

NLP專案的一生

ADL 2023 Fall - Recitation 2

林彥廷

你會學到什麼？

- 從資料到模型開發到 Demo 的開發流程
- 常用工具與套件

專案的一生



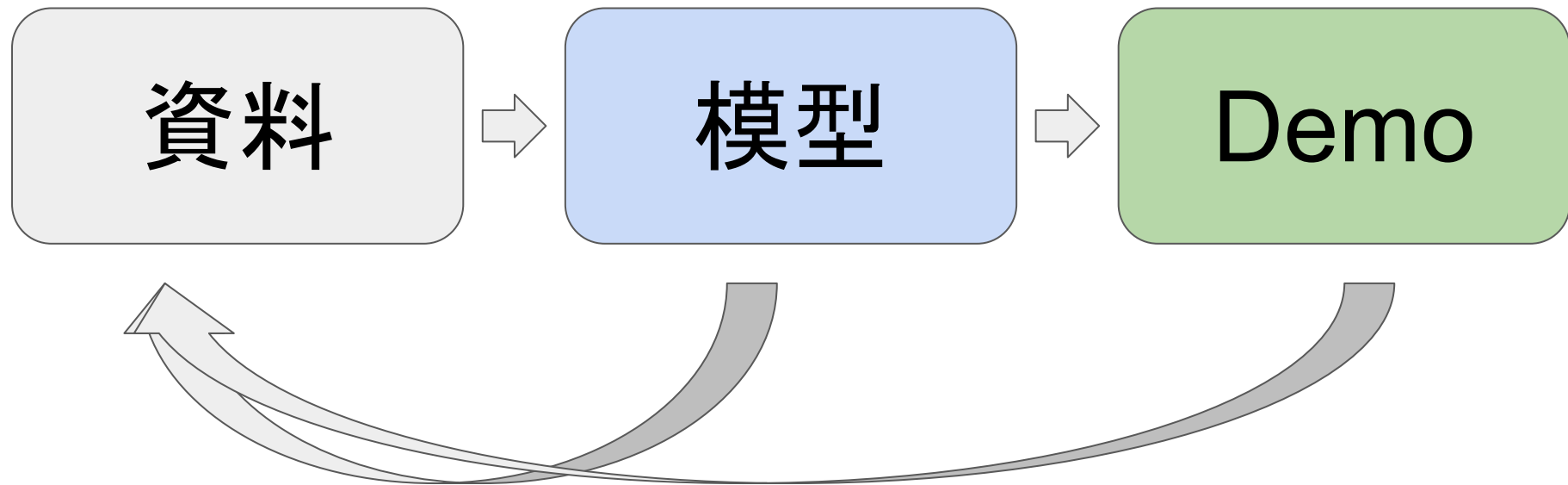
專案的一生

資料



emo

專案的一生



專案的一生



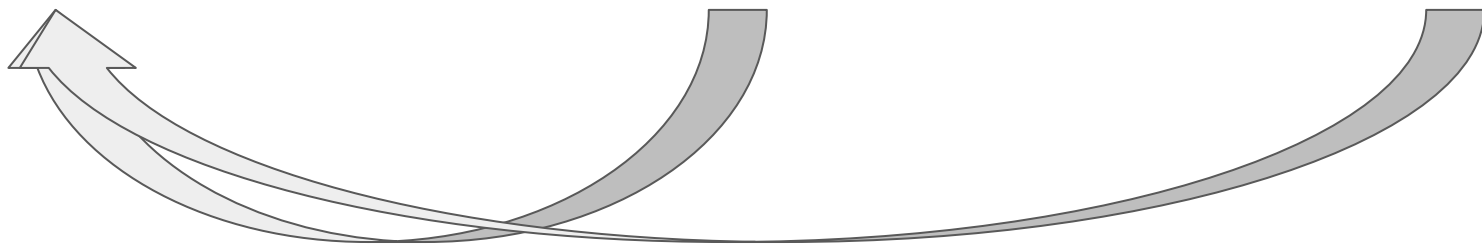
資料



模型



Demo



資料

- 資料收集
- 資料清洗、驗證
- 資料標注

資料

- 資料收集
 - 網路爬蟲、客戶資料、自行生成、GPT4!
- 資料清洗、驗證
- 資料標注

資料

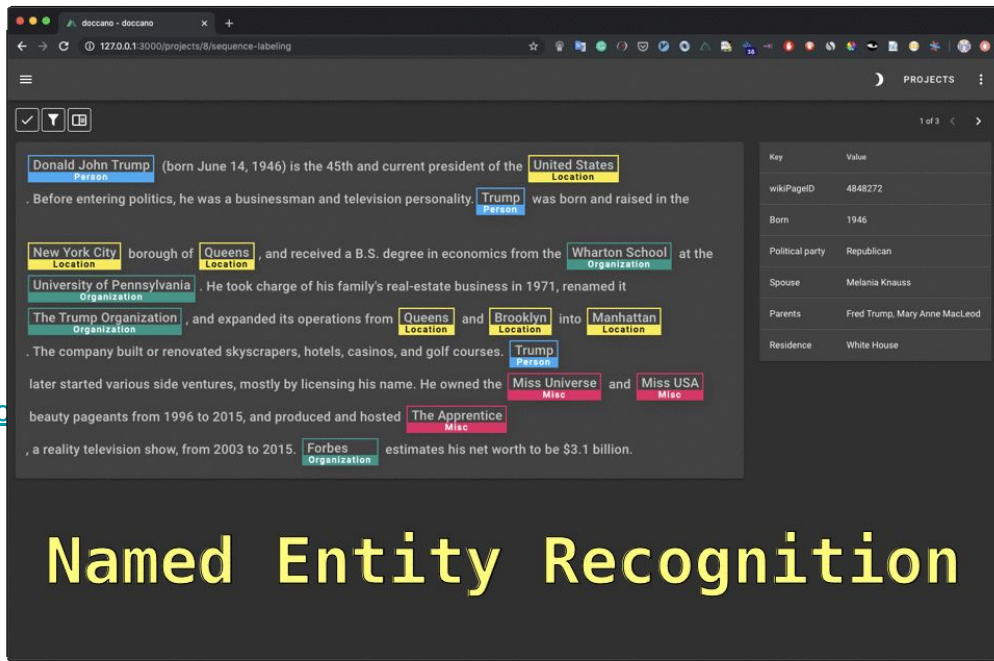
- 資料收集
 - 網路爬蟲、客戶資料、自行生成、GPT4!
 - [\[2207.11363\] Knowledge-Grounded Conversational Data Augmentation with Generative Conversational Networks](#)
 - [\[2302.05096\] Selective In-Context Data Augmentation for Intent Detection using Pointwise V-Information](#)

資料

- 資料收集
- 資料清洗、驗證
- 資料標注

資料

- 資料收集
- 資料清洗、驗證
- 資料標注
 - 開源標注平台: Docanno
 - [GitHub - doccano/doccano: Open source anno](#)



The screenshot displays the Docanno web interface for sequence labeling. The main text area shows a paragraph about Donald Trump with various entities highlighted and labeled. The labels include 'Person' (Donald John Trump), 'Location' (United States, New York City, Queens, Manhattan), 'Organization' (Wharton School, University of Pennsylvania, The Trump Organization, Forbes), and 'Misc' (Miss Universe, Miss USA, The Apprentice). The sidebar on the right shows a table of key-value pairs for the document.

