**No.12 Scene description(場景說明)**

Leader :黃宗德 406410091

Member: 陳威樺 406410061

Member: 王至弘 406410009

Member: 林于恩 406410050

Member: 吳啟民 609211014

Member: 鄭策丰 609211003

**1. 專案目標**

過去的程式大多只有對圖中的物件進行偵測，並簡單的舉出圖中所含的物件名稱，我們所期望的是基於物件偵測下，能夠對於一張圖甚至是一段影片中的內容，不僅僅只辨識出圖中的物件，更希望能夠言簡意賅的來描述圖中的內容。英語有句俗話叫"A picture is worth a thousand words"，其說明圖片相較於文字往往蘊含大量而豐富的資訊，該如何用簡短的文字來描述圖片、甚至影片中的故事，是項值得思考並且好好嘗試的問題。

我們期望這項功能能夠輔助人們理解圖片或影片中的資訊，察覺那些可能因為思考盲點上所無法覺察的事物，使得人們可以更加透徹及全面的理解整個圖片與影片中的脈絡，甚至對於某些藝術創作者，期望能夠提供他們在創作上些許的指引，成為他們因創作瓶頸而裹足不前的一盞苦海明燈，但藝術領域變化多端且難以捉摩，因此這是個充滿挑戰性的計畫，我們願意朝這方面做嘗試，即便最後結果不盡理想。

場景辨識，於輸入一張照片後，可以根據照片內容生成描述該照片的語句。(以下幫說明)，根據輸入的圖片，辨識出圖片中的物件，再以一段話描述其中不同物件之間的關係。

普通的物件辨識只能辨識出圖中的物件，例如:



希望能做到的目標: 除了辨識出物件外，還能以一句話描寫物件間的關係

**2. 專案實作方法，包含資料集來源、模型來源 (自己的，還是已有的)，資料集說明以及下載位置(link)、模型說明包含架構、參數等。**

**(1). 模型來源:**

**a.** <https://github.com/facebookresearch/detr>

**b.** <https://github.com/AaronCCWong/Show-Attend-and-Tell>

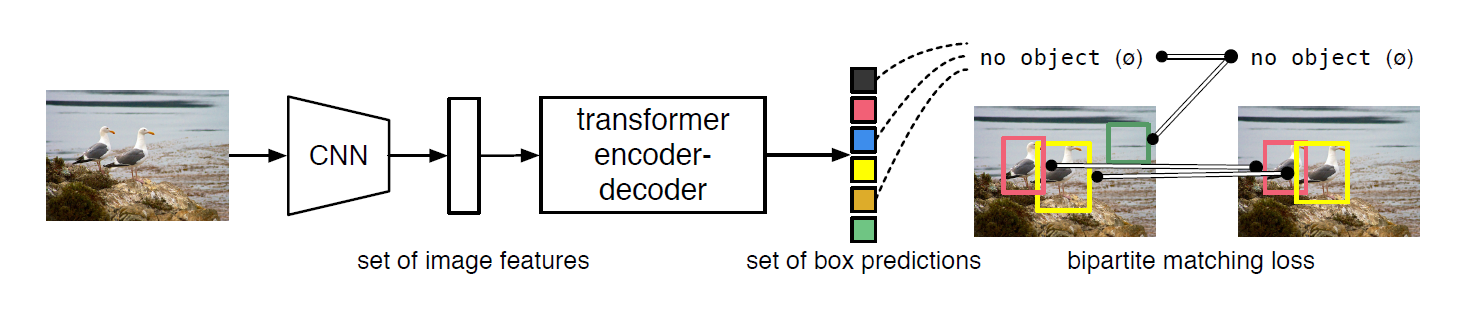
**(2). 資料及來源:**

**a.** <https://cocodataset.org/#download> (2014 dataset)

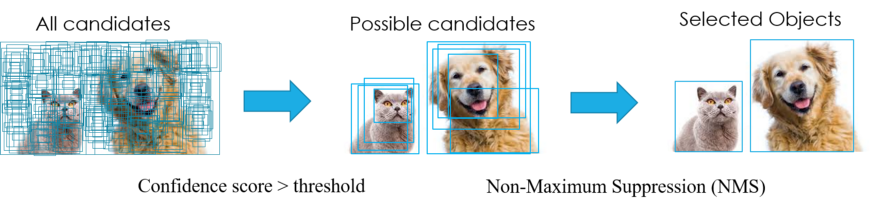
**b.** <https://tribhuvanesh.github.io/vpa/> (29種分類數據集)

