以深度學習實現手語辨識

專題生:王樸 、吳承威

指導教授：林政宏

[[1]](#footnote-1) *摘要*—這將國際手語轉換成手語。

*關鍵字—sign language、YOLOv3、LSTM*

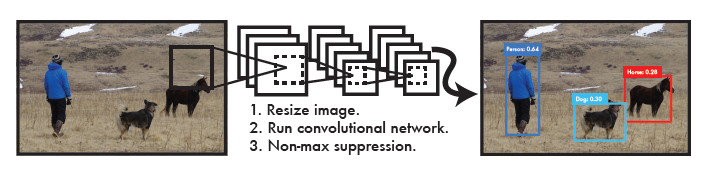
導論:很少有人能理解手語,造成聾啞人士和非聾啞人士溝通的障礙。此外書面交流非常麻煩，因為聾啞人士普遍寫作能力較差。況且書寫溝通是是非常緩慢的,無法即時的應對當下的狀況。因此我們決定使用深度學習的框架-YOLOv3和長短期記憶模型(Long-Short-Term Memory)做一個手語辨識的系統,用來將美國手語(American Sign Language,ASL)轉成文字,增進非聾啞人士與聾啞人士的溝通。

# 原理與架構簡介

## YOLOv3

1. 原理
2. YOLOv3偵測系統:

使用YOLOv3進行物件偵測,有三個步驟,分別為:重訂影像大小、卷積運算、非最大值抑制



1. Bounding box

YOLO將一張相片分成S\*S的網格(grid),每個grid cell 會去偵測是否含有物件(中心)。每個grid cell會預測B個Bunding boxes 、每個bounding box的confidence score以及C的conditional class probability(Pr(Classi|Object))。

每個bound box會有5個預測值,分別為:

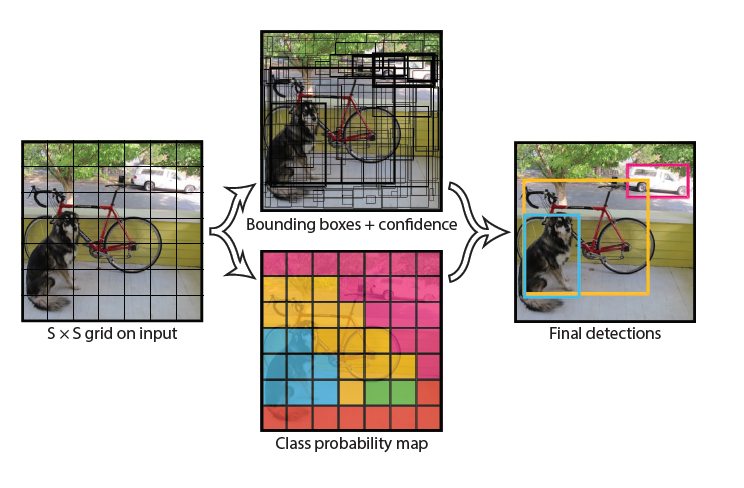
x:bounding box中心相對於grid cell邊界之x座標

y: bounding box中心相對於grid cell邊界之y座標

w:bounding box相對於整個影像之寬度

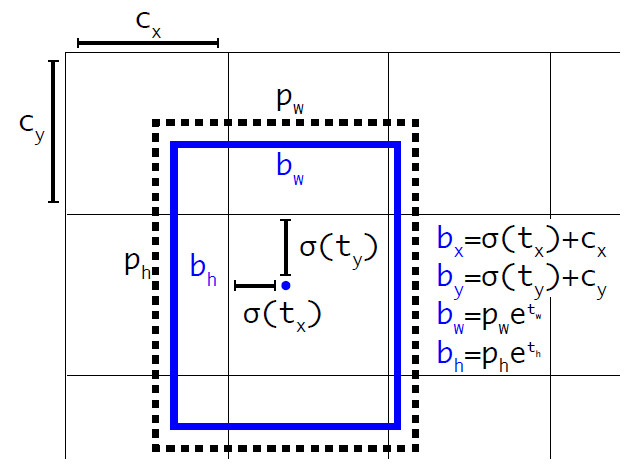
h: bounding box相對於整個影像之高度

Confidence:IOU



1. Anchorbox:

YOLOv3中,使用dimension cluster的方式來預測anchorbox。Network會預測4個座標tx,ty,tw,th。其預測會符合:



