Image Captioning Based on Fine-grained Relationships with Multiscale Regions of Interest

Liang-Yu Lin[1], Chow-Sing Lin[2,*]

Department of Computer Science & Information Engineering
National University of Tainan, Tainan, 70005, Taiwan

[1]E-mail : m10759006@stumail.nutn.edu.tw
[2]Email : mikelin@mail.nutn.edu.tw

Abstract

With the rapid development of machine learning, the technique of Image Captioning is becoming more and more advanced. Recent researches of Image Captioning introduce Region Proposal Networks (RPN) and Attention Mechanism. Through RPN, we can extract features of specific object region in the image and reduce the probability of noises being treated as visual features. Attention mechanism makes the models to focus more on the mapping of object and caption. However, the current research results have deficiencies. Both RPN and Attention Mechanism only focus on the single object region instead of fine-grained visual features. Aforementioned deficiencies cause mistakes that caption generator generates uncertain relationships. In this paper, to improve exquisiteness of relationship descriptions for Image Captioning, we propose the Image Captioning model which generates sentence with multi-scale regions of interest (ROIs) between two different objects. Our proposed architecture includes Region Proposal Networks, Fully Convolutional Neural Networks and Long Short-term Memory cells. Compared to the existing research results, we extract not only object regions but multi-scale ROIs between two different objects on visual features. Some of Multi-scale ROIs are noises that can be screened by utilizing Intersection-over-Union (IoU). Each ROI utilizes FCNN to extract the visual features, followed by obtaining sorted fusion features with fusion mechanism and sorting network, and lastly learning transformation between this features to a whole sentence by LSTM. Caption generator can focus on learning how to generate fine-grained attributes with hierarchical attribute supervisions on the training stage. The architecture proposed in this study can use more precise verbs to describe object actions on dynamic pictures. Furthermore, our architecture outperforms on metrics based n-gram.

Keywords: Image Captioning, Region Proposal Networks, Multi-scale ROIs, Long Short-term Memory cells

---

* Corresponding author: mikelin@mail.nutn.edu.tw
DOI：10.53106/222344892023101302003

利用多尺度感興趣區域之細微關係提供圖片字幕

林亮宇，林朝興*

國立臺南大學 資訊工程系

摘要

隨著機器學習的蓬勃發展，圖片字幕生成(Image Captioning)的技術愈來愈進步。近期的 Image Captioning 引入區域提取網路 (Region proposal Networks，RPN)與注意力機制(Attention Mechanism)。 Image Captioning 透過 RPN 提取圖片中特定的物件區域，可以降低雜訊被當作視覺特徵的機率；注意力機制讓模型更專注在物件到文字的轉換。但是目前研究成果還存在著缺陷，RPN 與注意力機制皆專注於單一物件區域。它們缺少物件與物件之間更細膩的視覺特徵。上述的缺陷導致字幕生成器生成不明確的關係描述。為了提高 Image Captioning 生成關係描述的細膩度，本研究提出透過不同物件之間多尺度感興趣區域之關係特徵的 Image Captioning 模型。本研究架構有 RPN、全卷積神經網路 (Fully Convolutional Neural Networks，FCNN)以及長短期記憶 (Long Short-term Memory，LSTM)單元。相較於現有的研究成果， 在視覺特徵上，除了物件區域外，我們將進一步提取不同物件之間的多尺度 ROIs。由於某些多尺度 ROIs 是屬於雜訊，因此利用並交比(Intersection-over-Union)進行篩選。每一個 ROI 都先經由 FCNN 萃取出視覺特徵，再通過融合機制與排序網路獲得已排序的融合特徵，最後利用 LSTM 學習此特徵到完整句子的轉換。在訓練過程中額外透過階層式屬性的輔助監督，使字幕生成器能夠針對如何生成細膩的屬性進行學習。本研究提出的架構能夠在動態的圖片上，使用更精確的動詞描述物件動作。並且在基於 n-gram 的方法上，獲得更高的分數。

關鍵詞：圖片字幕生成，區域提取網路，多尺度感興趣區域，長短期記憶單元

1. 緒論

隨著時代進步加上科技日新月異，許多生活化的應用需要機器分析影像[1]–[3]。透過生成趨近於人類視覺概念的自然語言描述，以提供完備的輔助系統，例如：將視覺影像注入聊天機器人(ChatBot)，並幫助視障人士感知周圍的視覺環境；當人在家中找不到鑰匙，ChatBot 可以透過影像分析提供鑰匙周遭的環境描述，使人類迅速解決此問題。為了提供上述的輔助功能，圖片字幕生成(Image Captioning) 已成為相當熱門的研究議題。

圖片字幕生成是屬於跨領域研究議題，它結合了電腦視覺(Computer vision)與自然語言處理(Natural Language Processing，NLP)。電腦視覺主要讓電腦分析影像，並學習透過人類的視覺概念理解其影像內容。由於電腦視覺必須同時處理複雜結構的運算，因此需要大量的計算資源；自然語言處理是將自然語言中複雜結構轉換到電腦能夠理解的數據，讓電腦可以在數據與自然語言之間互相轉換。由於自然語言是屬於擁有時間結構的語言，因此自然語言處理需要透過含有時間序列結構的相關技術提取特徵。但是，傳統硬體的計算資源不足且缺乏有效的技術，導致在電腦視覺與與自然語言處理上出現瓶頸。

至今，機器學習(Machine Learning，ML) 的蓬勃發展，迎來了許多嶄新的技術，例如：卷積神經網路(Convolutional Neural Networks，CNN) 與遞歸神經網路(Recurrent Neural Networks，RNN)。卷積神經網路一般包含卷積層(Convolutional layer) 與池化層(Pooling layer)。卷積層提取影像中的高階視覺特徵，池化層則是減少特徵樣本數，可以降低模型學習的複雜度；遞歸神經網路用來生成含有時間序列結構的完整句子。結合上述技術，即可建出圖片字幕生成架構。

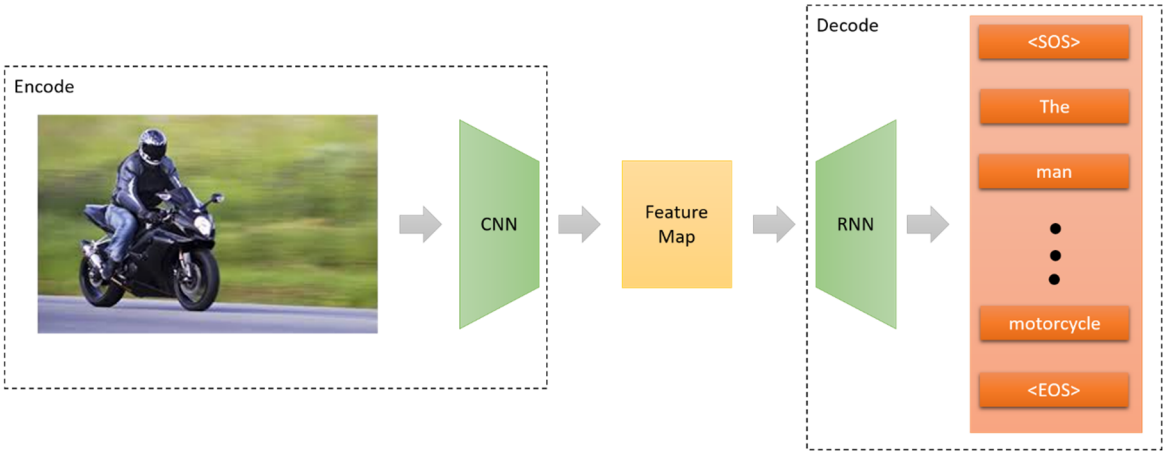


圖 1 Encoder-decoder 圖片字幕生成架構圖

圖片字幕生成陸續出現[4]–[12]，其常見架構幾乎都是基於編碼器(Encoder)-解碼器(Decoder)，如圖 1。編碼器負責提取靜態影像的特徵圖(Feature Map)，通常使用卷積神經網路。解碼器則負責將特徵圖以含有時間序列的結構生成完整圖片字幕，通常使用遞歸神經網路。在訓練階段上，模型透過預先定義的損失函數(loss function)更新參數，並學習輸入靜態影像到輸出完整圖片字幕的映射。除了單一風格的圖片字幕生成，另外也有針對學習生成不同風格的自然語言描述[9]–[12]，例如：MSCap[9]利用生成對抗網路(Generative Adversarial Networks，GAN)生成多種風格的自然語言描述，像是“Humorous”、“Romantic”、“Positive”或“Negative”。但是，上述研究主要學習整張圖片到句子的映射，對於圖片字幕中不曾被描述的物件區域也會被當作視覺特徵，導致視覺特徵含有雜訊的機率極高[13]。