**(3). 模型說明:**

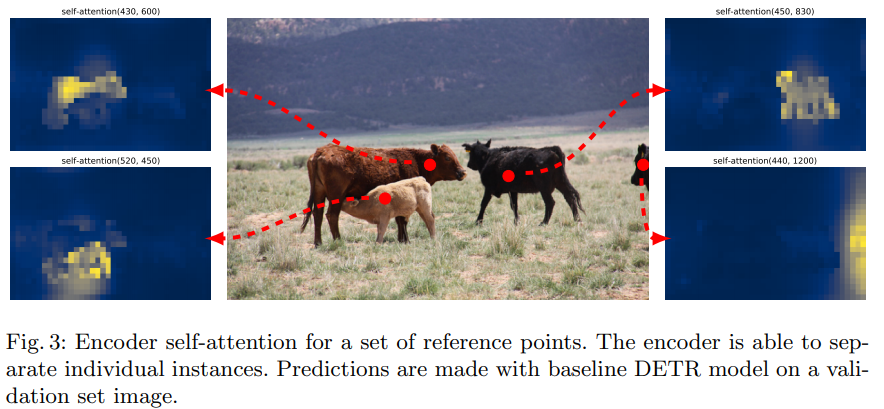
基於BERT 架構的DETR模型，於基礎架構上，不同於其他模型，輸入圖形會先經過傳統CNN找出其特徵，再將其特徵輸入至auto-encoder的網路中，以此降低訓練參數數量，降低訓練難度。



於此模型中，並不像其他的圖形辨識模型，需要先做dense object detection來確定需要被描述的物體範圍，所以在DETR模型中就不需要考慮到原始輸入的解析度以及recepive field。但是實際上DETR是將object detection 視為 set detection，set包含預測的物體種類以及該bounding box，且在DETR模型中對於重複的bounding box 的獎勵並不高(讓bounding box的相關性盡可能不重複)(如下圖)



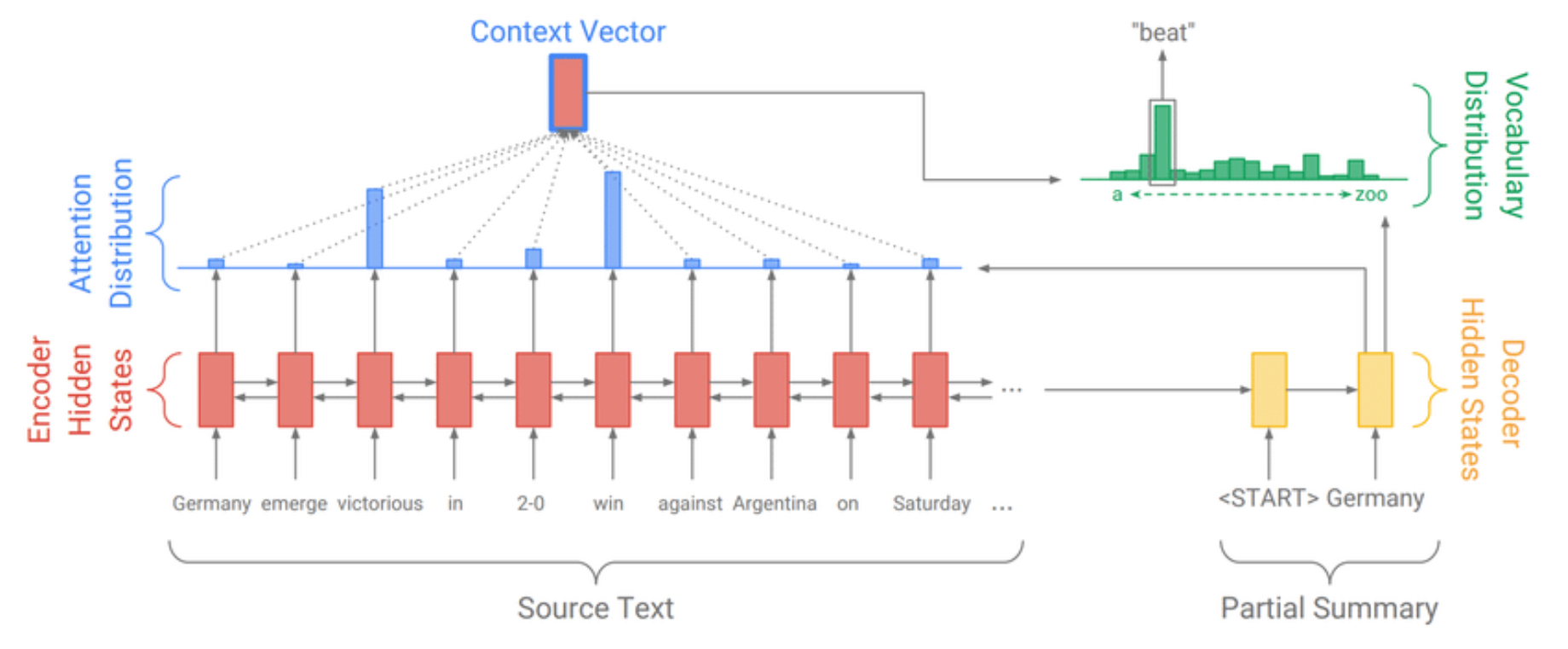
因此在encoder最後一層的attention layer做可視化處理，可以發現encoder已經可以分辨出不同的物體，且盡可能不讓bounding box 不重疊。(如下圖)