以Anchorbox對boundong進行預測,可以解決多尺寸輸入的問題。

1. class-specific confidence score

class-specific confidence score=

Pr(Classi|Object)\*Pr(Object)\* IOU= Pr(Classi)\* IOU

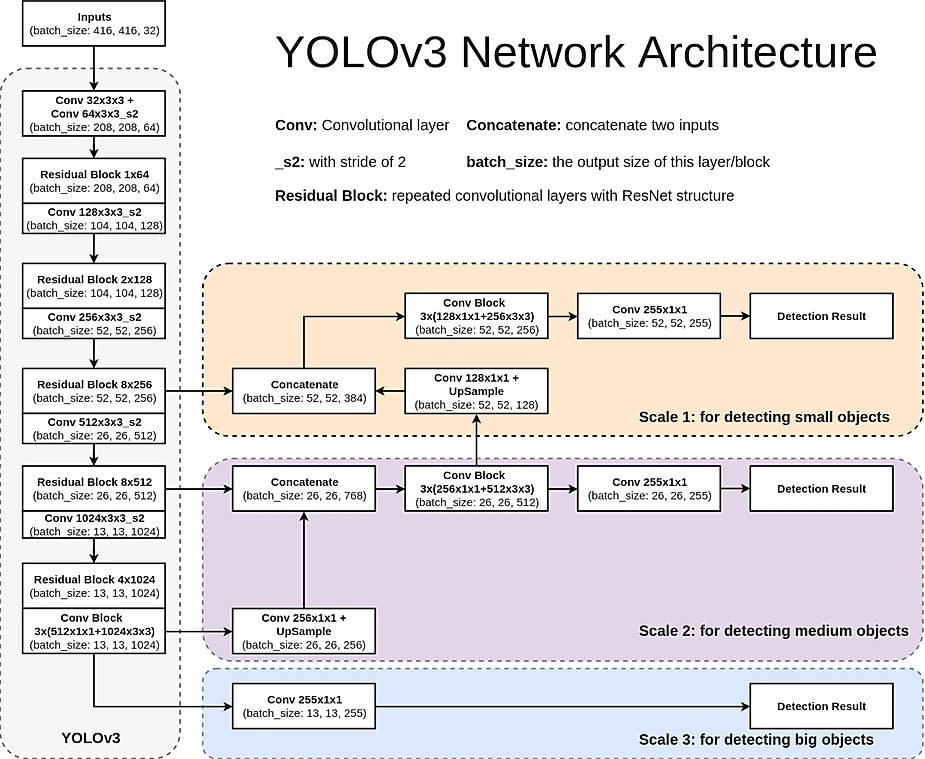
理想狀態下,若有包含物件,則Pr(Object) = 1;若不包含物件,則Pr(Object) = 0。class-specific confidence score反映了一個bounding box 中出現某種class的機率,以及bounding box 與ground的接近程度。

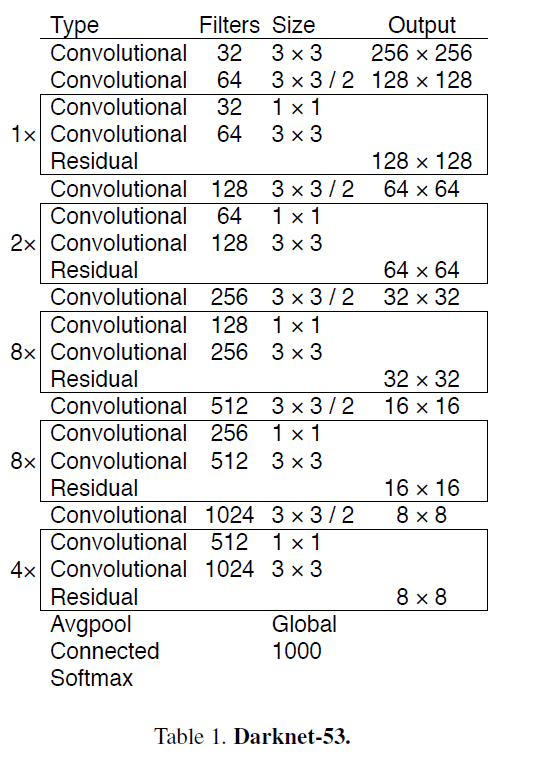
1. Class Prediction

我們使用多標籤分類來預測bounding box 中含有的class,並使用independent logistic classifier。訓練時使用binary croos-entropy loss做class 預測。

