实验 4 BERT 环境配置

组员：陈德康、危弘毅、徐昌华、朱宣玮、郑坤武

组长　学号： 202300130236　**班级：**数据 23
组员 2 学号： 202300130235　**班级：**数据 23
组员 3 学号： 202300130228　**班级：**数据 23
组员 4 学号： 202300300198　**班级：**数据 23
组员 5 学号： 202200150184　**班级：**公信 22

2025 年 11 月 7 日

1 实验目的

本实验旨在配置基于 PyTorch 的 BERT 模型运行环境，为后续自然语言处理任务提供基础支撑。实验重点包括：

- 搭建支持 BERT 模型训练的深度学习环境，包括 PyTorch、Transformers 等核心库
- 验证关键组件的版本兼容性，确保 CUDA 加速功能正常可用
- 准备数据处理和模型评估所需的辅助工具库
- 建立完整的 NLP 实验技术栈，为后续文本分类、情感分析等任务奠定基础

2 实验环境

- **操作系统：** Linux (AutoDL 容器环境)
- **Python 版本：** 3.8
- **深度学习框架：**
  - PyTorch 2.4.1+cu121 (CUDA 12.1)
  - Transformers 4.46.3
- **数据处理库：**
  - Datasets 3.1.0
  - Pandas 2.0.3
  - NumPy 1.24.2
- **机器学习工具：**
  - Scikit-learn 1.3.2
  - tqdm 4.67.1 (用作进度条工具)

3 具体实验步骤与结果分析

首先配置 PyTorch 深度学习框架，选择与 CUDA 12.1 兼容的版本 2.4.1。该版本提供了强大的 GPU 加速支持，包含完整的 CUDA 工具链，如 cublas、cudnn、cusparse 等计算库，确保 BERT 模型能够在 NVIDIA GPU 上高效运行。通过 pip show 命令验证安装成功，显示 PyTorch 已正确安装在 Python 3.8 环境中。

```
root@autodl-container-1f554fbcbc-3e3343a5:/# pip show torch
Name: torch
Version: 2.4.1+cu121
Summary: Tensors and Dynamic neural networks in Python with strong GPU acceleration
Home-page: https://pytorch.org/
Author: PyTorch Team
Author-email: packages@pytorch.org
License: BSD-3
Location: /root/miniconda3/lib/python3.8/site-packages
Requires: networkx, nvidia-cublas-cu12, nvidia-cuda-nvrtc-cu12, nvidia-cuda-runtime-cu12, nvidi
a-cuda-cupti-cu12, nvidia-cudnn-cu12, nvidia-cusparse-cu12, nvidia-nvtx-cu12, nvidia-cusolver-c
u12, triton, sympy, jinja2, nvidia-curand-cu12, typing-extensions, filelock, fsspec, nvidia-cuf
ft-cu12, nvidia-nccl-cu12
Required-by: torchvision, torchaudio
```

图 1: 基础库安装

还需要安装 Hugging Face Transformers 库版本 4.46.3，这是运行 BERT 模型的核心依赖。该版本支持丰富的预训练模型和先进的 NLP 架构，提供了 BERT 及其变体的标准接口。安装过程中同时配置了必要的依赖项，包括 tokenizers 用于高效的分词处理、safetensors 用于模型安全加载、huggingface-hub 用于模型下载管理。验证显示所有依赖包均已正确安装，位置在 conda 环境的 site-packages 目录下。

```
root@autodl-container-1f554fbcbc-3e3343a5:/# pip show transformers
Name: transformers
Version: 4.46.3
Summary: State-of-the-art Machine Learning for JAX, PyTorch and TensorFlow
Home-page: https://github.com/huggingface/transformers
Author: The Hugging Face team (past and future) with the help of all our contributors (https://
github.com/huggingface/transformers/graphs/contributors)
Author-email: transformers@huggingface.co
License: Apache 2.0 License
Location: /root/miniconda3/lib/python3.8/site-packages
Requires: regex, numpy, filelock, requests, safetensors, tqdm, huggingface-hub, pyyaml, packagi
ng, tokenizers
Required-by:
```