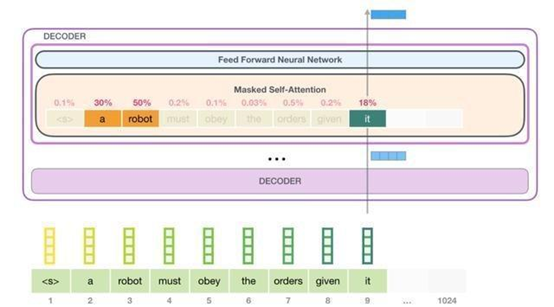
而在轉換成文字的部分，與許多傳統模型(比較舊)不一樣，傳統模型為了處理句子中間不同單字的關聯性(EX: **A robot must obey the orders given it by human beings except where such orders would conflict with the First Law**

上文若無全文，無法得知此三個代名詞(紅字)確切意思為何?)

因此發明出Attention 的處理方式，即每一個不同的set中decoder的部分會先透過attention 計算一個context vector，而這個context vector來源自encoder再輸入單字後的context vector 的組合。(如下圖)

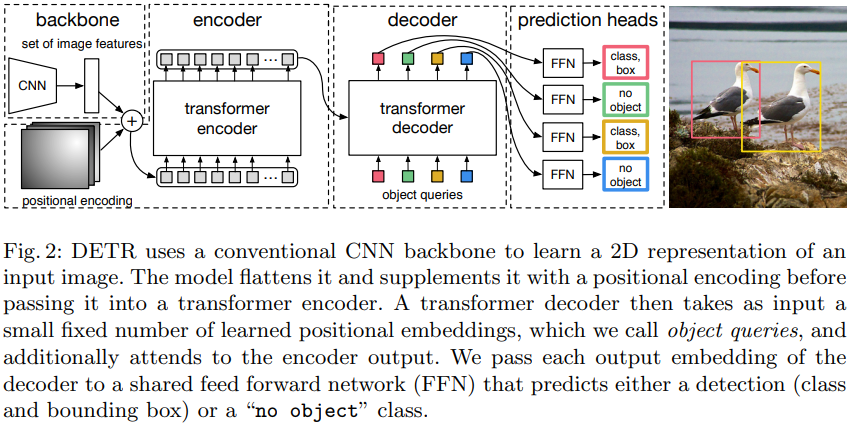


Attention雖然讓context vector變得更合理，但事實上這個context vector在字詞與字詞間的聯繫依然是比較薄弱的，也就是在attention前，產生context vector的過程。因此DETR在transformer 中使用self-attention，而Self-attention的加入，讓模型可以直接考慮輸入與輸出文字自己本身的字詞間關聯，不容易因為語句或文本過長導致效果變差。(如下圖)



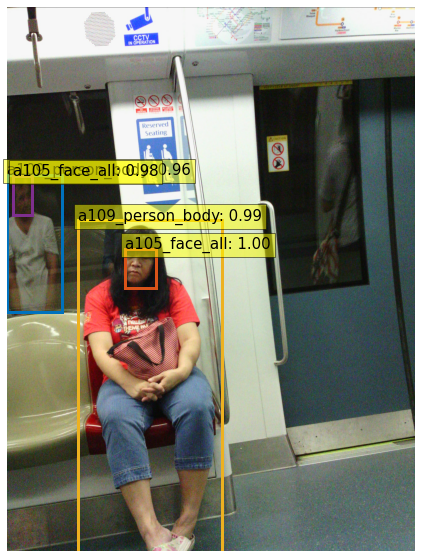
因此此模型大致執行方式:

