「教師跨領域研究社群」-AI 技術應用交流餐會

時間：2022-11-18
地點：工學大樓 E819 室

報告內容：

主題：

ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision 介紹：

動機：

目前參數量最小的多模態 Transformer 方法。ViLT 使用預訓練的 ViT 來初始化交互的 transformer，這樣就可以直接利用 interaction layer(交互層)來處理視覺特徵，不需要額外增加一個視覺 encoder (如 Faster-RCNN)。

貢獻：

第一個基於 patch projection 的多模態預訓練模型，也是第一個使用 patch projection 來做 visual embedding 的方法。

證明了可以將 BERT 的方法和 Vison Transformer 結合起來用於多模態(multimodal) transformer。

模型架構(word patch alignment)、(cont.)：

在預訓練期時，模型共有三種有預訓練任務：

- Image Text Matching (ITM)：

找出 patch 與文字是否有對應，輸出是 True/False。

- Masked Language Modeling：把 15%的詞做 Masked, 預測 Masked 的詞。
- Word Patch Alignment：跟 ITM 一起, 先計算 patch 與文字的相似度，找出對應文字與 patch 配對。

下流任務介紹：

- Visual Question Answering
- Natural Language for Visual Reasoning.
- Image Text Retrieval

「教師跨領域研究社群」-AI 技術應用交流餐會

時間：2022-11-18
地點：工學大樓 E819 室

實驗結果：如簡報所呈現。

重點整理：

視覺和語言預訓練（VLP）提高了各種聯合視覺和語言下游任務的性能。在本文中提出了一個最小的 VLP 模型，視覺和語言轉換器 (ViLT)，在某種意義上說，使用無卷積方式簡化視覺輸入與文本輸入的處理。論文提出 ViLT 比以前的 VLP 模型快數幾倍與具有競爭性或更好的下游任務性能。

報告者講解及師生專注聆聽的畫面。

報告結束，老師上台進行點評及補充。

報告開始的介紹及全神貫注的師生。

會議開始前老師的開場白。