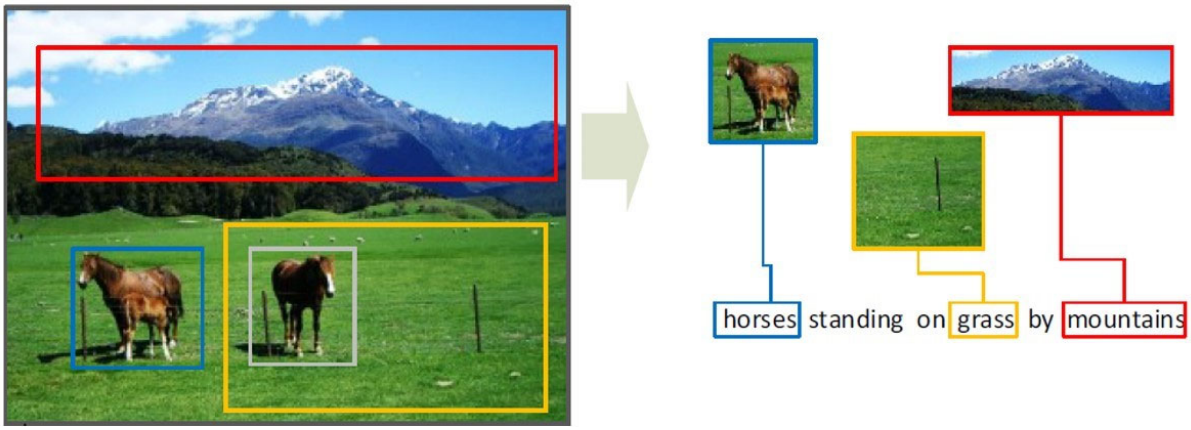


圖 2 圖片中的物件區域提取範例

為了克服上述問題，學者於圖片字幕生成中引入特定視覺區域提取之技術[13]–[16]。如圖 2[13]，首先透過物件提取網路(Region Proposal Network，RPN)[17]對圖片進行物件偵測，提取圖片中的個別物件區域。接著，將這些個別物件區域進一步萃取視覺特徵，讓模型針對如何從特定的視覺特徵映射到特定的自然語言描述進行訓練。但是，目前的圖片字幕生成所使用的視覺特徵不夠完整，導致關係描述不明確，甚至生成錯誤的動作描述。如圖 3[15]，可以發現生成的完整句子，在“woman”與“table”之間的關係描述生成“sitting”，但“sitting”並不符合圖片中物件的動作。我們發現主要的問題在於先前圖片字幕生成模型[15]忽略不同物件之間更加細膩的視覺特徵，像是不同物件之間的多尺度(Multi-scale)區域。由於上述問題，導致字幕生成器容易受到大量訓練資料的影響。如圖 4，模型生成“man”與“motorcycle”的關係描述上，由於在訓練階段中學習生成“ride”遠大於“stand”動作描述的次數，造成高機率生成“ride”的動作描述。

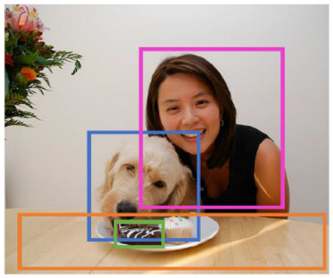


圖 3 圖片字幕生成實驗範例

為了解決上述缺少忽略不同物件之間更細膩的視覺特徵的問題，我們認為除了透過物件提取網路提取個別物件區域，必須加入物件聯集區域以及多尺度區域。如圖 5(a)所示，透過物件提取網路進行物件偵測，我們可以提取出圖片中的個別物件區域，並且用邊界框(Bounding box) 標示出個別物件的位置與邊界。接著將兩個邊界框的邊界進行聯集，即可得到兩物件的聯集區域。以男孩與飛盤為例，圖 5(b)即為其中一個聯集區域。最後，多尺度區域則為兩物件聯集區域之間各種不同大小的區域。圖 5(c)中藍色的邊界框即為男孩與飛盤之間的多尺度區域。本研究提出一個方法來有效地提取聯集區域中所有多尺度區域。並且為了避免包含過多雜訊的多尺度區域也被用來提取特徵，我們透過計算並交比(Intersection over Union，IOU) 並且設定閥值來篩選多尺度區域。

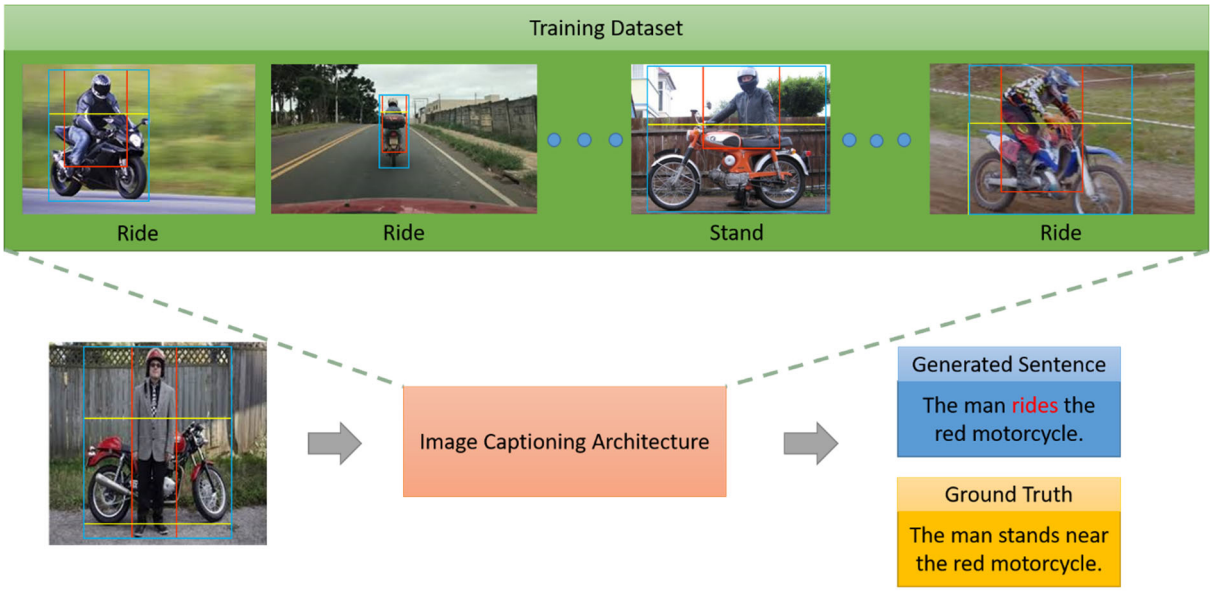


圖 4 生成的關係描述不符合圖片之範例

(a)個別物件區域 **(b)物件聯集區域 (男孩與飛盤)** **(c)多尺度區域 (男孩與飛盤)**

圖 5 物件區域範例

本研究提出透過多尺度感興趣區域之細微關係提供圖片字幕的深度學習神經網路架構，包含區域提取網路、全卷積神經網路(Fully Convolutional Neural Networks，FCNN)以及長短期記憶(Long Short-term Memory，LSTM)單元[18]。透過區域提取網路提取圖片中各式各樣的物件區域讓模型可以學習不同物件的視覺特徵到物件描述的映射。可以減少雜訊被當作視覺特徵的機率，例如：圖片中物件分別為 a、b 與 c 物件，當對應此圖片的特定描述只包含 a 與 b 物件時，可以透過區域提取網路提取 a 與 b 物件的視覺區域讓模型學習特定物件到特定描述的映射。除了物件區域，同時也在不同物件之間的聯集區域中提取多尺度區域。可以用來表示更加細膩的關係特徵並加強學習圖片到自然語言描述的映射。由於某些多尺度區域是屬於多餘或雜訊的區域，因此使用並交比篩選出重要的區域。提取完的所有物件區域先透過全卷積神經網路中的卷積層萃取特定區域的視覺特徵，再經由融合機制將所有視覺特徵進行整合，以及利用排序機制(Sorting Mechanism)重整視覺特徵的順序，最後以含有時間序列結構的長短期記憶單元學習此所有排序完成特徵到完整句子的映射。另外，為了有效地使模型產生的圖片字幕更貼切，我們在訓練階段中額外使用粗略到細膩描述的階層式屬性來描述詞語屬性。

本研究的主要貢獻如下列幾點：

1. 額外提取在不同物件之間聯集區域中的多尺度 ROIs
2. 篩選並且融合多尺度 ROIs 特徵
3. 字幕生成器在生成動態圖片字幕時使用更貼切的動詞

本論文內容組織如下。第一章為說明圖片字幕生成研究背景與目的，並簡述本論文的方法；第二章探討之前在圖片字幕的相關研究；第三章為講述本研究提出的方法；第四章為實驗結果與分析；第五章為總結與未來改善的方向。

