

自然語言處理

Multimodal Learning

Instructor: 馬豪尚

Multimodal Learning 多模態學習

- › Jeff Dean在2019年底NIPS大會上的一個採訪報道，講到了未來機器學習趨勢：多任務和多模態學習將成為突破口
- › 模態 (modal) 是事情經歷和發生的方式，我們生活在一個由多種模態 (Multimodal) 資訊構成的世界
 - 包括視覺、聽覺、文字、嗅覺資訊等等
 - 當研究的問題或資料集包含多種這樣的模態資訊時我們稱之為多模態問題
 - 研究多模態問題是推動人工智慧更好的了解和認知我們周圍世界的關鍵。



多模態

- › 多模態即是從多個模態表達或感知事物，可能有以下形式：
 - 描述同一物件的多媒體資料
 - › 如網路環境下描述某一特定物件的影片、圖片、語音、文字等資訊
 - 來自不同感測器的同一類媒體資料
 - › 如醫學影像學中不同的檢查設備所產生的影像數據，包括B超(B-Scan ultrasonography)、電腦斷層掃描(CT)、核磁共振等
 - 具有不同的資料結構特徵、表示形式的表意符號與資訊
 - › 如描述同一物件的結構化、非結構化的資料單元
 - › 描述同一數學概念的公式、邏輯符號、函數圖及解釋性文字
 - › 描述同一語意的詞向量、詞袋、知識圖譜以及其它語意符號單元等



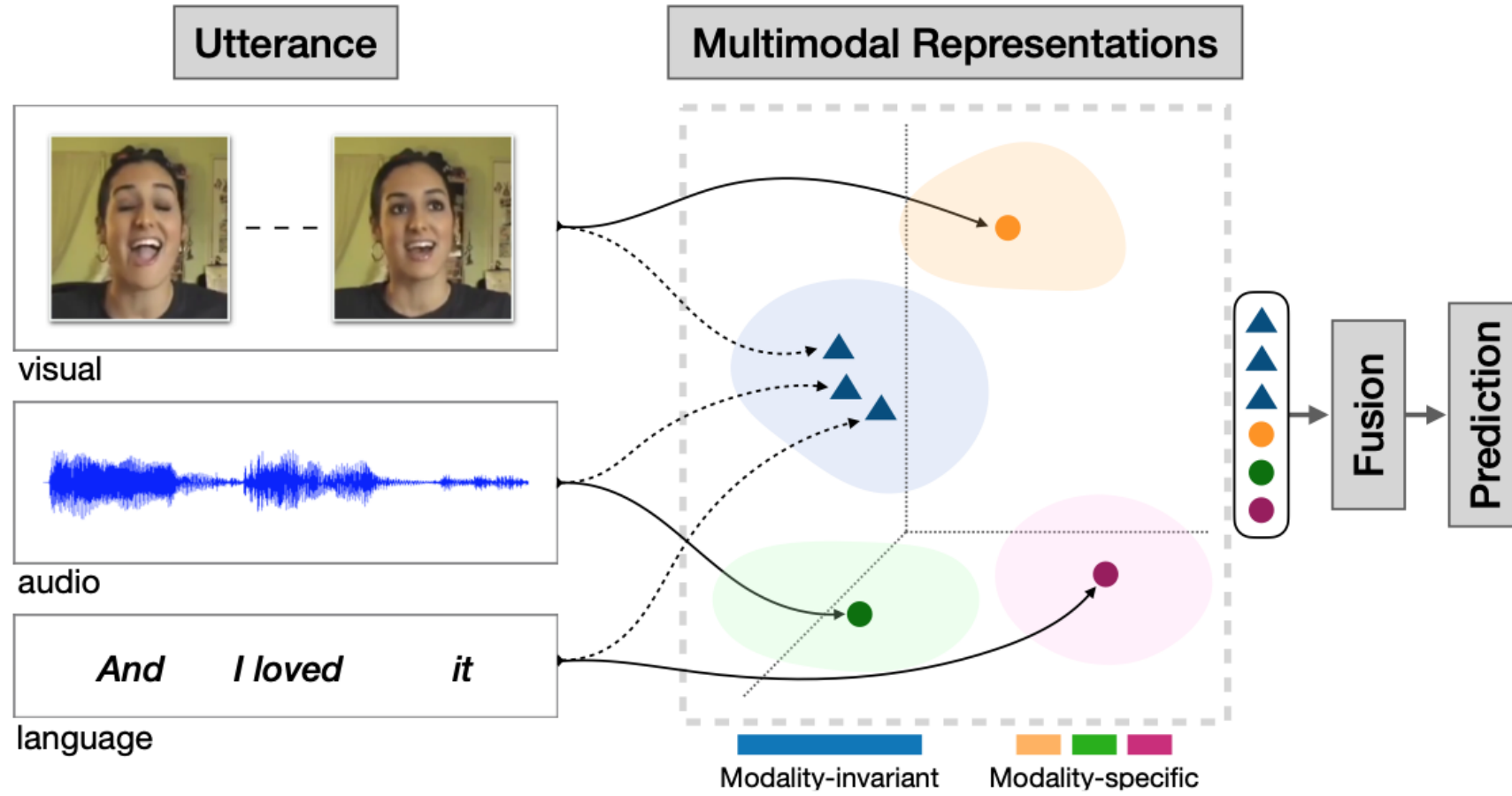
It snowed in the evening. Flakes of snow were drifting down. If you walked in the snow, you can hear a creaking sound.