图 2: Transformer 模型库安装

安装 Datasets 库版本 3.1.0，用于高效加载和处理大规模文本数据集。该库提供了标准化的数据接口，支持多种数据格式的读取和预处理，特别适合 BERT 模型的训练数据准备。安装过程中集成了 pyarrow 作为后端引擎，提供了列式存储的高效数据访问，同时包含 pandas、numpy 等数据处理组件的完整支持，为数据预处理流水线提供了完整的技术支撑。

```
root@autodl-container-1f554fbcbc-3e3343a5:/# pip show datasets
Name: datasets
Version: 3.1.0
Summary: HuggingFace community-driven open-source library of datasets
Home-page: https://github.com/huggingface/datasets
Author: HuggingFace Inc.
Author-email: thomas@huggingface.co
License: Apache 2.0
Location: /root/miniconda3/lib/python3.8/site-packages
Requires: numpy, huggingface-hub, pyyaml, tqdm, requests, pandas, dill, filelock, fsspec, packa
ging, multiprocess, xxhash, aiohttp, pyarrow
Required-by:
```

图 3: 数据集处理工具配置

配置 Scikit-learn 版本 1.3.2 作为机器学习工具库，用于模型评估、数据划分和传统机器学习算法的对比实验。该版本提供了稳定的分类评估指标和交叉验证功能，是模型性能评估的重要工具。同时安装 Pandas 2.0.3 用于结构化数据处理，NumPy 1.24.2 用于数值计算，两者共同构成了数据处理的基础设施，支持从原始数据到模型输入的完整转换流程。

```
root@autodl-container-1f554fbcbc-3e3343a5:/# pip show scikit-learn
Name: scikit-learn
Version: 1.3.2
Summary: A set of python modules for machine learning and data mining
Home-page: http://scikit-learn.org
Author: None
Author-email: None
License: new BSD
Location: /root/miniconda3/lib/python3.8/site-packages
Requires: joblib, threadpoolctl, scipy, numpy
Required-by:
```

图 4: 机器学习工具链完善

```
root@autodl-container-1f554fbcbc-3e3343a5:/# pip show pandas
Name: pandas
Version: 2.0.3
Summary: Powerful data structures for data analysis, time series, and statistics
Home-page: None
Author: None
Author-email: The Pandas Development Team <pandas-dev@python.org>
License: BSD 3-Clause License
```

图 5: Pandas 安装验证

验证 NumPy 1.24.2 的安装状态，这是整个 Python 科学计算生态的基础包。结果显示 NumPy 已正确安装，并被 transformers、torchvision、scikit-learn、pandas 等所有主要依赖包所使用，确保了数值计算的一致性和稳定性。该版本提供了高效的数组操作和线性代数运算，是张量计算和矩阵操作的基础。

```
root@autodl-container-1f554fbcbc-3e3343a5:/# pip show numpy
Name: numpy
Version: 1.24.2
Summary: Fundamental package for array computing in Python
Home-page: https://www.numpy.org
Author: Travis E. Oliphant et al.
Author-email: None
License: BSD-3-Clause
Location: /root/miniconda3/lib/python3.8/site-packages
Requires:
Required-by: transformers, torchvision, tensorboard, scipy, scikit-learn, pyarrow, pandas, matp
lotlib, datasets, contourpy
```

图 6: NumPy 安装验证

安装 tqdm 4.67.1 作为进度显示工具，在模型训练和数据处理过程中提供直观的进度反馈。该库被 transformers、huggingface-hub 和 datasets 等多个组件所依赖，在长时间运行的训练任务中能够有效监控进度，提升实验的可观察性和用户体验。

```
root@autodl-container-1f554fbcbc-3e3343a5:/# pip show tqdm
Name: tqdm
Version: 4.67.1
Summary: Fast, Extensible Progress Meter
Home-page: None
Author: None
Author-email: None
License: MPL-2.0 AND MIT
Location: /root/miniconda3/lib/python3.8/site-packages
Requires:
Required-by: transformers, huggingface-hub, datasets
```

图 7: NumPy 安装验证

4 实验总结与收获

本次实验成功配置了基于 PyTorch 的 BERT 模型运行环境，建立了从数据处理、模型训练到性能评估的完整技术栈。通过系统的版本管理和依赖检查，确保了深度学习组件的兼容性和稳定性。CUDA 加速环境的正确配置为大规模模型训练提供了硬件支持，而丰富的数据处理工具则为文本数据的预处理和特征工程提供了便利。该环境具有良好的扩展性，为后续的实验打下基础。