基於深度學習之唇形辨識應用與探討

(Lip Recognition Application and Exploration Based on Deep Learning)

組員：鄭惠銘、柯沛升、陳楷翰

專題動機與理念

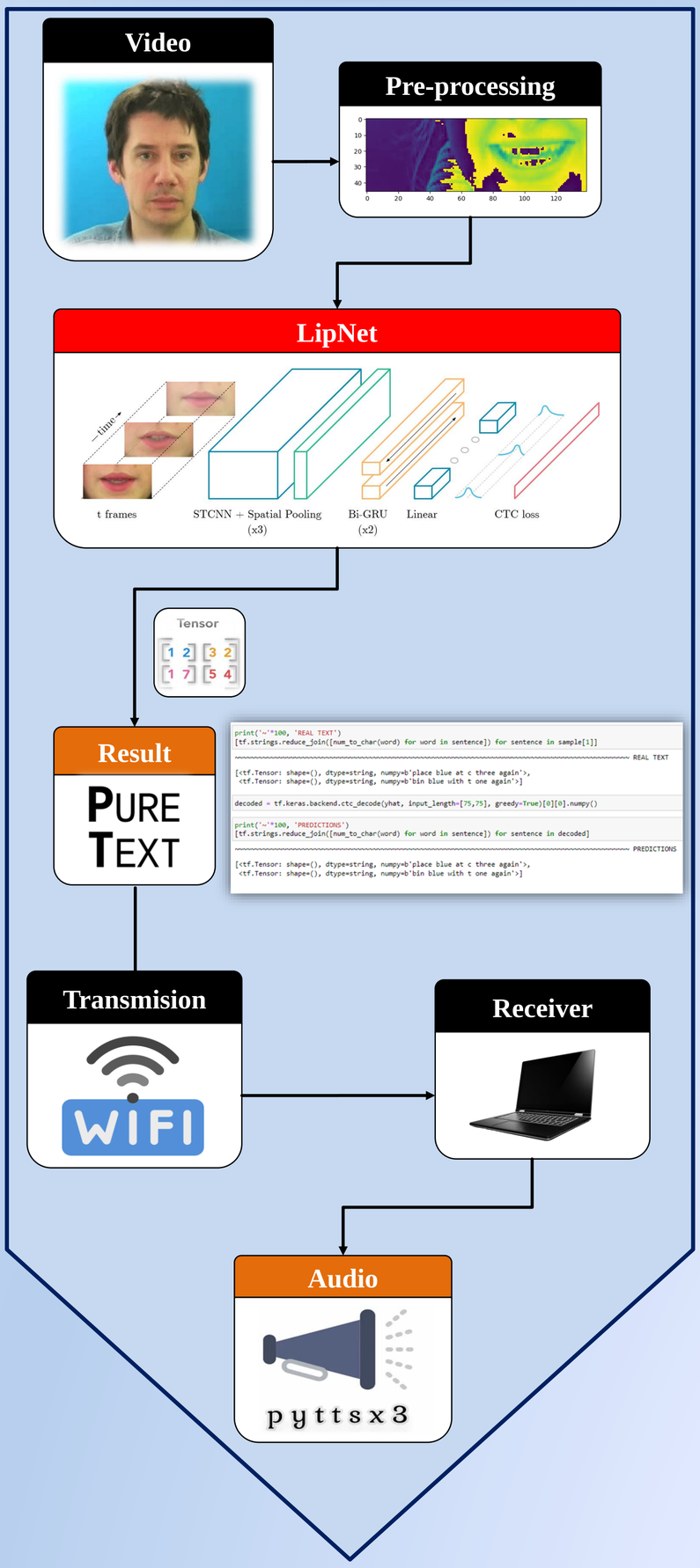
唇讀，語者無須出聲，即可由其唇部的運動來傳達其意圖。唇讀對日常生活具相當大的貢獻與影響，如某些**不便說話的場合**，或是因**老弱病殘而難以成聲**，抑或是社會案件中的**犯罪對話影像解密**，唇讀都能一一對症下藥。

在至今大學生涯中，我們鮮少接觸深度學習，甚至尚無深度學習課程，然而如今AI熱潮卻風靡全球。為了與世界接軌，更考量到未來求學生涯與職涯所需，我們打算**藉由專題實作接觸AI領域**，希望能一探深度學習的真貌。

專題摘要

我們根據**LipNet**論文建構唇讀模型，並針對**GRID**中的影片進行測試與訓練。我們將影像輸入至**三層STCNN**的3D conv、**3D maximum pooling**中進行唇部特徵提取，再將特徵向量輸入到**兩層Bi-GRU**當中轉換為tensor，以減少參數及避免梯度消失，最後將tensor輸入到**CTC loss function**當中，讓model自動將視覺特徵與文字訊息對齊。經LipNet運算過後，將可得到一串**文字訊息**(即語者所述語句)。最終，我們將取得的文字訊息，以UDP協議，經由wifi傳輸至遠方電腦，並將之轉換為語音播出，如此即完成**唇讀至發聲**之任務。