1. Feature Extractor:

使用Darknet-53做特徵擷取。Darknet-53為YOLOv2中Darknet-19再加上許多residual network,一共有53層卷積層

1. 架構圖:



## LSTM

1. 原理:

在LSTM的input sequence 中每層會做以下計算:

*it*​=*σ*(*Wii*​*xt*​+*bii*​+*Whi*​*h*(*t*−1)​+*bhi*​)

*ft*​=*σ*(*Wif*​*xt*​+*bif*​+*Whf*​*h*(*t*−1)​+*bhf*​)

*gt*​=tanh(*Wig*​*xt*​+*big*​+*Whg*​*h*(*t*−1)​+*bhg*​)

*ot*​=*σ*(*Wio*​*xt*​+*bio*​+*Who*​*h*(*t*−1)​+*bho*​)

*ct*​=*ft*​∗*c*(*t*−1)​+*it*​∗*gt*

​*ht*​=*ot*​∗tanh(*ct*​)​

其中:

ht:hidden state at time t

*ct*​::the cell state at time t

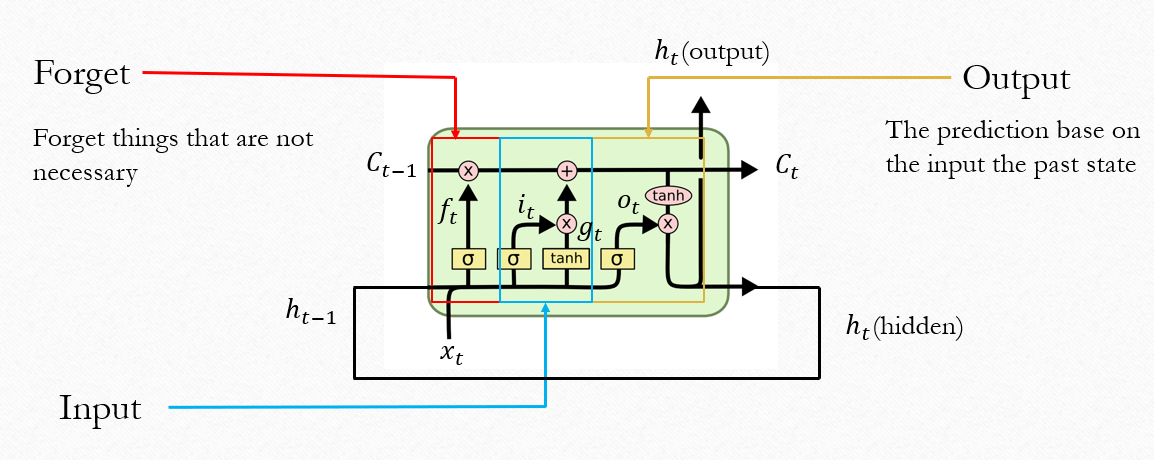
*xt*​ is the input at time t

*h*(*t*−1)​:the hidden state of the layer at time t-1 or the initial hidden state at time 0