2. 相關研究

2.1. 圖片字幕生成

圖片字幕生成可以透過分析靜態影像生成自然語言描述，目前將機器學習應用於該領域的研究相當普遍，像是運用近期熱門的生成對抗網路(Generative Adversarial Networks，GAN)於非監督式學習(Unsupervised learning)的圖片字幕生成[4]。此研究的架構圖如圖 6[4]所示。首先透過 CNN 萃取圖片的視覺特徵，再將此特徵輸入生成器(Generator)生成圖片的自然語言描述，最後使用鑑別器(Discriminator)鑑別此描述的真實性。上述兩組元件設計基於 LSTM cells。在訓練階段中，利用各種不同的損失更新模型的參數，達到生成最符合圖片的自然語言描述。

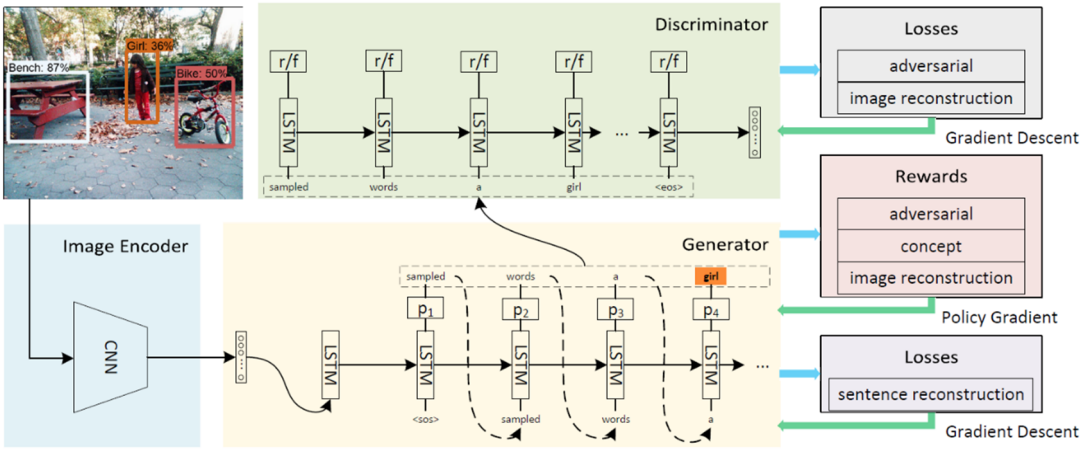


圖 6 非監督式學習圖片字幕生成架構圖

此架構所使用的視覺特徵是影像中的全局區域，這將導致高機率萃取含有雜訊的視覺特徵，像是句子中沒有被描述的視覺區域將被當作視覺特徵。若利用外部語料庫當作模型學習自然語言的知識庫，由於語料庫上並沒有可用的視覺特徵提供訓練導致模型只能生成語料庫中的完整句子。但是，此句子卻不是針對輸入圖片的自然語言描述。 如圖 7[4]依據圖片所生成的完整句子頂多只能描述物件的顏色或材質但沒有更加細膩的動作描述，例如：“looking”、“eating”…。主要原因是非監督式學習使圖片字幕生成模型不能針對學習動作的視覺特徵到自然語言描述的映射。因此，我們認為在圖片字幕生成上，必須透過監督式學習(Supervised Learning)的訓練方式針對學習不同物件之間的視覺特徵到關係描述的映射。

圖 7 圖片字幕生成實驗範例

2.2. 區域提取網路

區域提取網路(Region Proposal Network，RPN)[17]屬於物件偵測(Object Detection)應用，可以學習提取靜態影像中特定的區域，該區域被稱為感興趣區(Region of Interest，ROI)。可控制的圖片字幕生成[15]引入區域提取網路到圖片字幕生成中，架構圖如圖8[15]。使用 Faster R-CNN[17]提取靜態影像中的感興趣區域，在輸入生成字幕模型之前，給予感興趣區域排序訊號，依據給予的順序生成圖片的描述。但是，不同物件之間的關係特徵被考慮得不夠完整，也就是針對不同物件之間的動作描述並沒有特定的視覺區域作為映射，導致生成的圖片字幕僅描述物件之間的關係而缺少了貼切的動作描述。圖9[15]為可控制的圖片字幕生成的生成範例，我們可以發現他們僅用了介係詞描述物件之間的關係，而缺少動作描述，像是以動詞“walking”來描述物件“woman”的動作。

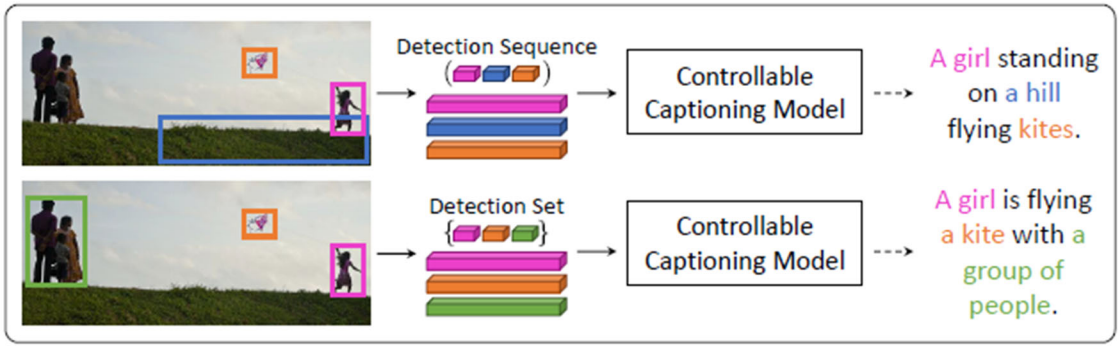


圖 8 可控制的圖片字幕生成架構圖

圖 9 可控制的圖片字幕生成範例

2.3. 注意力機制

Zhang 等人[13]分別針對靜態影像與完整句子進行去雜訊的前處理，並且引入注意力機制(Attention Mechanism)，架構圖如圖 10[13]。靜態影像透過目前最新的區域提取網路架構 Faster R-CNN[17]提取圖片中的物件區域；完整句子使用 Stanford NLP[13]工具提取主要描述。利用長短期記憶單元與額外的注意力機制學習物件區域與主要描述之間的映射。區域提取網路確實達到去除雜訊的效果，但在不同物件之間的關係上，缺少更細膩的視覺特徵。透過注意力機制，模型可以更專注在個別物件與詞彙的關係上。然而，忽略物件與物件之間的關係特徵，導致生成的動作詞彙不夠貼切。如圖 11[13]，實驗範例可以發現“man”與“food”的動作描述只生成“holding”，但更加細膩的動作描述是“eating”。

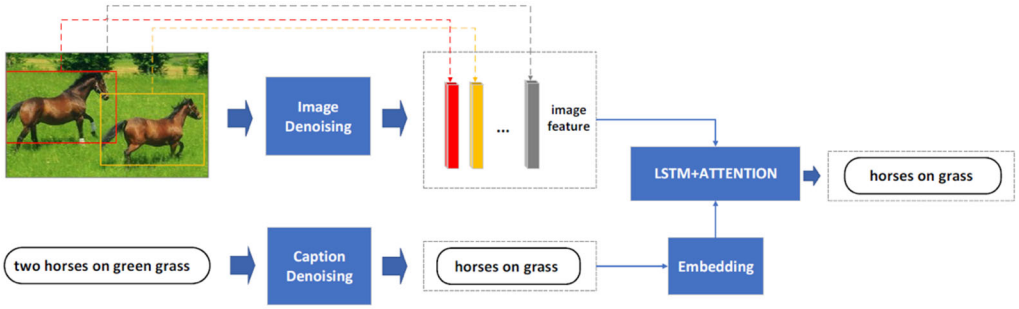


圖 10 圖片字幕生成整體架構圖

圖 11 圖片字幕生成實驗範例

2.4. 總結

圖片字幕生成的整體架構從只能學習整張圖片到完整句子的映射，進一步發展到可以透過區域提取網路所提取的感興趣區域生成可控制的自然語言描述[15]，並額外引入注意力機制[13]到此研究上。我們發現目前的研究成果有兩個缺點，一個缺點是在區域提取網路所提取的感興趣區域上，缺少不同物件之間更加細膩的視覺特徵；另一個則是注意力機制不能完整表示不同物件之間的關係特徵。由於上述的兩個主要缺點，導致字幕生成器在生成關係描述上不夠明確。我們需要探討的議題如下：

1. 從靜態影像中提取更細膩的關係特徵
2. 篩選多尺度區域
3. 字幕生成器生成更細膩的動作關係描述
4. 多尺度區域與圖片字幕生成在描述準確率的改善

我們與先前研究[13], [15], [19], [20]一樣利用區域提取網路提取靜態影像中的個別物件區域，但不同之處是除了提取個別物件本身之外，額外提取在物件與物件之間的聯集區域上的多尺度區域並篩選掉多餘的區域以及雜訊。多尺度區域可以加強學習不同物件之間更細膩的關係描述，提供更符合圖片的自然語言描述。字幕生成器在訓練階段中進行階層式屬性的輔助監督，以學習含有細膩屬性的描述。我們將分析多尺度區域對圖片字幕生成在描述準確率的改善，並且與先前研究比較圖片字幕生成的準確率。