使用CNN找出feature map -> 找出不重複set detection-> 放進transformer -> 得到多個final set -> 輸出評價最好的set



Fine tune方式:

因為本DETR只能找出信心值最佳的bounding box，因此目標是能夠標示出帶有物體訊息的bounding box。(見下圖結果)

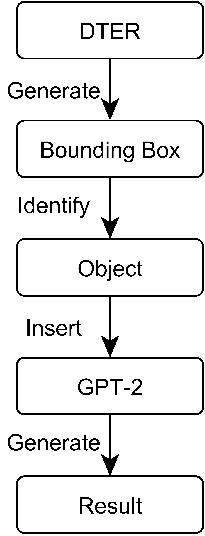


**3. 實驗設計，包含每一個實驗的說明 (此部分若還沒有開始，請寫規劃)。**

依據NLP的測量準確率方式，將會使用BLEU進行模型正確率評估

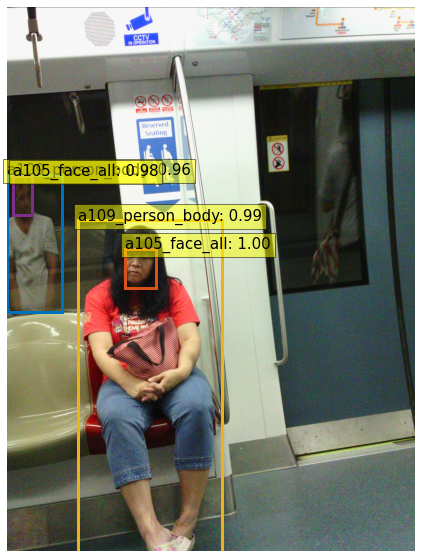
，因為時間緣故，最終驗證將會使用dataset中的圖片進行實驗，加上自行於網路上抓的圖片進行demo。

主要關注attention的重點有無錯誤，是否有重要字詞未出現

**4. 實驗結果，包含使用的量化指標 (evaluation metrics) (若還沒開始，請寫規劃)。**

(1). 專案進行結果:

a. 依據期初計畫，將會使用DETR生成bounding box，接著利用fine tune方式把bounding box之中的物體辨識出來，再利用GPT-2模型進行文本生成。(如右圖)

b. 以本次進度，可以成功把DETR中生成的bounding box加上物體的名稱，讓辨識出來的bounding box能夠有名稱進行註解(如右圖)

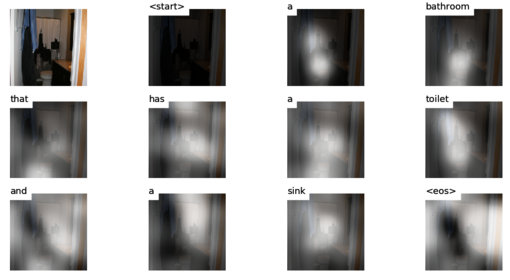
c. 但是再將結果輸入進GPT-2模型以生成字句時遇到問題。因GPT-2模型當時設計時是以補完字句為主要目的設計，因此不支援輸入斷續名詞讓模型補完字句。因此若以右圖為例子，生成最多為: a person, sitting, chair。但是我並無法只用這三個做為輸入讓GPT-2模型幫我生成 a person is sitting on the chair。因此進度卡在無法將生成bounding box 內容與生成字句方式連接起來。

d. 可能解決辦法有兩種: 一種是找到有更多分類的資料集，讓物體內容能夠更細分，EX: a person加入更多動作或者形容詞，a person is sitting, a person is running … ，讓生成的描述更加接近一般語句，讓後續放入GPT-2模型生成的結果更佳，但是找不到相關資料集。