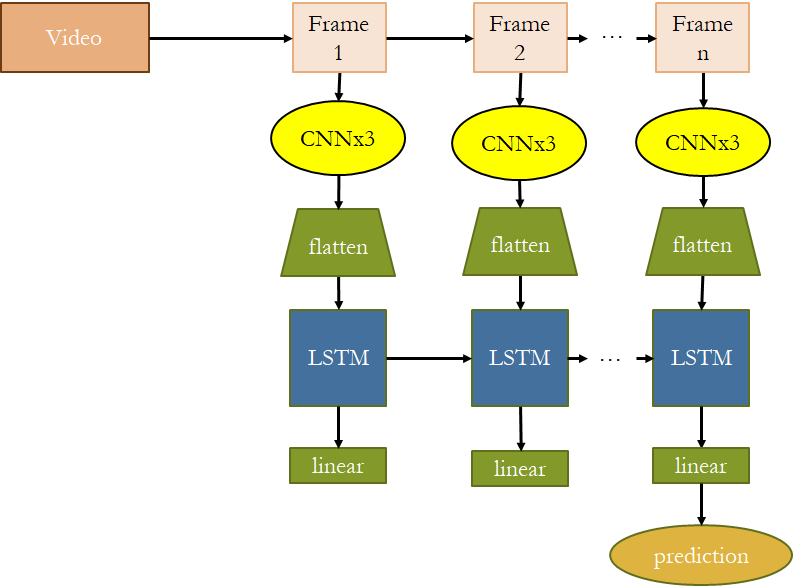
*it:input*​

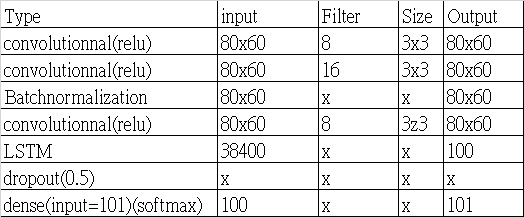
*ft:forget*​

*gt:cell*​

*ot*​:output

1. 架構圖



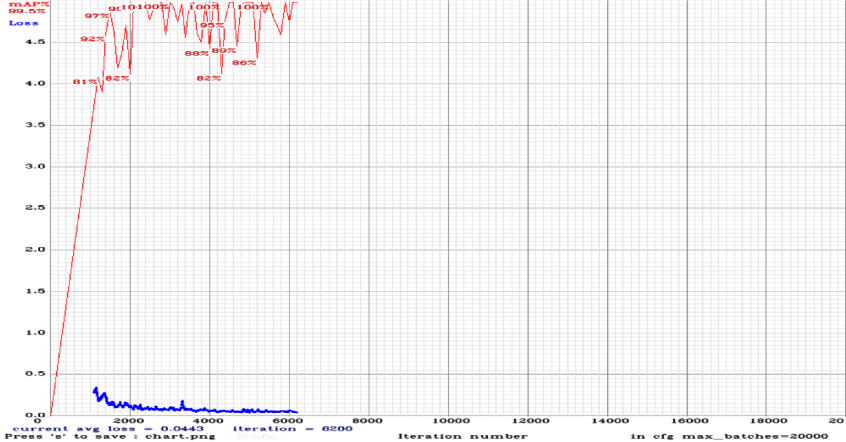


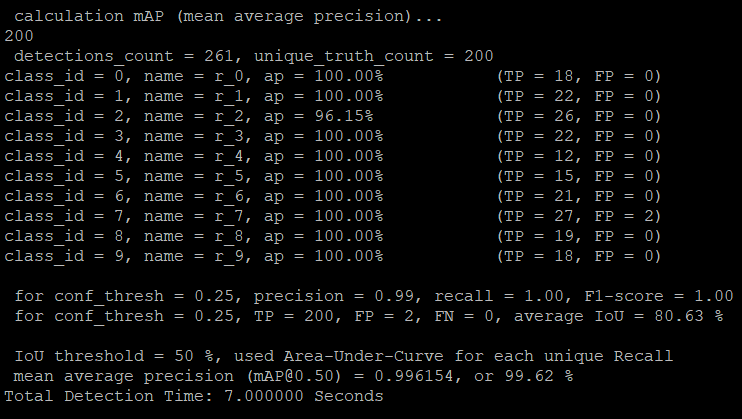
# 實驗步驟、過程、結果

1. YOLOv3:
2. 實驗步驟及過程:
3. 收集data:使用ipone7對不同人、不同背景之下進行拍攝,每一種手語影像各100
4. Label:對每張影像進行標籤化
5. Training:進行訓練,共訓練了6200 iterations

