110學年度大學部專題競賽

國立清華大學資訊工程學系
Department of Computer Science, National Tsing Hua University

擴展視訊會議功能之機器學習模型研究與實作

陳伯瑾　黃寶萱　陳柏均

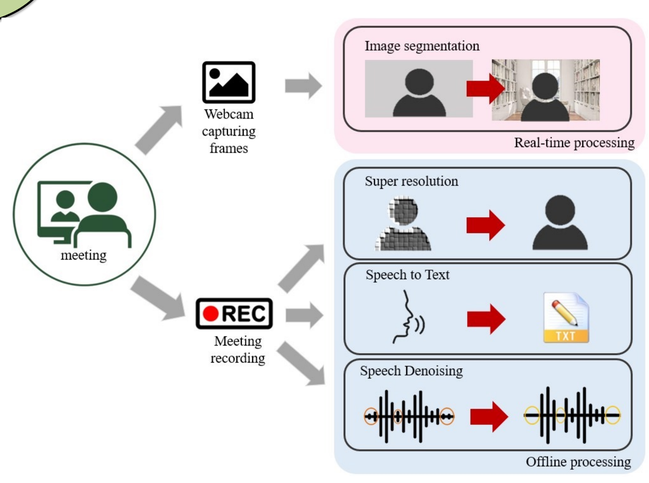
研究動機與目標

2019年年底，新型冠狀病毒爆發後快速的蔓延至世界各國，企業、學校等紛紛實施在家工作、遠距教學，線上通訊顯然已成為疫情下的趨勢，人們對視訊會議軟體的使用需求快速上升。

本研究以現有視訊會議軟體的架構為設計主軸，目標是利用機器學習模型與深度學習模型對影音進行不同的處理：更換會議背景(instance segmentation)、提高影像解析度(super resolution)、會議內容文字記錄(speech to text)、去除影片背景雜音(speech denoising)，並將我們的線上會議系統以網站的方式呈現，提供更完善的視訊會議品質。

研究方法與實作

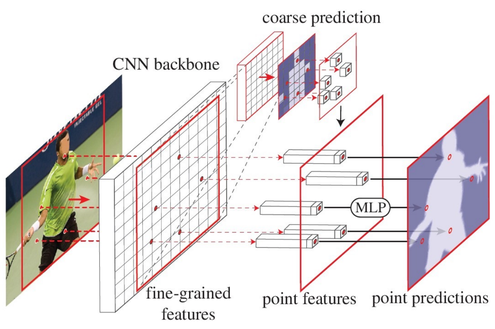
本研究整體架構上，針對instance segmentation部分，會藉由Http Request將WebRTC取得的使用者影像畫面傳入Model進行人像與背景切割處理。針對會議記錄影音優化部分，則會藉由mediaRecorder及Python urlopen() 將會議影音紀錄存放到電腦本機的資料夾中，Super Resolution會將紀錄的影像部分提取處理，Speech Denoising則是對音訊部分做處理，前兩者完成後，會做結合產生新的處理過的會議紀錄，而Speech to text也是提取音訊，並將轉換出的文字內容以文字檔同樣儲存在電腦本機資料夾中。



(系統架構圖)

- **Image Instance Segmentation**

我們利用PointRend model將webRTC拍攝回傳的即時影像分割出人像與背景，由ResNet-101 + FPN提取影像的feature map，經過RPN提出RoIs，再利用RoIAlign layer調整每個RoI的大小，最後經過mask branch得到影像中的object masks，切割出的人像與模糊處理後的背景疊加在一起，達到背景模糊/更換背景的功能，保護與會者的隱私。



(Model Mask branch架構)

- **Super Resolution**

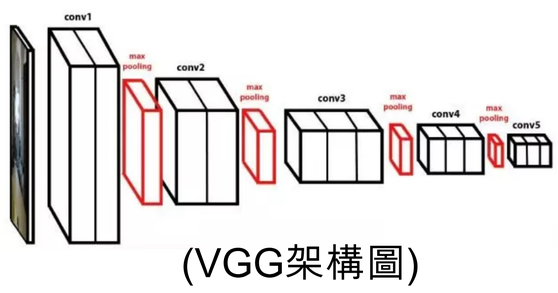
利用Real-ESRGAN模型實作提高影像解析度的功能，Generator network架構主要由數層convolutional layers、activation layers(PReLU)組成，並結合Dense Block的架構；Discriminator由U-net結合 Spectral normalization layer，以每個像素的角度判別真偽，並透過second-order degradation model生成更複雜的data以訓練模型，達到更好的成效。

(Generator)

(Discriminator)

- **Speech to text**

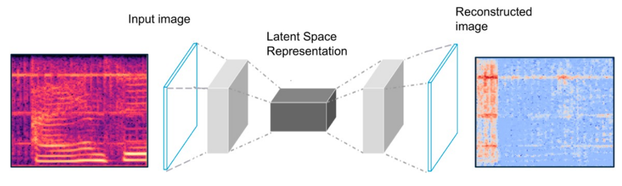
本系統主要為中文Speech to text處理，使用Google API完成。另外，比較了使用VGG model進行處理的表現差異。此架構首先會把input的audio轉換為spectrogram形式，並且使用訓練圖形資料時，常見的VGG進行訓練。Model架構有十層左右的convolution layers，loss function使用CTC (Connectionist Temporal Classification)。



(VGG架構圖)

- **Speech Denoising**

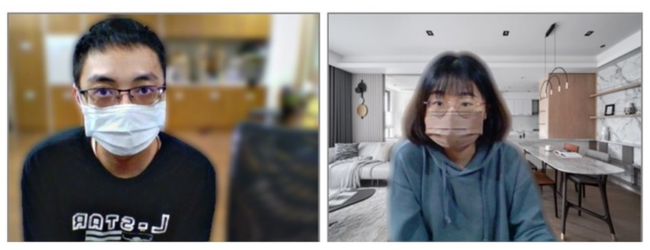
將音訊的spectrogram提取出來，比較U-net與DSTN-LSTM兩種model的降噪效果。由於U-net整體表現上較穩定，因此系統主要以U-net為降噪應用的Model。model有10層convolution layers，使用了”Huber loss”，可以根據不同狀況，選用L1 loss或L2 loss進行weight update。



(U-net架構圖)

研究結果

Instance Segmentation



(背景虛化結果)　(置換背景結果)

Speech to Text

(音訊檔)

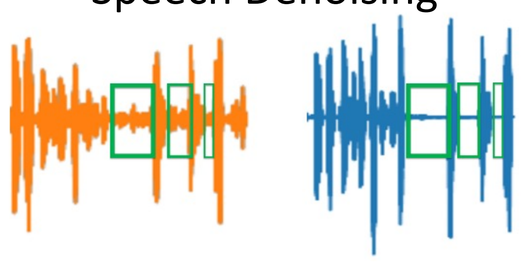
Recognized text:
測試錄音測試錄音測試測試

(轉換文字結果)

Super Resolution

(低解析度圖)　(增強影像解析度)

Speech Denoising



(原始音訊)　(降低車輛噪音)