3. 利用多尺度感興趣區域之細微關係提供圖片字幕

為了進一步提高圖片字幕生成在不同物件之間的關係描述細膩度，我們提出利用不同物件之間的多尺度區域來提取更細膩的關係特徵。首先，我們將介紹本研究提出的圖片字幕生成深度學習神經網路整體架構。接著，詳細說明如何透過區域提取網路提取圖片中的個別物件區域與多尺度區域，以及多尺度區域的篩選機制。接著說明特徵提取與融合的步驟，我們透過外觀相似度(Appearance similarity) 的融合機制整合所有多尺度區域特徵以獲得關係特徵。最後，經由排序機制將所有特徵進行重整，透過長短期記憶單元將帶有時間序列結構的物件特徵與關係特徵針對圖片到完整句子的映

射進行訓練，並且我們在訓練階段中引入階層式屬性的輔助監督。

3.1. 深度學習神經網路整體架構

本研究提出的圖片字幕生成深度學習神經網路架構如圖 12。透過區域提取網路提取輸入靜態影像中個別物件區域、物件聯集區域以及多尺度區域，使模型能夠提取不同物件之間更加細膩的關係特徵。經由 FCNN 萃取所有感興趣區域的特徵圖，分別為物件特徵f^{obj}、聯集特徵f^{union}以及多尺度區域特徵F^{multi_scale}。由於f^{union}與F^{multi_scale}都屬於不同物件之間的關係特徵，因此利用特徵融合機制取得融合的關係特徵f^{RLAT}。利用排序機制將所有的物件特徵與關係特徵根據真實描述的順序進行排序並合併所有特徵。長短期記憶單元先以含有時間序列結構的機制提取特徵，再透過每一個時間 t 的特徵丟入 Softmax 激活函數進行自然語言描述的生成。在訓練階段上，字幕生成器的損失指標除了計算完整句子，同時也針對階層式屬性的輔助監督。此監督分別為粗略與細膩屬性的損失計算，用來更新第一層與第二層的 LSTM cells 參數。

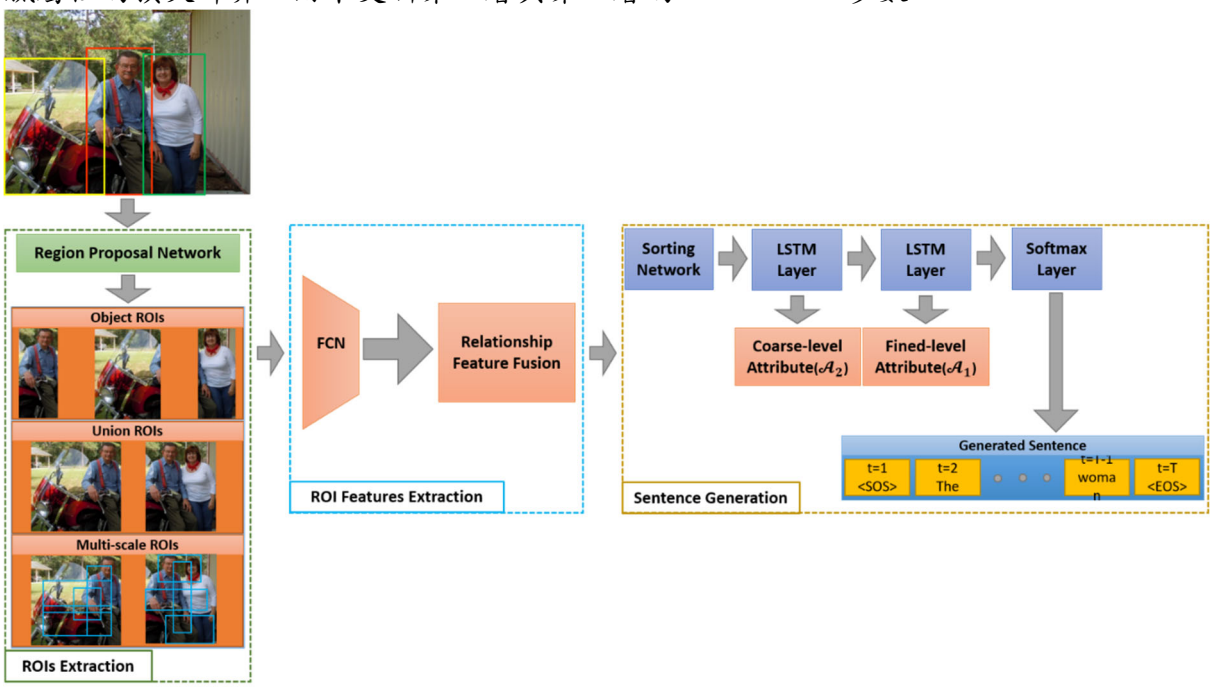


圖 12 基於多尺度區域之靜態影像生成字幕整體網路架構

3.2. 提取與篩選感興趣區域

3.2.1. 感興趣區域

本研究將所有感興趣區域分為三個類別，圖 13 為三個感興趣區域類別的範例圖。圖 13((a)為個別物件區域，13(b)為物件聯集區域，13(c)為多尺度區域。我們接下來將分別詳細說明每種感興趣區域的提取方法，其中因為多尺度區域中可能含有雜訊，我們將透過並交比來篩選多尺度區域。

(a)個別物件區域 (b)物件聯集區域 (c)多尺度區域

圖 13 視覺區域提取範例

3.2.2. 個別物件區域提取

我們使用目前最新的區域提取網路架構 Fast R-CNN 以及 VGG-16 模型來提取圖片中的個別物件區域。透過區域提取網路對圖片進行物件偵測，我們可以提取出圖片裡的物件的位置以及大小，並且記錄所有個別物件的邊界框。如圖 13(a)所示，我們將紀錄所有由物件提取網路所標註的邊界框，以提取出所有人、男孩、飛盤以及腳踏車等個別物件區域。

3.2.3. 計算物件聯集區域

透過物件提取網路提取出所有個別物件區域後，我們將所有個別物件區域兩兩聯集，即可計算出所有物件聯集區域。如圖 14 所示，偵測圖中的物件後，將物件區域兩兩一組，計算並且紀錄所有物件邊界框聯集區域。圖 14 中，顯示了三組物件聯集區域，分別為腳踏車與男孩聯集區域、飛盤與男孩聯集區域以及飛盤與男人聯集區域。

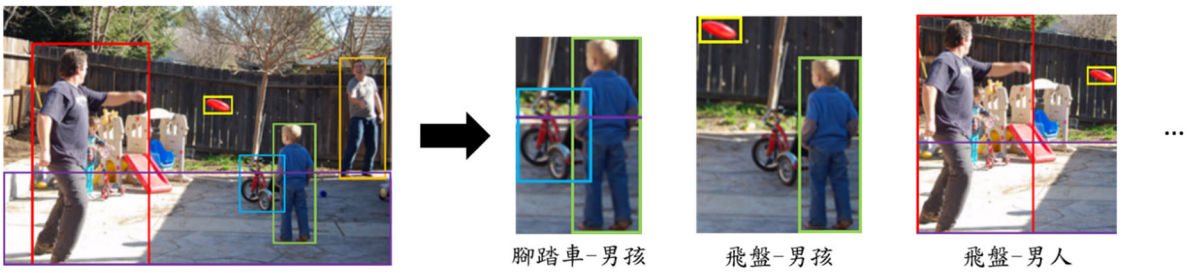
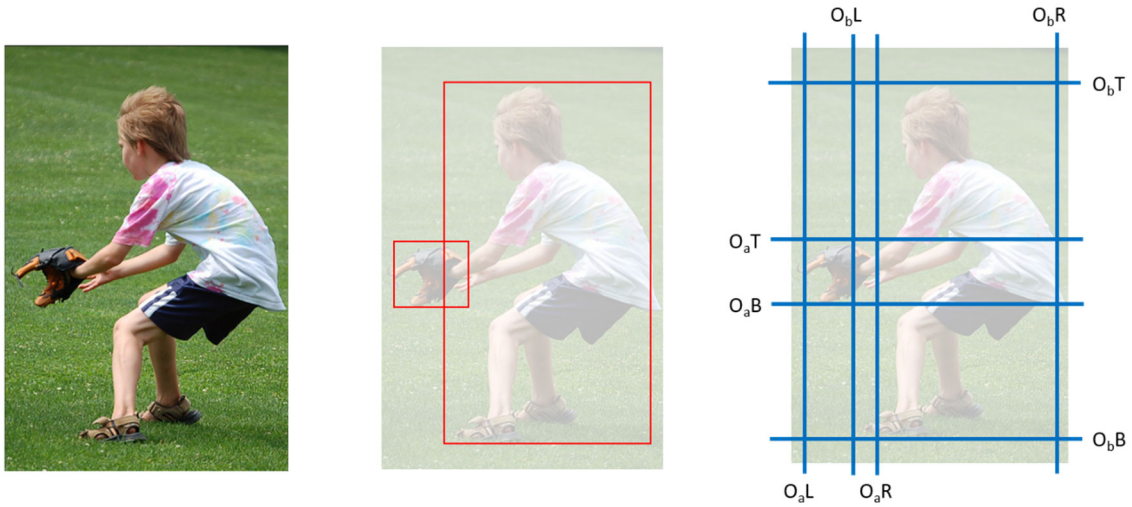