Multimodal Learning 多模態學習

- › 多模態學習是從多種模態的資料中學習並且提升自身的演算法，它不是某一個具體的演算法，它是一類演算法的總稱
- › 從語意感知的角度切入，多模態資料涉及不同的感知來源如視覺、聽覺、觸覺、嗅覺所接收到的訊息
- › 在資料層面理解，多模態資料則可被看作多種資料類型的組合
 - 圖片、數值、文字、符號、音訊、時間序列
 - 集合、樹、圖等不同資料結構所組成的複合資料形式
 - 來自不同資料庫、不同知識庫的各種資訊來源的組合

Multimodal Learning 多模態學習

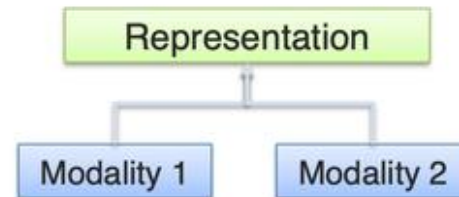


Multimodal Learning Challenge

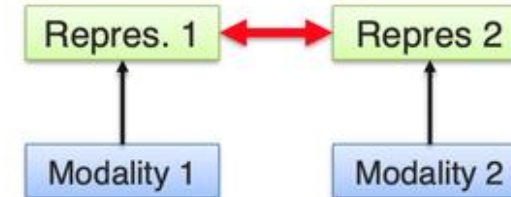
- › 第一個挑戰是學習如何總結多模態資料
 - 單模態的表徵負責將資訊表示為電腦可以處理的數值向量或進一步抽象化為更高層的特徵向量
 - 多模態表徵是指透過利用多模態之間的互補性，剔除模態間的冗餘餘性，從而學習到更好的特徵表示

Definition: Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.

① Joint representations:

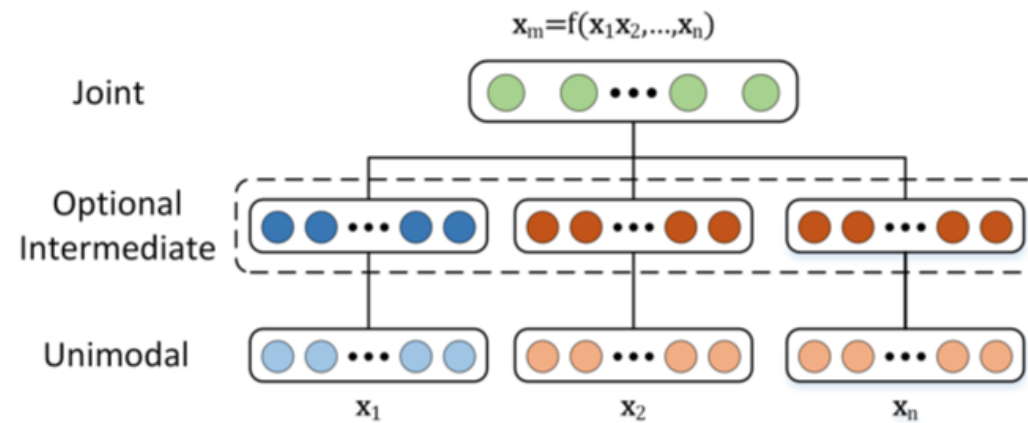


② Coordinated representations:



Joint Representation 聯合特徵學習













- › 聯合表徵 (Joint Representation) 將多個模態的資訊一起映射到一個統一的多模態向量空間
- › Joint結構著重捕捉多模態的互補性，融合多個輸入模態 x_1, x_2 獲得多模態特徵 $x_m = f(x_1, \dots, x_n)$ ，進而使 x_m 完成某種預測任務



(a) Joint representation

Joint Representation 聯合特徵學習

- 在應用階段輸入圖片，利用條件概率 $P(\text{文本}|\text{圖片})$ ，生成文本特徵，可以得到圖片相應的文本描述
- 而輸入文本，利用條件概率 $P(\text{圖片}|\text{文字})$ ，可以產生圖片特徵，透過檢索最靠近該特徵向量的兩個圖片實例，可以得到符合文字描述的圖片

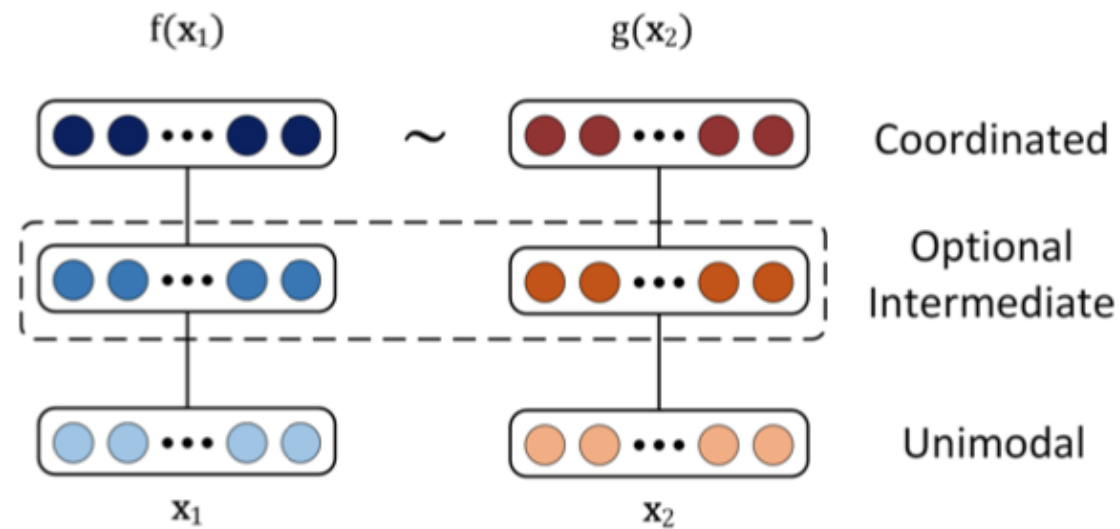
Image	Given Tags	Generated Tags	Input Tags	Nearest neighbors to generated image features	
	pentax, k10d, kangarooisland, southaustralia, sa, 300mm, australia, australiansealion	beach, sea, surf, strand, shore, wave, seascape, sand, ocean, waves	nature, hill, scenery, green, clouds		
	< no text >	night, lights, christmas, nightshot, nacht, nuit, notte, longexposure, noche, nocturna	flower, nature, green, flowers, petal, petals, bud		
	aheram, 0505, sarahc, moo	portrait, bw, balckandwhite, people, faces, girl, blackwhite, person, man	blue, red, art, artwork, painted, paint, artistic, surreal, gallery, bleu		
	unseulpixel, naturey crap	fall, autumn, trees, leaves, foliage, forest, woods, branches, path	bw, blackandwhite, noiret blanc, bianconero, blancoynegro		