Key	Value
wikiPageID	4848272
Born	1946
Political party	Republican
Spouse	Melania Knauss
Parents	Fred Trump, Mary Anne MacLeod
Residence	White House

Named Entity Recognition

資料

- 資料收集
- 資料清洗、驗證
- 資料標注

○ [Scale AI](#)

Nucleus

Create Dataset +

MS COCO


- Overview
- Charts
- Slices (12)
- Autotags

Models >

- Jobs
- Guides
- Docs

Other Products

MS COCO ds_bwm61zzb8mjksanms4wg



MS COCO

See All Insights


ITEMS	SLICES	AUTOTAGS	MODEL RUNS
123289	12	0	2

Object Class Distribution

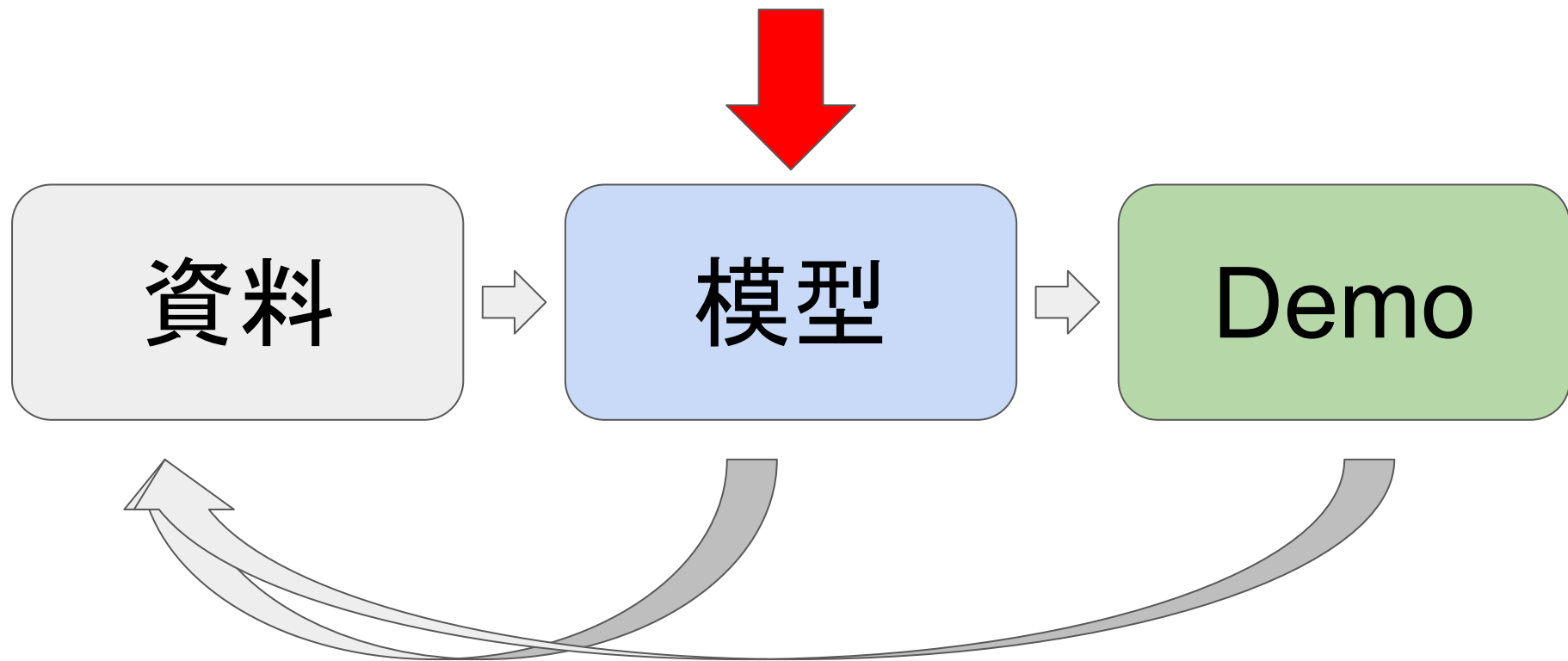
Ground Truth

Q Search

☒ Showing 20 of 80 Results



專案的一生



模型

- NLP 可以做什麼事情？

模型

- NLP 可以做什麼事情？
- NLP guy 如何看待問題？

模型

- 分類**整句**
- 分類句中每個**單詞**
- **生成**文本內容
- 從文本中**提取**答案

模型

- 分類**整句**
- 分類**單詞**
- **生成**內容
- 文本**提取**

模型

- 分類整句：
- - 情感分析：「這部電影很棒！」(正面)
- - 垃圾郵件檢測：「您已經贏得一百萬美元！」(垃圾郵件)