實現流程



實現原理

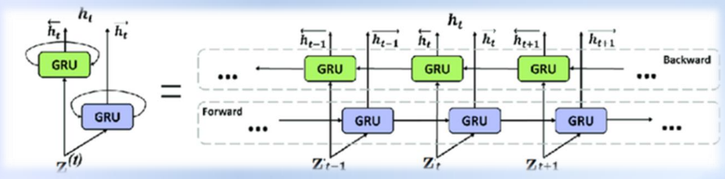
●藉由STCNN，將影片中的每一幀圖像分別轉換為**特徵向量**

$$[\text{stocnv}(\mathbf{x},\mathbf{w})_{c'tij}] = \sum_{c=0}^{C}\sum_{t'=1}^{k_t}\sum_{i'=1}^{k_w}\sum_{j'=1}^{k_h} w_{c'ct'i'j'}\, x_{c,\,t+t',\,i+i',\,j+j'}$$

●將特徵向量輸入Bi-GRU轉換為tensor

(⊙ 為element-wise multiplication；$\mathbf{z}$ 為STCNN之輸出)

$$[\boldsymbol{u}_t, \boldsymbol{r}_t]^T = sigm(\boldsymbol{W}_z\boldsymbol{z}_t + \boldsymbol{W}_h\boldsymbol{h}_{t-1} + \boldsymbol{b}_g)$$
$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{U}_z\boldsymbol{z}_t + \boldsymbol{U}_h(\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) + \boldsymbol{b}_h)$$
$$\boldsymbol{h}_t = (\mathbf{1} - \boldsymbol{u}_t) \odot \boldsymbol{h}_{t-1} + \boldsymbol{u}_t \odot \tilde{\boldsymbol{h}}_t$$



●使用CTC loss 作為模型訓練的損失函數

$$p(y|\boldsymbol{x}) = \sum_{u \in \boldsymbol{B}^{-1}(y) s.t. |u|=T} p(u_1, \ldots, uT|\boldsymbol{x})$$, where T is the number of time-steps in the sequence model.

定義與分析

●Learning Rate Definition

```
def scheduler(epoch, lr):
    if epoch < 30:
        return lr
    else:
        return lr * tf.math.exp(-0.1)
```

●Accuracy Definition

$$\textbf{Accuracy} = \frac{\text{預測正確語句數}}{\text{總預測語句數}}$$

●模型準確率

```
真實語句： lay red with f three again
預測語句： lay red with three again
1/1 [==============================] - 4s 4s/step
真實語句： place blue at v five again
預測語句： place blue at v five again
1/1 [==============================] - 4s 4s/step
真實語句： set sp blue with b three soon
預測語句： set s blue with b three soon
1/1 [==============================] - 4s 4s/step
真實語句： lay red with r eight now
預測語句： lay red with r eight now
1/1 [==============================] - 4s 4s/step
真實語句： bin red with n two please
預測語句： bin red with two please
1/1 [==============================] - 3s 3s/step
真實語句： place blue with d zero please
預測語句： place blue with d zero please
1/1 [==============================] - 3s 3s/step
真實語句： bin white with n nine again
預測語句： bin white with n nine again
模型準確性： 92.0 %
```

➡ 模型準確率最高可達92%

神經網路參數

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv3d (Conv3D)             (None, 75, 46, 140, 128)  3584

 activation (Activation)     (None, 75, 46, 140, 128)  0

 max_pooling3d (MaxPooling3D  (None, 75, 23, 70, 128)  0
 )

 conv3d_1 (Conv3D)           (None, 75, 23, 70, 256)   884992

 activation_1 (Activation)   (None, 75, 23, 70, 256)   0

 max_pooling3d_1 (MaxPooling  (None, 75, 11, 35, 256)  0
 3D)

 conv3d_2 (Conv3D)           (None, 75, 11, 35, 75)    518475

 activation_2 (Activation)   (None, 75, 11, 35, 75)    0

 max_pooling3d_2 (MaxPooling  (None, 75, 5, 17, 75)    0
 3D)

 time_distributed (TimeDistr  (None, 75, 6375)         0
 ibuted)

 bidirectional (Bidirectiona  (None, 75, 256)          6660096
 l)

 dropout (Dropout)           (None, 75, 256)           0

 bidirectional_1 (Bidirectio  (None, 75, 256)          394240
 nal)

 dropout_1 (Dropout)         (None, 75, 256)           0

 dense (Dense)               (None, 75, 41)            10537

=================================================================
Total params: 8,471,924
Trainable params: 8,471,924
Non-trainable params: 0
```

結論與未來展望

透過LipNet 的實現，我們取得了**92%的準確度**，由於交談中未必要所有字詞皆正確也能使人明白其意，因此該準確度應可視為具實用價值。然而，由於**僅用一位男性訓練**LipNet模型，且為用**slicing function**進行唇部區域選取，故而不足亦無法處理隨手拍攝之影片。此外，在處理影片到遠端播放語音所需的時間近10秒之久，因此在**延遲**方面的表現不盡理想。在未來，我們希望能進一步訓練模型、以**YOLO**自動抓取唇部區域、**優化影片預處理程序**，再改以**5G傳輸**，將前述之缺陷盡可能臻至完美。

透過該專題，我們**初步了解深度學習知識與實現**，更藉由該專題得到了參與Synopsys 競賽的寶貴經驗，可謂收穫甚滿。實在非常感謝教授的指導與其他學長的幫助!