圖 14 物件聯集區域範例

3.2.4. 提取多尺度區域

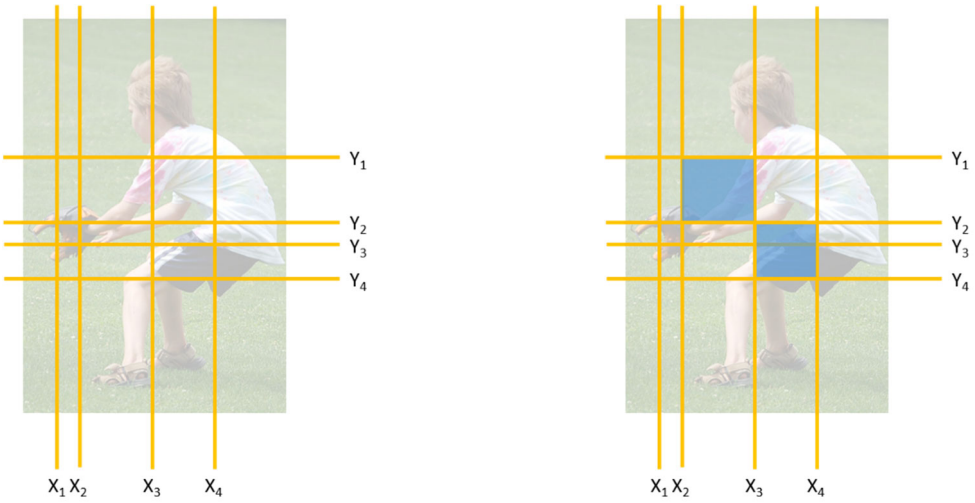
多尺度區域為物件聯集區域內各種不同大小的區域。我們將在此小節詳細說明如何在一組物件聯集區域之中，提取所有多尺度區域。以圖 15(a)的聯集區域為例，此聯集區域為棒球手套與男孩的聯集區域。我們將棒球手套定義為物件 a，以符號O_a表示。男孩則定義為物件 b，以符號O_b表示。圖 15(b)為O_a與O_b的邊界框。透過個別物件的邊界框，我們可以分別獲得兩個個別物件區域在橫軸與縱軸的邊界。如圖 15(c)所示，我們以O_aL表示O_a的左邊界，以O_aR表示O_a的右邊界，以O_aT表示O_a的上邊界，以O_aB表示O_a的下邊界。物件 b 以相同的符號表示。



(a) (b) (c)

圖 15 個別物件邊界

接著，我們將會在個別物件區域的縱軸與橫軸邊界之間尋找兩個參考邊界，將其平分為三等分。最後，我們總共會獲得兩個物件在縱軸與橫軸上的四個參考邊界。將四個參考邊界排序並且編號，我們可以在橫軸與縱軸上，分別獲得四個參考邊界。如圖 16(a)所示，排序後的橫軸參考邊界以X_1、X_2、X_3以及X_4表示，縱軸則以Y_1、Y_2、Y_3以及Y_4表示。獲得參考邊界後，則可以提取參考邊界之間的所有區域當作多尺度區域。圖 16(b)中，顯示兩個多尺度區域的範例。



(a) (b)

圖 16 多尺度區域計算

3.2.5. 篩選多尺度區域

透過 3.2.4 節的方法，我們可以提取物件聯集區域中的所有多尺度區域。但是在所有多尺度區域中，部分多尺度區域可能包含過多或過少的個別物件區域。過多的個別物件區域過於接近個別物件區域，可能妨礙深度學習模型的訓練。而過少的個別物件區域不足以被用來表示物件與物件之間的關係特徵。因此，我們必須過濾上述兩種情形的多尺度區域，以確保多尺度區域品質。

多尺度 ROIs 的篩選機制參考先前研究[21]，他們根據亂數(Random)與鄰近(Nearest)的選取方式進行實驗評估，發現鄰近的效能相對較好，因此我們選擇鄰近。它主要是根據邊界框並交比來篩選多尺度區域計算兩個區域之間的交集與聯集區域之比。首先定義 a 與 b 物件區域為R_a與R_b，聯集區域為$R_{a_b}^{Union}$，多尺度區域為$R_{a_b}^{multi_scale}$。利用邊界框並交比篩選多尺度 ROIs 如圖 17，圖 17(a)為兩個物件的聯集區域$R_{a_b}^{Union}$。在多尺度區域若未同時包含兩個物件的區域，則被視為雜訊，像是$\mathrm{IoU}\left(R_{a_b}^{multi_{scale}}, R_a\right) = 0$或$\mathrm{IoU}\left(R_{a_b}^{multi_{scale}}, R_b\right) = 0$，如圖 17(b)至圖 17(d)。若多尺度區域包含過多的聯集區域也視為雜訊，也就是$\mathrm{IoU}\left(R_{a_b}^{multi_{scale}}, R_{a_b}^{Union}\right) = 0$，如圖 17(e)。其中閥值 h 經由實驗評估後，設定閥值為 0.5。

圖 17 聯集與多尺度 ROIs

3.3. 感興趣區域的特徵提取與融合

3.3.1. 視覺特徵提取

透過 3.2 節的方法，我們可以獲得所有感興趣區域。我們將所有感興趣區域以集合表示$\mathrm{R} = \{r_i | \mathrm{i} = 1, 2, \ldots, \mathrm{N}\}$，其中 N 為個別物件、聯集區域以及多尺度區域的總數。所有感興趣區域經由全卷積神經網路萃取高階的視覺特徵，架構使用 VGG-16[22]。它是由多層的卷積層與最大池化層所組成。r_i經由 VGG-16 模型萃取視覺特徵可以獲得f_i，如公式 3.1。我們將視覺特徵集合定義為$\mathrm{F} = \{f_i | \mathrm{i} = 1, 2, \ldots, \mathrm{N}\}$。其中F可以分為個別物件視覺特徵、物件聯集視覺特徵以及多尺度視覺特徵。

$$f_i = VGG\_16(r_i) \tag{3.1}$$

為了特徵融合的說明，我們以圖片中包含兩個物件為例，並且定義其符號。兩個物件分別定義為 a 物件O_a與 b 物件O_b。在所有視覺特徵集合F則可以分為兩個個別物件視覺特徵(a 物件f_a^{obj}與 b 物件f_b^{obj})、一個物件聯集視覺特徵($f_{a\_b}^{uniob}$)以及多尺度視覺特徵$\mathrm{F}_{\mathrm{a_b}}^{\mathrm{multi_scale}} = \{\mathrm{f}_i^{\mathrm{multi_scale}} | \mathrm{i} = 1, 2, \ldots, \mathrm{k}\}$。其中多尺度視覺特徵$\mathrm{F}_{\mathrm{a_b}}^{\mathrm{multi_scale}}$為一個集合，裡面包含 k 個經過篩選後並且提取視覺特徵的多尺度視覺特徵。

3.3.2. 視覺特徵融合

獲得所有視覺特徵後，我們會將包含不同物件區域所提取出的視覺特徵進行融合，來計算物件之間的關係特徵f^{RLAT}。由於聯集與多尺度視覺特徵都屬於不同物件之間的關係特徵，因此會以聯集與多尺度視覺特徵進行融合。我們將$\mathrm{f}_{\mathrm{a_b}}^{\mathrm{union}}$與$\mathrm{F}_{\mathrm{a_b}}^{\mathrm{multi_scale}}$的視覺特徵合併表示為$\mathrm{F}_{\mathrm{a_b}}^{ROIs} = \{\mathrm{f}_i^{ROIs} | \mathrm{i} = 1, 2, \ldots, \mathrm{k}+1\}$。

視覺特徵的融合機制參考先前研究的實驗評估[21]，分別針對使用外觀相似度(Appearance similarity)、全連接層(Fully connected layer)、平均池化層(Average-pooling layer)以及最大池化層(Max-pooling layer)的融合技術進行評估。在上述的評估中，外觀相似度得到相對較好的效能。因此我們選擇此方法融合聯集與多尺度視覺特徵。由於考慮到兩個物件的關係特徵，因此計算方式被重新定義為公式 3.2。

$$\mathrm{f}\left(f_a^{obj}, f_a^{obj}, \mathrm{F}_{\mathrm{a\_b}}^{ROIs}\right) = \sum_{i=1}^{k+1} G\left(f_a^{obj}, f_a^{obj}, f_i^{ROIs}\right) f_i^{ROIs} \tag{3.2}$$

