機器人創新巔峰：
端到端大模型驅動實現人類級反應速度

在最新的機器人科技研究中，「端到端大模型驅動」已成為一個重要的發展里程碑。這種技術的實施意味著，機器人的反應速度正在獲得驚人的提升，已經接近於人類的速度。在這項進展的核心，是一種深度學習的模型，能夠不僅理解語言和圖像，還能夠自主學習並內化這些訊息來指導其行為。

當我們說一個機器人是「端到端」驅動的時候，我們指的是從感測器收集的數據到最終行為的輸出，都是由一個統一的 AI 模型來控制的。Figure 機器人就是這種創新的體現。它的機載攝像頭以 10hz 的頻率捕捉圖像，而其深度神經網絡則以驚人的 200hz 的速度輸出多達 24 個自由度的動作，這種速度的提升不僅顯著，而且標誌著機器人動作速度開始與人類相匹敵。

這些進步背後，是由 OpenAI 訓練的多模態模型（VLM），它結合了視覺和語言的理解能力。機器人通過其攝像頭和麥克風收集的數據，被輸入到 VLM，由它來處理整個對話的歷史記錄，生成語言回應，然後透過文本到語音技術，將這些回應傳達給人類。這意味著，機器人所執行的行為是基於自身學習和經驗，而非依賴遠程操作或人為的直接指揮。

這個創新的系統架構使得機器人能夠更加流暢地與人類互動，而不會有人工干預的滯後性。機器人現在可以自己決定在接到命令時應運行哪些學習過的行為模式，它的 GPU 加載特定的神經網絡權重來執行策略。這種方法不僅提高了效率，也提升了機器人的自主性和適應能力。

Figure 創始人的談話確認了這種速度的進步，這種進步不僅反映在機器人技術的進化上，更是對於未來機器人與人類世界互動方式的一個預示。OpenAI 的多模態模型所展示的多樣性能力，成為了機器人能與世界交互的關鍵，正如我們從各種演示視頻中所見的那些令人印象深刻的互動瞬間。隨著這些技術的持續發展，我們期待未來機器人能夠更加自然和高效地與我們的世界融合。

參考資料：

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., ... & Zitkovich, B. (2023). Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818.

https://zhuanlan.zhihu.com/p/668907606

- 機器人動作轉化為一種可以用文字描述的形式
- 轉換為文字標記並與網路規模的視覺語言資料集一起進行訓練
- 在模型進行預測或「推理」時，它會接收到一些文字標記作為輸入，並需要決定相對應的機器人動作。這裡的並透過「去標記」將文字標記轉化回具體的機器人動作指令的過程
- 實現閉環控制：系統的輸出會反饋到輸入端以進行調整和優化
- 利用視覺語言模型的骨幹和預訓練
- 泛化、語義理解和推理轉移到機器人控制中

基於多模態模型的端對端機器人控制模型