模型

- 分類句中每個單詞：
- - 語法成分：「他(代詞)跑(動詞)得(副詞)快(形容詞)」
- - 命名實體：「台灣(地點)的蔡英文(人)是總統(職位)」

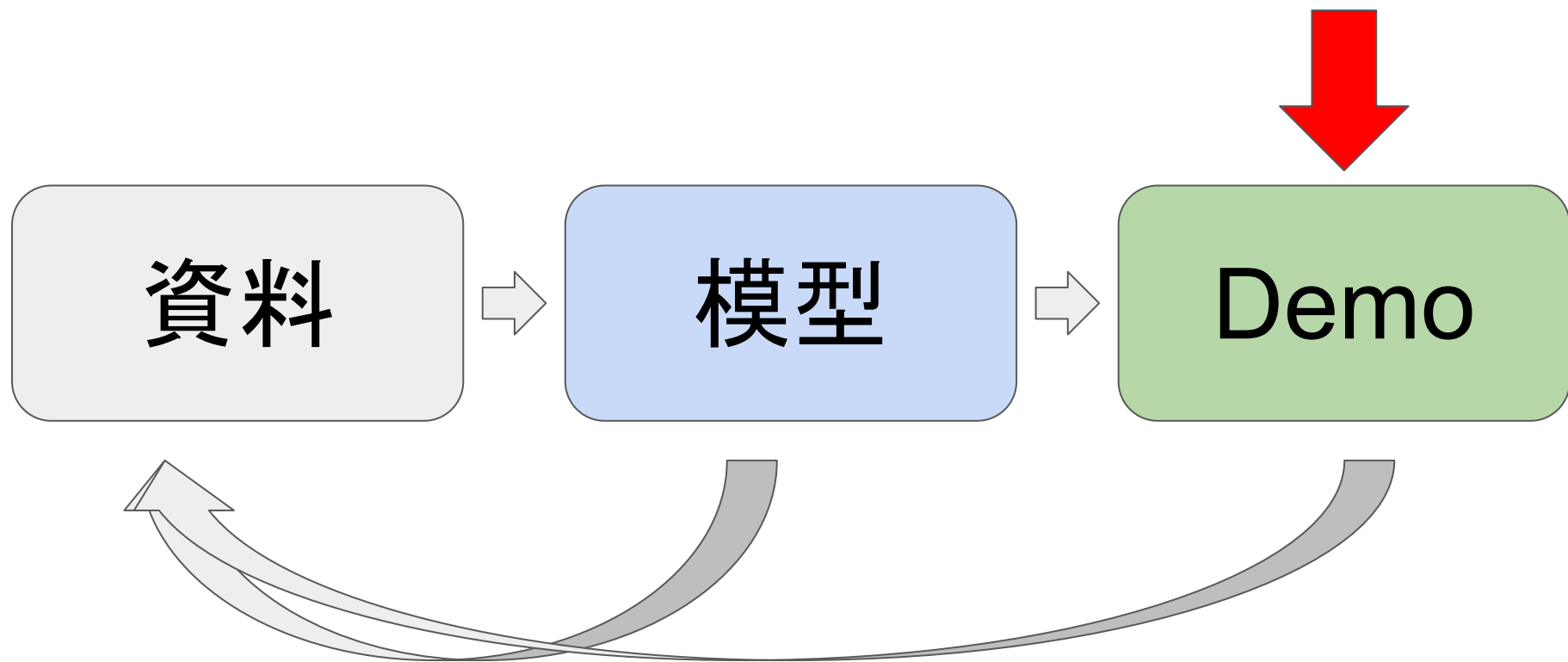
模型

- 生成文本內容：
 - - 自動生成: 提示「天氣如何」, 回答「今天很晴朗」
 - - 填充遮蔽單詞: 「他是一個__醫生__」(填充: 醫生)
 - - 翻譯: 「你好」(中文) → 「Hello」(英文)
 - - 摘要: 「這部電影講述了一個男孩的成長故事」 → 「男孩的成長故事」