視覺特徵融合函數 f 的輸入為物件f_a^{obj}，f_a^{obj}以及物件聯集與多尺度合併集合$\mathrm{F}_{\mathrm{a_b}}^{ROIs}$的視覺特徵，函數將會計算集合$\mathrm{F}_{a_b}^{ROIs}$中每一個感興趣區域與兩個個別物件的外觀相似度，其中計算方式為公式 3.3 所示。

$$G\left(f_a^{obj}, f_a^{obj}, f_i^{ROIs}\right) = \frac{\exp\left(f_a^{obj}, f_i^{ROIs}\right) + \exp\left(f_a^{obj}, f_i^{ROIs}\right)}{\sum_{i=i}^{k+1}\left(\exp\left(f_a^{obj}, f_i^{ROIs}\right) + \exp\left(f_a^{obj}, f_i^{ROIs}\right)\right)} \tag{3.3}$$

外觀相似度計算函數G針對感興趣區域特徵f_i^{ROIs}與個別物件特徵分別做內積並經過指數函數。再將兩個輸出值相加並根據所有感興趣區域特徵f_i^{ROIs}的外觀相似度做正規化，最後獲得感興趣區域與兩個個別物件的外觀相似度。

3.4. 長短期記憶單元生成字幕

3.4.1. 排序機制

因為長短期記憶單元的輸入與輸出與時間序列相關，我們透過排序機制將物件區域特徵f^{obj}以及關係特徵f^{RLAT}，根據完整句子描述的先後順序調整特徵順序。關係特徵為句子中不同物件之間的動作次序，如圖 18，“man”與“skateboard”之間的關係特徵$f_{a_b}^{RLAT}$被排在物件特徵f_b^{obj}之前，“man”與“road”之間的關係特徵$f_{a_c}^{RLAT}$則被排在物件特徵f_c^{obj}之前，依照此形式進行視覺特徵的排序，最後得到$\mathrm{F}^{sorted} = \{\mathrm{f}_a^{obj}, f_{a_b}^{RLAT}, f_b^{obj}, f_{a_c}^{RLAT}, f_c^{obj}\}$。

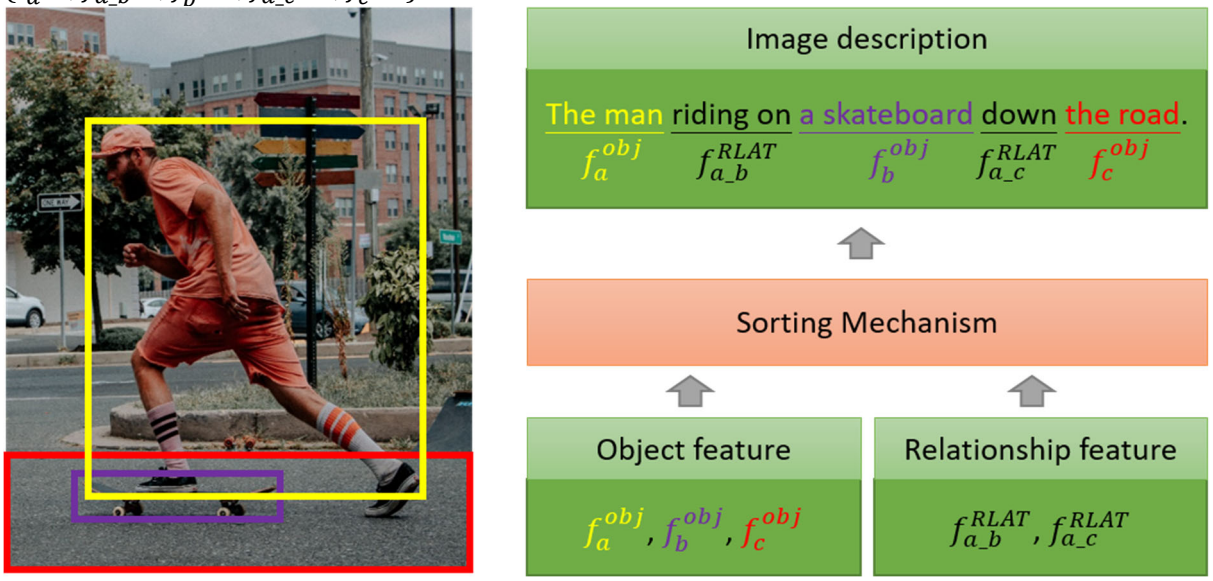


圖 18 區域視覺特徵排序範例

3.4.2. 長短期記憶單元

本研究利用長短期記憶單元學習視覺特徵F^{sorted}到自然語言描述的映射，其計算方法以公式 3.4 與公式 3.5 表示：

$$h_t = LSTM(F^{sorted}, h_{t-1}) \tag{3.4}$$

$$y^t = Softmax(h_t) \tag{3.5}$$

首先將F^{sorted}與前一個長短期記憶單元輸出特徵h_{t-1}輸入長短期記憶單元，得到時間 t 的特徵h_t，最後利用 Softmax 激活函數獲得時間 t 的文字y^t。Softmax 主要針對學習目標的類別，達到增加模型學習的速率，因此選擇 Softmax 當作架構的輸出層。在生成字幕的輔助監督指標上受到[21]的影響。為了使字幕生成器能夠提高針對細膩屬性生成的準確率，除了計算完整句子的損失函數，同時也加入語言屬性的損失進行額外的階層式輔助監督，分為粗略階級與細膩階級的損失計算。粗略階級屬性 A2 像是“person”或“cloth”等等，細膩階級屬性 A1 像是“person”底下的“person”、“girl”或“persons”等等的屬性描述，如圖 19[21]。在詞彙分割的工具上，我們參考[21]使用自然語言處理工具(Natural language processing toolkit，NLTK)。

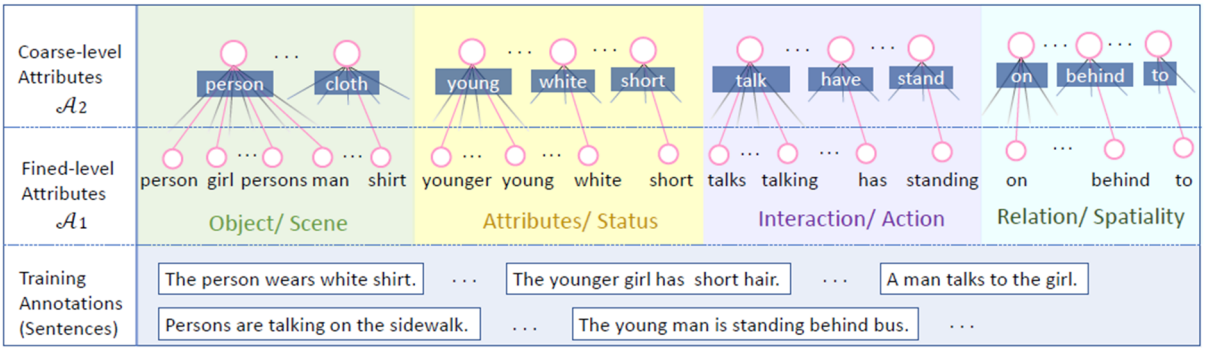


圖 19 階層式屬性與完整句子的表示

損失函數分別為粗略階級 A2、細膩階級 A1 與完整句子，計算公式 3.6、3.7 以及 3.8 表示：

$$L^{A_2}\left(y^{A_2} - \hat{y}_i^{A_2}\right) = -\sum_{i=1}^{n^{A_2}} y^{A_2} \log(\hat{y}_i^{A_2}) \tag{3.6}$$

$$L^{A_1}\left(y^{A_1} - \hat{y}_i^{A_1}\right) = -\sum_{i=1}^{n^{A_1}} y^{A_1} \log(\hat{y}_i^{A_1}) \tag{3.7}$$

$$L^{sen}\left(Y^{sen} - \hat{Y}^{sen}\right) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n^{word}} y^t \log(\hat{y}_i^t) \tag{3.8}$$

y^{A_2}與y^{A_1}分別為 Ground Truth 的粗略與細膩屬性，$\hat{y}_i^{A_2}$與$\hat{y}_i^{A_1}$分別為預測的粗略與細膩屬性，$Y^{sen} = \{y^t | \mathrm{t} = 1, 2, \ldots, \mathrm{T}\}$為 Ground Truth 的完整句子，$\hat{Y}^{sen} = \{\hat{y}_i^t | \mathrm{t} = 1, 2, \ldots, \mathrm{T}\}$為預測的完整句子，T 為完整句子的長度，$n^{A_2}$，$n^{A_1}$與$n^{word}$分別表示粗略屬性、細膩屬性與總辭彙的類別個數。損失函數計算使用交叉熵(Cross-Entropy)[23]，它主要利用概率的數值計算損失指標，因此適用於分類問題。由於我們使用階層式的輔助監督，因此必須分層定義長短期記憶單元的損失函數。其計算方法以公式 3.9 以及公式 3.10 表示：

$$\arg\min_{\theta^{f\_l}} L^{f\_l} = \arg\min_{\theta^{f\_l}} (L^{A_2} + L^{A_1} + L^{sen}) \tag{3.9}$$

$$\arg\min_{\theta^{s\_l}} L^{s\_l} = \arg\min_{\theta^{s\_l}} (L^{A_2} + L^{sen}) \tag{3.10}$$

