講者個人簡介

梁伯嵩博士於國立交通大學電子工程系畢業(電子 83)， 並於交通大學電子研究所獲得碩士與博士學位 (電研 85, 電博 91) ，並於臺灣大學管理學院 EMBA 商學組畢業。

目前任職於 聯發科技，擔任 前瞻技術平台 資深處長，負責 聯發科技前瞻研發中心 (MediaTek Advanced Research Center, 簡稱 MARC) 相關工作，並兼任 臺灣大學資訊工程學系 與 臺灣大學重點科技研究學院 合聘之客座教授，以及 陽明交通大學產學創新研究學院 智能系統研究所之教授級專業技術人員。曾榮獲「中華民國十大傑出青年」(科技發展類)、三度獲得經濟部智慧財產局 「國家發明創作獎」發明獎 (一金二銀)、經濟部技術處「產業科技發展獎 - 傑出青年創新獎」、中華民國資訊月 「傑出資訊人才獎」、ACM 台北/台灣分會與中華民國資訊學會「李國鼎青年研究獎」等榮譽。並為國內外 83 件發明專利的主要發明人。

題目

AI 運算架構 , 大規模神經模型 與 生成式 AI

AI Computing Architecture, Large-Scale Neural Network and Generative AI

摘要

隨著半導體科技的飛速成長，運算平台的基礎架構，已從 Bit 為主導的數位運算，已進入 Neurons 所啟發的人工智慧運算。AI 的神經網路架構，近幾年有很大的進展。 尤其在 Transformer 架構興起之後，啟動大型神經網路模型的增長。神經網路的參數數目從 Alexnet 的 6000 萬 (60M)，成長到 OpenAI GPT-3 的 1750 億 (175B)，而且已有數兆參數規模之研究。因為 LLM (Large Language Model) 的成熟，衍生許多令人驚豔的全新應用，像是文字產生高品質的影像、影片、音樂甚或程式設計，讓生成式 AI (Generative AI) 進入全新的境界。而最近對話式 AI - ChatGPT 的崛起，短短兩個月內吸引上億使用者，讓全世界都深刻的感受到 AI 的浪潮，即將顛覆人類的工作與生活模式。

但是，要讓大型神經網路的 AI 能力湧現 (Emergence) ，需要極大量的運算需求，短短七年間成長三十萬倍以上，遠超過半導體摩爾定律的成長速度，讓 AI 快速進入 Large-Scale Era。為了提供算力，百億億級 AI 超級電腦 (Exascale AI Supercomputer) 需求殷切，在此也會討論世界超級電腦的進展。此外，大型神經網路先透過 Pre-Train 訓練 AI Foundation Model，再 Fine Tune 進行下游任務訓練，兩階段訓練方式也漸漸成為主流。但是訓練 AI Foundation Model 的語料，十分重要。若其中隱含著偏見或毒化的資訊，將會讓這個瑕疵影響到所有的下游 AI 任務。因此 AI 模型的安全性，也成為新的資安議題。

AI 的運算需求，將成為 AI 進展與普及的關鍵，這是 IC 半導體產業的全新機會。台灣身處半導體產業鏈的關鍵角色，怎樣從 IC 的製造者，也同時成為 AI 應用的領先群，利用 AI 計算帶來更廣大的產業經濟與社會應用的價值，讓台灣的科技產業更上一層樓，是我們共同需要探索的課題。