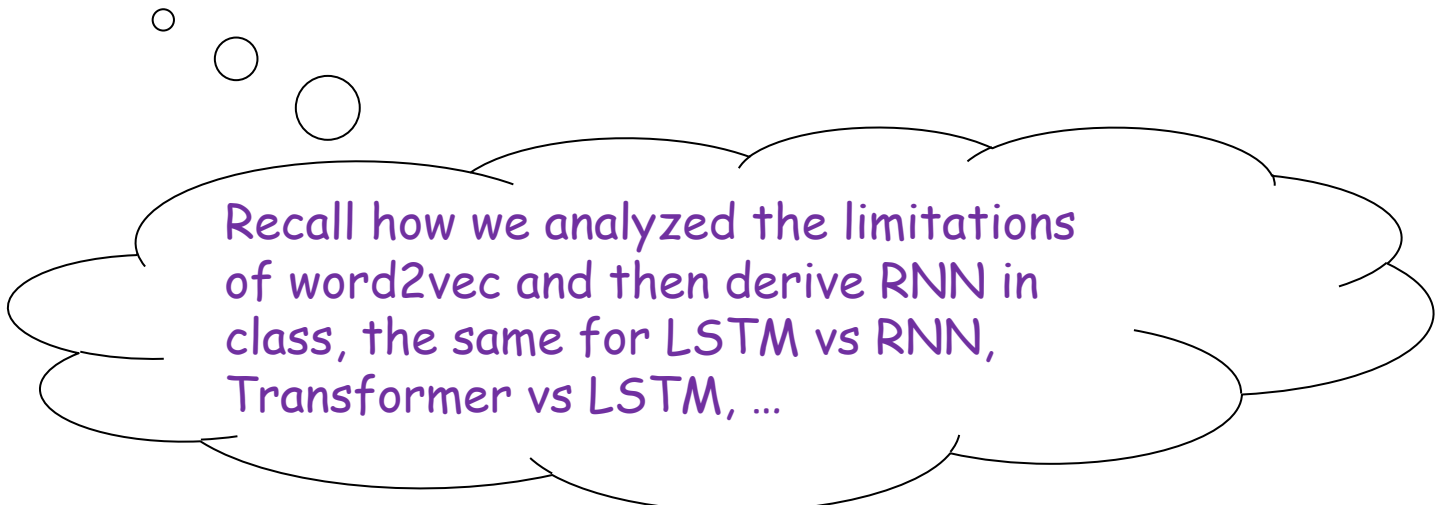




- **New Thoughts**

Provide your thoughts after reading this paper, such as:

- Limitations of this paper
- Application of the technology to an interesting task?
- Your new ideas with some details



Recall how we analyzed the limitations of word2vec and then derive RNN in class, the same for LSTM vs RNN, Transformer vs LSTM, ...



# Tips

---



- Your programs should be written in such a way that the TA can easily verify the results reported by you.
- Your presentation should be clear and comprehensive so that customers (TAs) will buy (give high score to) your product.