Coordinated Representation 協同特徵學習

- › 協同特徵學習 (Coordinated Representation) 將多模態中的每個模態分別映射到各自的表示空間，但映射後的向量之間滿足一定的相關性限制 (例如線性相關)
- › Coordinated結構並非尋求融合而是建模多種模態資料間的相關性，它將多個(通常是兩個)模態映射到共同空間
 - 表示為： $f(x_1) \sim g(x_2)$
 - 其中 \sim 表示某一種不同模態之間的關係

Coordinated Representation 協同特徵學習

- 類神經網路的最佳化目標是這種關係(通常是相似性，即最小化cosine距離等量測)



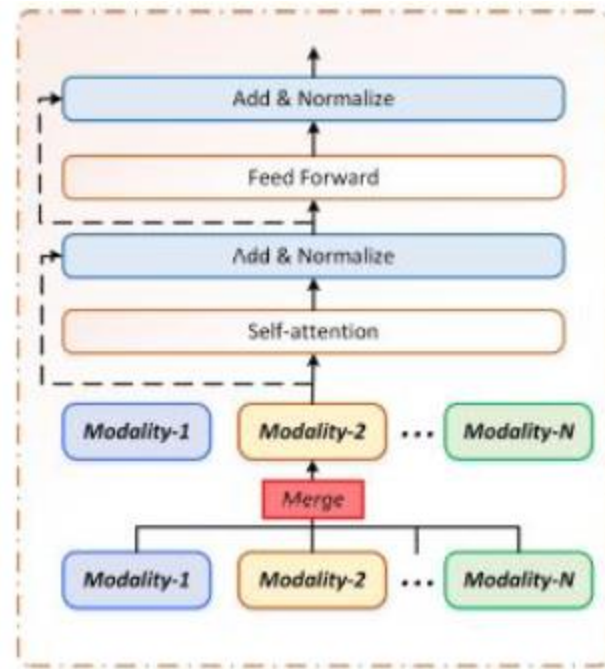
(b) Coordinated representations

Multimodal Learning Challenge

- › 第二個挑戰涉及如何將資料從一種模態轉換到另一種模態，不僅資料是異質的，模態之間的關係通常是開放式的或主觀的
 - 圖像描述任務中，存在多種描述圖像的正確方法，並且可能不存在完美的翻譯
 - 語言翻譯任務中，多重答案是正確的，決定哪個翻譯比較好往往是主觀的
- › 評估標準
 - 人工評價是最理想的評估，但是耗時耗錢，需要多樣化打分人群的背景以避免偏見
 - 自動化指標是視覺描述領域常用的替代方法，包括BLEU，Meteor，CIDEr，ROUGE等，但它們被證實與人的評價相關性較弱
 - 基於檢索的評估和弱化任務(例如：將影像描述中一對多映射簡化為VQA中一對一的映射)也是解決評估困境的手段