L^{f_l}、L^{s_l}分別為第一層與第二層長短期記憶單元的損失函數，θ^{f_l}、θ^{s_l}為第一層與第二層長短期記憶單元的參數。L^{f_l}的損失計算是屬於整體的損失，考慮完整句子損失之外，也加入粗略與細膩屬性損失。L^{s_l}則主要針對細膩屬性的損失與完整句子損失。在訓練階段中，透過損失函數的梯度(gradient)計算更新不同層的長短期記憶單元參數θ^{f_l}、θ^{s_l}，直到能使L^{f_l}、L^{s_l}最小化或模型收斂。字幕生成器使用兩層長短期記憶單元分別針對粗略與細膩的屬性進行訓練，架構圖如圖 20。

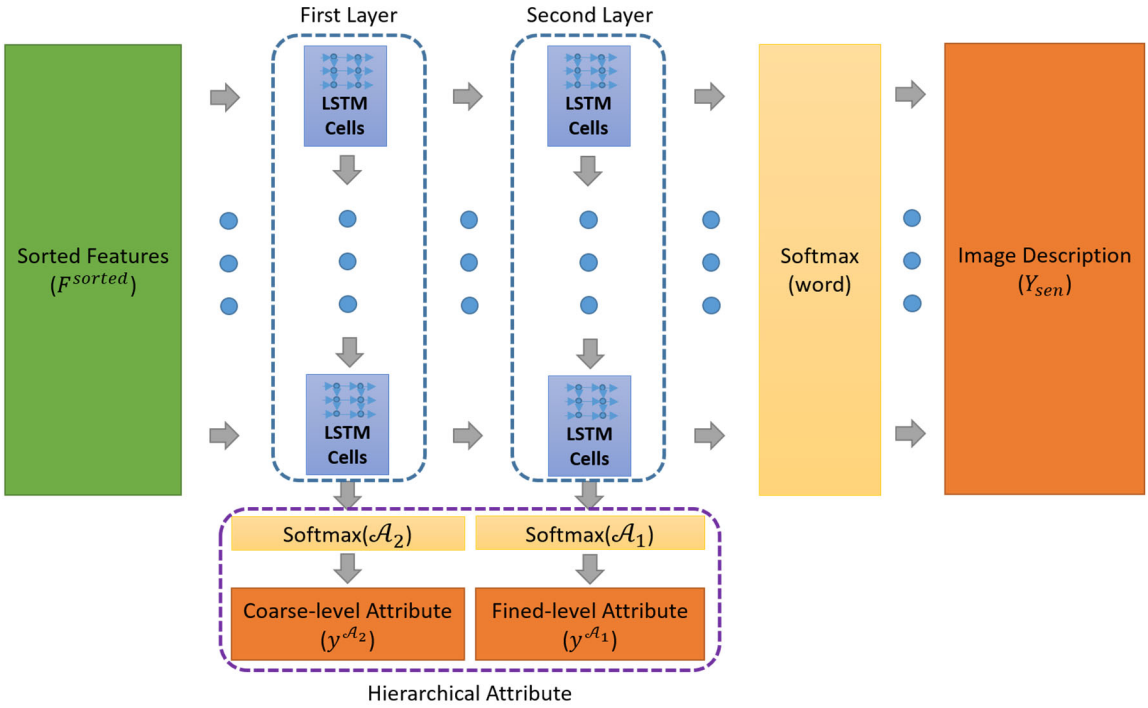


圖 20 LSTM 字幕生成器架構圖

4. 實驗結果

資料集。本研究使用的資料集為 MS-COCO[23]，並且使用 Karpathy split 方法[24]來切分資料集。將完整資料集分為三個部分，分別為訓練集、驗證集以及測試集，訓練集包含 113287 張圖片，而驗證集以及測試集皆為 5000 張圖片。資料集中每張圖片最多有 5 個圖片字幕。

基準線。我們將本研究提出的利用多尺度感興趣區域之細微關係特徵提供圖片字幕架構與可控制的 Image Captioning(SCT)[15]進行比較。SCT 透過區域提取網路提取物件區域，並在萃取物件影像特徵之後加入控制訊號，透過長短期記憶單元學習影像到文字的映射，進而產生圖片字幕。

評估指標。我們參考了 MS-COCO 評估伺服器[25]提供的四種基於 n-gram 的評估指標，分別為 BLEU[26]、ROUGE[27]、METEOR[28]以及 CIDEr[29]，並且加入 SCT 使用的 SPICE[30]，評估圖片字幕在 Scene Graph 上的相似度。另外，為了評估物件之間的細微關係，我們額外加入了 TIGEr[31]指標。

4.1. 圖片字幕生成範例

圖 4.3 顯示了一些我們的模型所產生的圖片字幕範例。由圖 21(a)與圖 21(b)兩個圖片字幕範例中我們可以發現，得益於 RPN 完整地提取圖片中的物件區域，讓我們的模型在描述個別物件上更加完整。在圖 21(c)與圖 21(d)中可以看到，我們的模型在描述動作更加細膩，例如圖 21(c)中，我們描述斑馬的動作是 walking，而 SCT 則是用 standing 來描述斑馬的動作。我們認為這是因為我們加入多尺度 ROIs 影像特徵，讓模型可以透過更細微的影像區域學習影像到動作描述之間的映射。然而，對於靜態的圖片，尤其是物件上有密集的小物件，例如圖 21(e)與圖 21(f)，我們的模型往往會重複輸出相同的詞語，因而降低了評分。我們認為這是因為經過 RPN 提取物件區域時，同時提取了很多小物件，而我們也計算了許多主要物件與小物件之間的多尺度 ROIs 影像特徵，並且輸入模型，導致模型重複輸出相同的詞語。

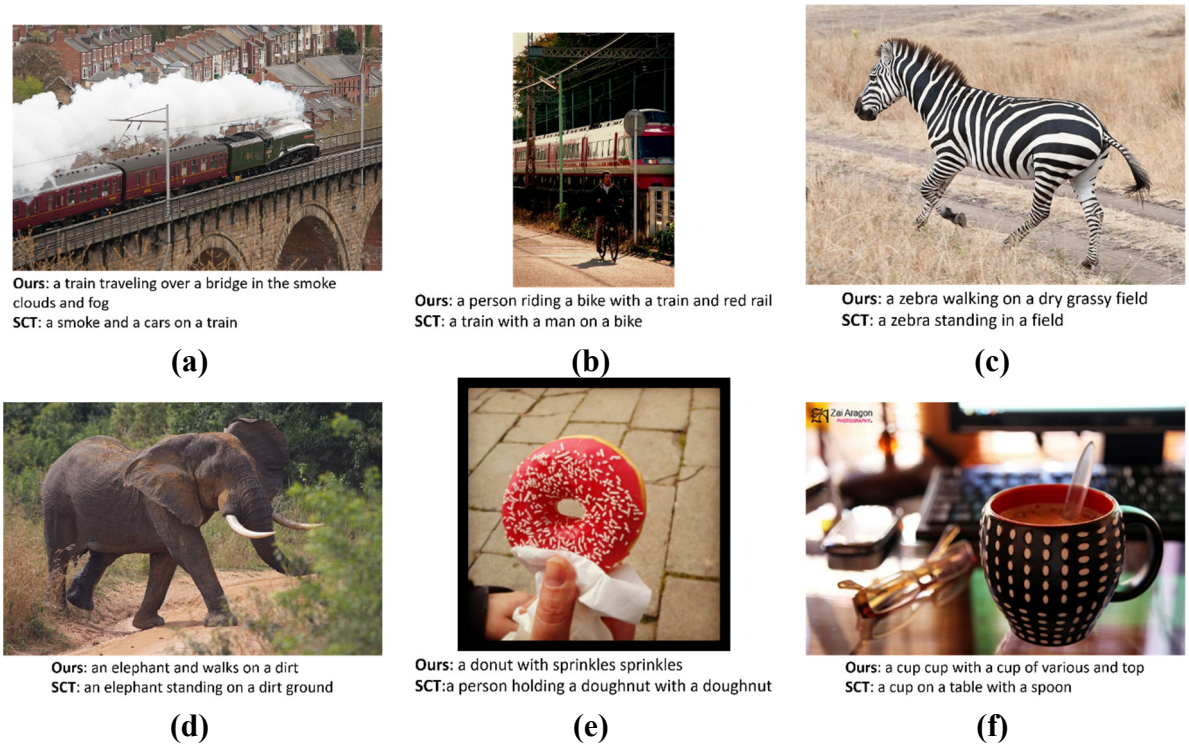


圖 21 圖片字幕生成範例

4.2. 定量結果

表 1 為我們的模型所生成的圖片字幕在六項指標上的分數，並且同時列出與 SCT 比較的兩個沒有加入控制訊號的方法分別為 FC-2K[32]與 Up-Down[33]的成果進行比較。首先，表 2 中前四個評估指標為基於 n-gram 的評估指標。SCT 透過區域提取網路與控制訊號，提取出物件區域。接著透過控制訊號，讓模型學習用較好的順序輸出物件描述。使得 SCT 的圖片字幕在四項基於 n-gram 的評估指標上皆優於沒有使用控制訊號的 FC-2K 與 Up-Down。而我們在四項基於 n-gram 的評估指標上有三項指標效能優於 SCT[15]，分別為 BLEU_4、METEOR 以及 ROUGE。BLEU_4 與 ROUGE 分別著重候選句與參考句中 n-gram 的精確度(precision)與召回率(recall)。得益於區域提取網路與多尺度區域影像特徵，我們的模型輸出的圖片字幕對於物件描述與關係描述可以更精準。同樣是沒有使用控制訊號的方法，我們的模型在 BLEU_4 與 ROUGE 指標的效能優於 SCT。METEOR 指標考慮了同義詞，並且鼓勵連續詞的匹配。我們認為透過多尺度區域