模型






- 從文本中提取答案：
- - 問題：「首都是什麼？」，上下文：「法國的首都是巴黎」，
答案：「巴黎」
- - 問題：「誰發明了電燈泡？」，上下文：「湯瑪斯·愛迪生發明了電燈泡」
答案：「湯瑪斯·愛迪生」




專案的一生




Demo






<http://twllm.com>


 Spaces |  yentinglin/Taiwan-LLaMa2  like 45  Running 

 App  Files  Commun

 Tw Llama Demo - a Hugging Face Space by yentinglin huggingface.co

Language Models for Taiwanese Culture

 [Online Demo](#) •  [HF Repo](#) •  [Twitter](#) •  [\[Paper Coming Soon\]](#) •  [Github Repo](#)



Taiwan-LLaMa is a fine-tuned model specifically designed for traditional mandarin applications. It is built upon the LLaMa 2 architecture and includes a pretraining phase with over 5 billion tokens and fine-tuning with over 490k multi-turn conversational data in Traditional Mandarin.

Key Features

- Traditional Mandarin Support:** The model is fine-tuned to understand and generate text in Traditional Mandarin, making it suitable for Taiwanese culture and related applications.
- Instruction-Tuned:** Further fine-tuned on conversational data to offer context-aware and instruction-following responses.
- Performance on Vicuna Benchmark:** Taiwan-LLaMa's relative performance on Vicuna Benchmark is measured against models like GPT-4 and ChatGPT. It's particularly optimized for Taiwanese culture.
- Flexible Customization:** Advanced options for controlling the model's behavior like system prompt, temperature, top-p, and top-k are available in the demo.

Model Versions

Different versions of Taiwan-LLaMa are available:

- **Taiwan-LLaMa v1.0 (This demo):** Optimized for Taiwanese Culture
- **Taiwan-LLaMa v0.9:** Partial instruction set
- **Taiwan-LLaMa v0.0:** No Traditional Mandarin pretraining

The models can be accessed from the provided links in the Hugging Face repository.

Try out the demo to interact with Taiwan-LLaMa and experience its capabilities in handling Traditional Mandarin!

Demo



FastAPI



gradio

專案的一生

資料



emo

專案的一生



以前的
作業一



分類整句



Gradio

資料

- 以前的作業一有分類整句、單詞的任務
- <https://github.com/adamlin120/ADL-HW1-NTU-2021/tree/master/data/intent>

```
[  
  {  
    "text": "i need you to book me a flight from ft lauderdale to houston on southwest",  
    "intent": "book_flight",  
    "id": "train-0"  
  },  
  {  
    "text": "my check engine light is on and i need to take a look at it",  
    "intent": "schedule_maintenance",  
    "id": "train-1"  
  },  
  {  
    "text": "is the company party on my list of reminders",  
    "intent": "reminder",  
    "id": "train-2"  
  },  
]
```

資料

- 以前的作業一有分類整句、單詞的任務
- <https://github.com/adamlin120/ADL-HW1-NTU-2021/tree/master/data/intent>
- 今天只用“分類整句”

資料

- 以前的作業一有分類整句、單詞的任務
- <https://github.com/adamlin120/ADL-HW1-NTU-2021/tree/master/data/intent>
- 今天只用“分類整句”
- 上傳到 [Huggingface Dataset](#)

資料

訓練集

驗證集

測試集

資料

模型訓練時...

訓練集

可以看

驗證集

測試集

資料

模型訓練時...

訓練集

可以看

驗證集

不可以看

測試集

資料

模型訓練時...

訓練集

可以看

驗證集

不可以看

測試集

絕對不可以看

資料

模型訓練時...

怎麼測試？

訓練集

可以看

驗證集

不可以看

測試集

絕對不可以看

資料

模型訓練時...

怎麼測試？

訓練集

可以看

驗證集

不可以看

訓練時驗證

測試集

絕對不可以看

資料

模型訓練時...