參數調整:

1. Batch=32
2. Subdivisions=64
3. Learning rate=0.001
4. Testing :進行測試,以非訓練所用之影像進行測試,看神經網路是否可以正確框出物件所在之處以及其預測為某類別之機率
5. 實驗結果:





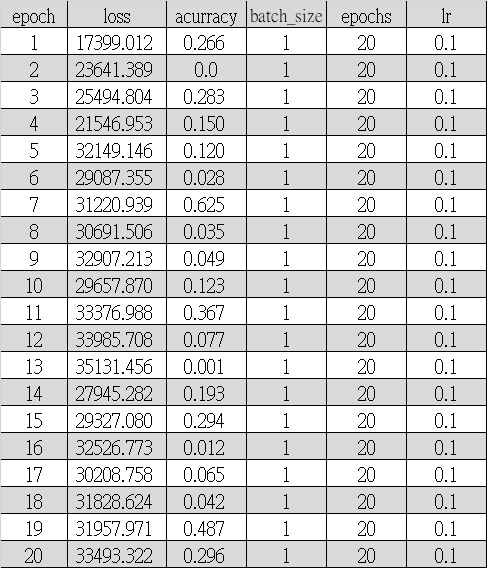
1. 結果討論:

由上圖可知,MAP非常高,可能是因為Train 和test的資料相似度太高,資料度變異性不大。Train到6200 iteration還沒有收斂,可能得要train到10000個iteration。最後因為資料量不夠,很可能造成overfitting。。

1. LSTM:
2. 實驗步驟及過程:
3. 收集data:使用UCF101 dataset
4. Training:進行訓練,共訓練了20 epochs

參數調整:

1. Batch=1
2. Learning rate=0.1
3. Testing :進行測試,以非訓練所用之影片進行測試,看神經網路是否可以正確預測出為何種動作、準確率、損失函數
4. 實驗結果:



1. 討論:

我們測試了不同的learning rate,也新增了linear層用來降維,使用batch nomrmalization、dropuot層防止overfitting可是loss非常高,accuracy還是很低,可能CNN跟LSTM學的東西不同,應該要分開train才行。

# 競賽心得與競賽意見回復

王樸:這學期的專題學到很多東西,像是如何使用keras、pytorch、tensorflow,以及基本的應用舉凡MNIST,CIFAR。此外,自己拍攝手語資料,做資料的處理,再熟悉YOLO的理論以及使用方式,也花了很多的時間。另外,研究LSTM的運作原理、程式,也花了很多心力。全部的東西都只能靠自學,總覺得效率不是很高,不過實驗室的學長人都非常好,很樂意回答我們的各種問題,受益良多。第一個學期就要有報告成果、交論文式的報告有點困難,一方面為了要符合書面報告的格式要額外花很多時間,另一方面第一個學期比較是在確定自己專題的方向以及對背景知識熟悉,很難能有什麼實做的成果。

吳承威:

During this course I learn a lot thing that will be useful, this this start from start studding thing by myself and without the need of teachers then to learn things by doing . Also, in the prosses I learn how to an artificial intelligence works and how we can create, train and use. And at finally I also learn how to write an investigation paper, which in the future, will be very useful for future investigation or jobs.

# 誌謝

感謝實驗室各學長大力協助,關於神經網路架構、程式撰寫、資料收集、數學模型。參考資料

參考資料

[1] Surejya Suresh Sign Language Recognition System Using Deep Neural Network 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS),

[2] Pigou, L., Dieleman, S., Kindermans, P.J.: Sign Language Recognition Using Convolutional Neural Networks. In: ECCV ChaLearn Workshop on Looking at People (2014)

[3] Diniya A J &Arathi Remesh &Dona Jose &Jasmi Davis Volume 5, Issue 3, May-Jun-2019 Sign Language Translator Using CNN Model International Journal of Scientific Research & Engineering Trends

[4] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: CVPR (2016)

[5] ] J. Redmon and A. Farhadi. YOLO9000: Better, faster, stronger. In CVPR, 2017.

[6] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. CoRR, abs/1804.02767, 2018.

1. [↑](#footnote-ref-1)