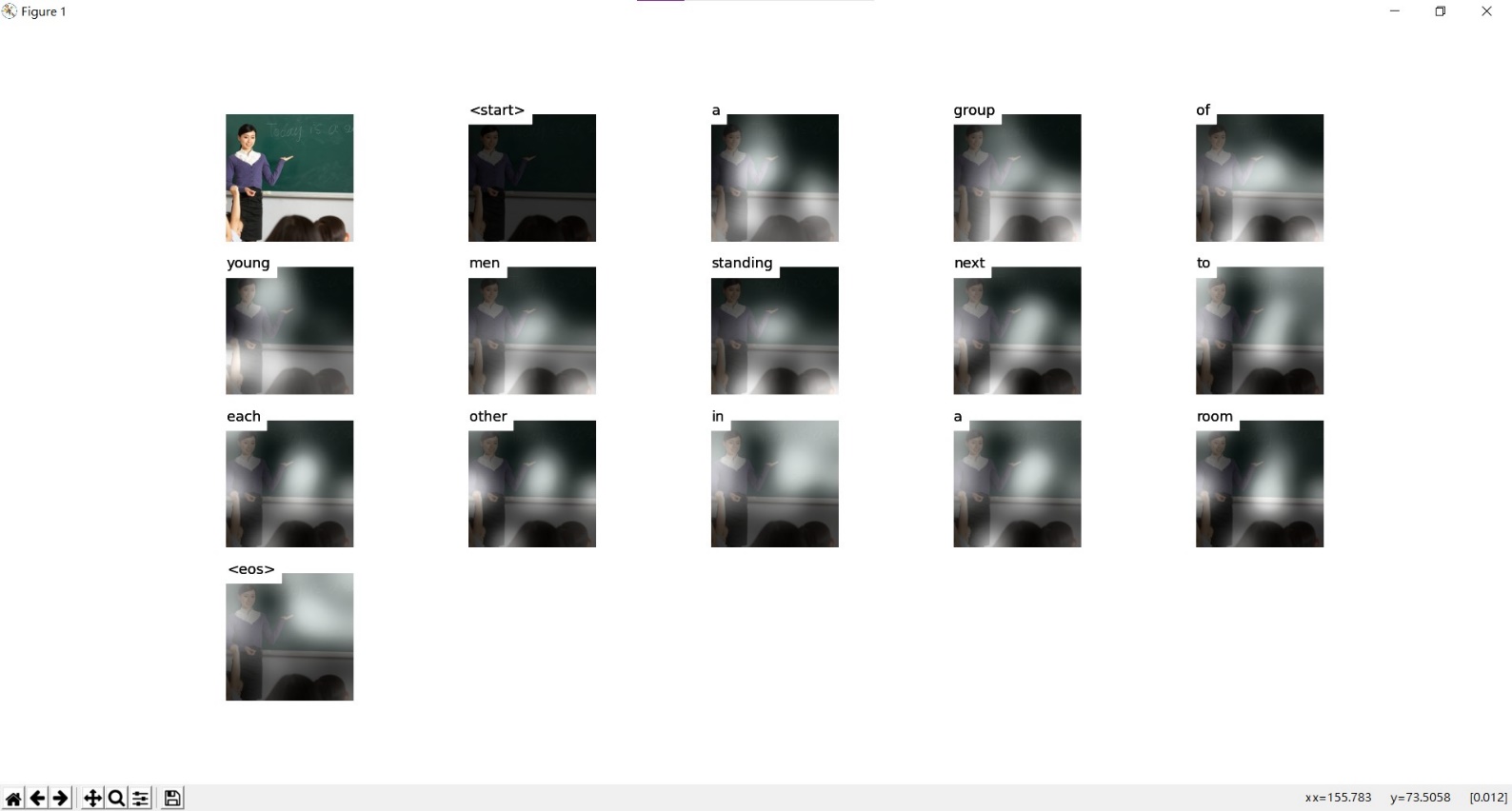
第二種方式為直接更改GPT-2模型，讓GPT-2能夠輸入特定斷續單字而能夠補完完整句子，但是這種做法是對GPT-2模型進行大改，因此只能從頭訓練，因為時間成本問題，這種做法也不可行。

e. 因此最終實現方式是使用其他大型計劃來實現場景辨識此專案，經過搜尋找找到Show-Attend-and-Tell，此計畫基本上就是場景辨識的完整版方案。因此直接使用Show-Attend-and-Tell的程式碼，使用coco 2014的dataset進行訓練，encoder使用VGG19，進行模型生成，而在這訓練過程中使用到的單字一共為10365個，數量偏少，因此在圖片內容物超出這些單字時精準度會比較不佳

(2). 實驗結果:

如果demo圖片是在訓練集中，且單字並沒有超出範圍時，生成字句結果不錯

但是當我自行抓一張圖片進行demo時，結果不慎理想(自行demo圖片如下圖)

以此張圖片為例，期望生成出A teacher is teaching some child in the classroom這種描述，至少生成出字句要包含teacher, teaching，描述方式隨意，但是demo結果非常不佳(見下圖)

(A group of young men standing next to each other in a room)

因此我懷疑有over fitting的傾向，但是有一些dataset 的圖片生成結果也不佳，因此這個模型出來的結果並不是有參考性的，因此沒有做成果評估(BLEU)