怎麼測試？

訓練集

可以看

驗證集

不可以看

訓練時驗證

測試集

絕對不可以看

訓練後驗證

資料

數量？

訓練集

最多

驗證集

~10~30%

測試集



~10~30%


資料





訓練集




驗證集

測試集

Datasets: yentinglin/ntu_adl_recitation   like 0

License:  apache-2.0

 Dataset card  Files and versions  Community  Settings

Dataset Viewer  [Auto-converted to Parquet](#)  API  Go to dataset viewer

Split

- ✓ train (15k rows)
- validation (3k rows)
- test (7.5k rows)

intent (string)	id (string)	text (string)
"book_flight"	"train-0"	"i need you to book me a flight from ft lauderdale to houston on southwest"
"schedule_maintenance"	"train-1"	"my check engine light is on and i need to take a look at it"
"reminder"	"train-2"	"is the company party on my list of reminders"
"are_you_a_bot"	"train-3"	"are you a human"
"todo_list_update"	"train-4"	"i need to do cleaning so add it to my to do list"
"oil_change_reminder"	"train-5"	"how do i change my oil and what kind do i need"

專案的一生



模型

- 分類整句：
- - 情感分析：「這部電影很棒！」(正面)
- - 垃圾郵件檢測：「您已經贏得一百萬美元！」(垃圾郵件)

```
[
  {
    "text": "i need you to book me a flight from ft lauderdale to houston on southwest",
    "intent": "book_flight",
    "id": "train-0"
  },
  {
    "text": "my check engine light is on and i need to take a look at it",
    "intent": "schedule_maintenance",
    "id": "train-1"
  },
  {
    "text": "is the company party on my list of reminders",
    "intent": "reminder",
    "id": "train-2"
  },
]
```

模型

- 分類整句：
- - 意圖偵測：「播放田馥甄的歌」（音樂播放）

模型

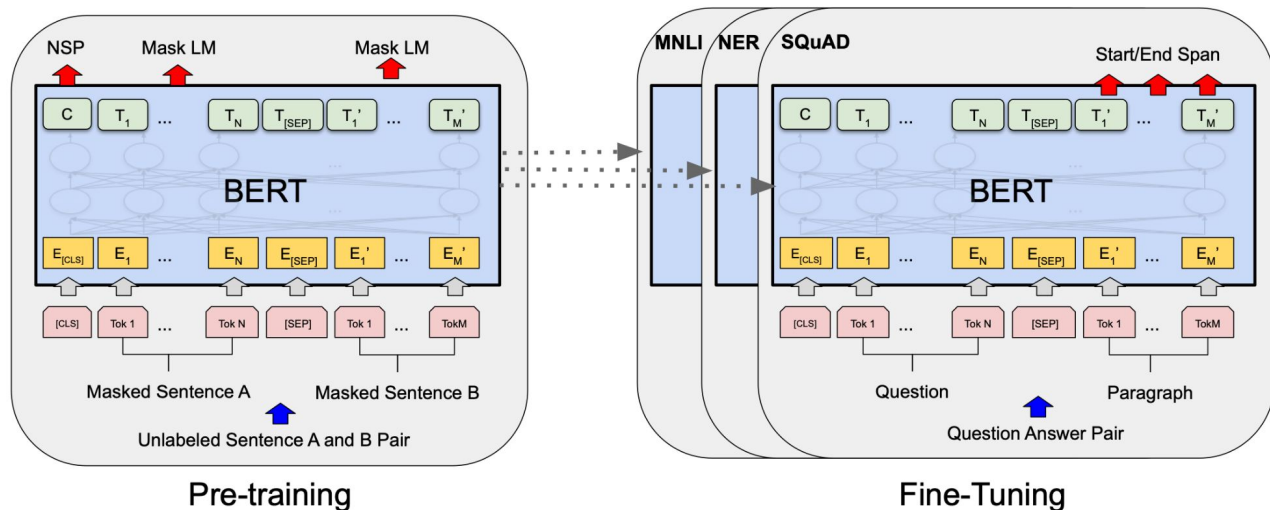


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

模型

- 萬事用 Huggingface



模型

Gradio