視覺特徵加上長短期記憶模型特性，使我們的模型更容易產生與參考句連續匹配的詞語。因此，我們的模型在 METEOR 指標優於 SCT。最後，在 CIDEr 指標上，我們的效能劣於 SCT。我們認為這可能與我們專注在物件與物件之間的特徵，加入過多多尺度區域視覺特徵，導致模型重複生成相同的詞。而 CIDEr 會對重複出現的詞扣分，所以我們的模型在 CIDEr 的分數上獲得較低的分數。或許在往後的研究加入場景圖(Scene Graph)，透過加上場景圖來調節多尺度區域視覺特徵，以減少模型重複生成同一個詞語的情形。

表 1 各項指標效能比較表

Method	BLEU_4	METEOR	ROUGE	CIDEr	SPICE	TIGEr
FC-2K[32]	10.4	17.3	36.8	98.3	25.5	71.6
Up-Down[33]	12.9	19.3	40.0	119.9	29.3	70.8
SCT[15]	20.9	24.4	52.5	**193.0**	**45.3**	73.7
Ours	**37.8**	**26.6**	**57.9**	105.2	25.4	**74.8**

另一方面，我們的模型在 SPICE 指標上的分數也低於 SCT。SCT 透過控制訊號調整物件輸出的順序，使 SCT 輸出的圖片字幕轉換成場景圖後，物件之間的關係能夠更接近參考句。因此，相較於沒有加入控制訊號的方法，SCT 在 SPICE 指標上優於 FC-2K 與 Up-Down。我們的模型透過影像上的多尺度區域視覺特徵，提取更細膩的影像特徵，讓模型輸出更貼切的詞語描述物件與動作。我們的模型犧牲了物件輸出順序而著重在更貼切的物件與動作描述，因此我們的模型在 SPICE 指標上效能劣於 SCT。最後，得益於區域提取網路提取物件區域以及多尺度區域視覺特徵，我們的模型更能夠專注在單一物件以及物件與物件之間聯集區域的影像特徵。因此，我們的模型在同時考慮文字與圖片匹配程度的 TIGEr 指標上優於其他方法。

5. 總結與未來研究

本研究除了在靜態影像中提取個別物件區域，額外提取並且篩選在不同物件之間聯集區域中的多尺度區域，使模型可以針對視覺特徵與關係描述的映射進行學習。除此之外，透過排序機制讓所有的視覺特徵對應字幕生成器所生成描述的時間序列，以利模型訓練進行。我們透過區域提取網路提取個別物件區域，另外加入物件與物件之間聯集區域的多尺度區域影像特徵，最後透過長短期記憶模型輸出與時間序列相關的圖片字幕。讓我們的模型能夠在基於 n-gram 的評估指標上獲得更高的分數，並且在同時比較文字匹配與考慮影像特徵的評估指標 TIGEr 上，也獲得較高的分數。

從目前的實驗數據來看，本研究透過多尺度區域影像特徵確實能夠有效地提升基於 n-gram 評估指標的分數，但是也因為過多的多尺度區域影像特徵導致模型輸出重複詞語。在未來的研究中，可能透過加入也能夠描述物件之間關係的場景圖(Scene Graph)，作為模型的輸入，以調節物件聯集區域多尺度區域特徵的數量，降低模型發生輸出重複詞語的可能性。並且透過加入場景圖學習文字上物件與物件之間的關係，以提升 SPICE 指標的分數。

致謝

本論文承蒙科技部計畫 MOST 109-2221-E-024-011 經費補助，對研究資源挹注助益頗多，特此致謝。

參考文獻

[1] J.Kim, T.Misu, Y.-T.Chen, A.Tawari, andJ.Canny, “Grounding Human-To-Vehicle Advice for Self-Driving Vehicles,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] K.Mori, H.Fukui, T.Murase, T.Hirakawa, T.Yamashita, andH.Fujiyoshi, “Visual explanation by attention branch network for end-to-end learning-based self-driving,” *IEEE Intell. Veh. Symp. Proc.*, vol. 2019-June, no. Iv, pp. 1577–1582, 2019.

[3] H.Li, P.Wang, C.Shen, andA.Van DenHengel, “Visual question answering as reading comprehension,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 6312–6321, 2019.

[4] Y.Feng, L.Ma, W.Liu, andJ.Luo, “Unsupervised Image Captioning,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] A. F.Biten, L.Gomez, M.Rusinol, andDi.Karatzas, “Good news, everyone! context driven entity-aware captioning for news images,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 12458–12467, 2019.

[6] Y.Zheng, Y.Li, andS.Wang, “Intention oriented image captions with guiding objects,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 8387–8396, 2019.

[7] A.Deshpande, J.Aneja, L.Wang, A. G.Schwing, andD.Forsyth, “Fast, diverse and accurate image captioning guided by part-of-speech,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 10687–10696, 2019.

[8] O.Vinyals, A.Toshev, S.Bengio, andD.Erhan, “Show and tell: A neural image caption generator,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 3156–3164, 2015.

[9] L.Guo, J.Liu, P.Yao, J.Li, andH.Lu, “MSCap: Multi-Style Image Captioning With Unpaired Stylized Text,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] K.Shuster, S.Humeau, H.Hu, A.Bordes, andJ.Weston, “Engaging image captioning via personality,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 12508–12518, 2019.

[11] C.Gan, Z.Gan, X.He, J.Gao, andL.Deng, “StyleNet: Generating attractive visual captions with styles,” *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 955–964, 2017.

[12] A.Mathews, L.Xie, andX.He, “SemStyle: Learning to Generate Stylised Image Captions Using Unaligned Text,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 8591–8600, 2018.

[13] Y.Zhang, Y.Ding, R.Wu, andF.Xue, “A Denoising Framework for Image Caption,” in *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 2019, pp. 825–832.

[14] J.Johnson, A.Karpathy, andL.Fei-Fei, “Densecap: Fully convolutional localization networks for dense captioning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4565–4574.

[15] M.Cornia, L.Baraldi, andR.Cucchiara, “Show, control and tell: A framework for generating controllable and grounded captions,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 8299–8308, 2019.

[16] D. J.Kim, J.Choi, T. H.Oh, andI. S.Kweon, “Dense relational captioning: Triple-stream networks for relationship-based captioning,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 6264–6273, 2019.

[17] S.Ren, K.He, R.Girshick, andJ.Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.
[18] S.Hochreiter andJ.Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[19] X.Li andS.Jiang, "Know More Say Less: Image Captioning Based on Scene Graphs," *IEEE Trans. Multimed.*, vol. 21, no. 8, pp. 2117–2130, 2019.
[20] X.Yang, K.Tang, H.Zhang, andJ.Cai, "Auto-Encoding Scene Graphs for Image Captioning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[21] G.Yin, L.Sheng, B.Liu, N.Yu, X.Wang, andJ.Shao, "Context and Attribute Grounded Dense Captioning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[22] K.Simonyan andA.Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv Prepr. arXiv1409.1556*, 2014.
[23] T.-Y.Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision -- ECCV 2014*, 2014, pp. 740–755.
[24] A.Karpathy andL.Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.
[25] X.Chen *et al.*, "Microsoft coco captions: Data collection and evaluation server," *arXiv Prepr. arXiv1504.00325*, 2015.
[26] K.Papineni, S.Roukos, T.Ward, andW.-J.Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
[27] C.-Y.Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
[28] S.Banerjee andA.Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.
[29] R.Vedantam, C.Lawrence Zitnick, andD.Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
[30] P.Anderson, B.Fernando, M.Johnson, andS.Gould, "SPICE: Semantic Propositional Image Caption Evaluation," in *Computer Vision -- ECCV 2016*, 2016, pp. 382–398.
[31] M.Jiang *et al.*, "Tiger: Text-to-image grounding for image caption evaluation," *arXiv Prepr. arXiv1909.02050*, 2019.
[32] S. J.Rennie, E.Marcheret, Y.Mroueh, J.Ross, andV.Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7008–7024.
[33] P.Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.