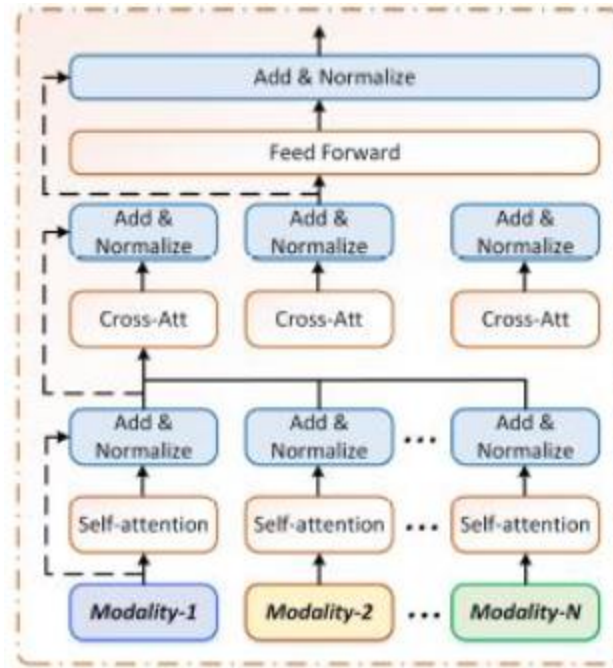
多模態學習模型的典型結構

- › Merge-attention：多個輸入模態調整為同一的特徵表示，多個模態的特徵在自注意力之前被合併，共同進入Transformer



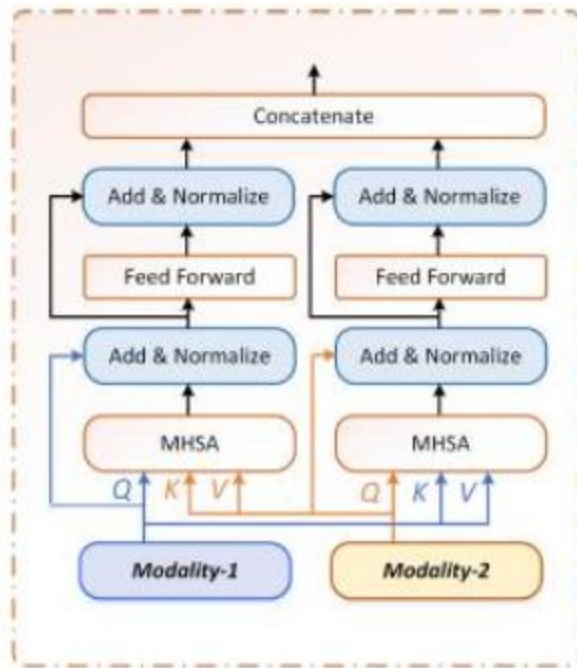
多模態學習模型的典型結構

- › Co-attention：每個輸入模態都具備私有自注意力通道，用於模態獨立特徵的導入，然後再使用共同的交叉注意力層融合多模態特徵



多模態學習模型的典型結構

- › Cross-attention：對於多模態任務，將圖像與語言分別結合，實現圖文資訊的相互嵌入與問答



多模態應用 ImageBind

- › META 開源型多模態學習的革新之作，將 6 種感知緊密結合
- › 使用圖片生成聲音 (Using an image to retrieve audio)
 - 想像一下，你有一張火車的圖片，但你想知道火車的聲音是什麼樣子的。有了 ImageBind，你只需要給它看一下圖片，它就能想起火車的聲音，然後進行生成。
- › 使用文字搜尋圖片和聲音 (Using text to retrieve images and audio)
 - 你知道，有時候我們想要找到一些圖片或聲音，但我們不知道該怎麼形容它們。只要把你想要的東西用文字描述一下，比如「小狗叫聲」，它就會幫你找到一些可愛的小狗叫聲和相關的圖片。這樣，你就能更快地找到你想要的資料了。
- › 使用聲音生成圖片 (Using audio to generate an image)
 - 最後，我們來看一個非常酷的功能：使用聲音生成圖片！想像一下，你聽到了一個很有趣的聲音，比如鸚鵡的叫聲，但你不知道鸚鵡長什麼樣子。這時，你可以讓 ImageBind 幫你畫一張鸚鵡的圖片。只要讓它聽聽鸚鵡的叫聲，它就能想象出鸚鵡的模樣，然後把它畫出來。

多模態應用 GPT-4

- › GPT-4模型是基於GPT-3.5建構的，增加了視覺語言模型的模組（在圖形Transformer階段完成的視覺預訓練模型）
- › 為了預訓練模型在多模態領域進行初步微調，首先會在文本資料集和多模態資料集中抽取問題，由人類去標註給出高品質答案，然後用這些人工標註好的資料來微調GPT-4初始模型