Day 22

深度學習與電腦視覺

學習馬拉松

cupoy 陪跑專家：楊哲寧

深度學習理論與實作

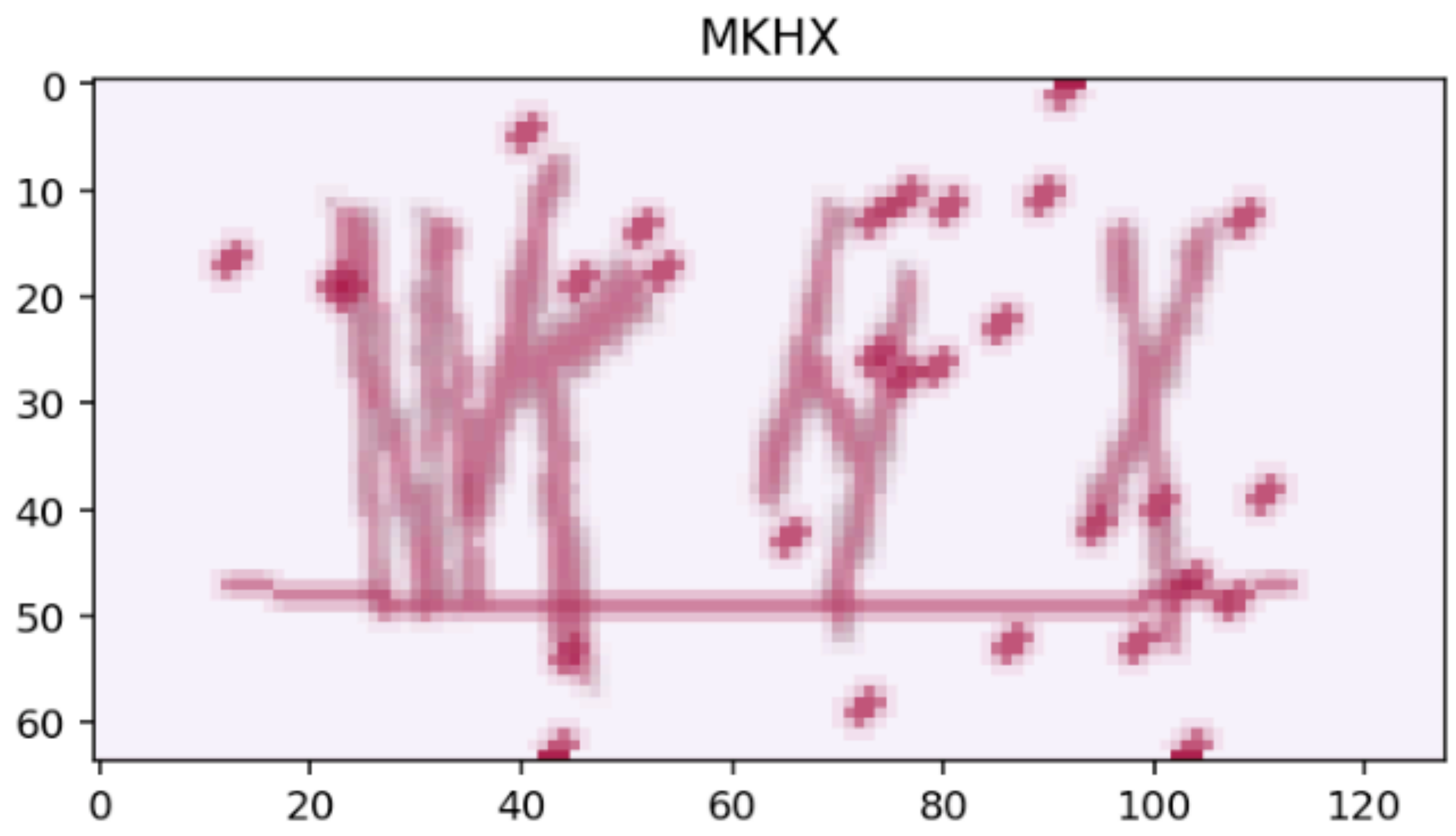
驗證碼識別

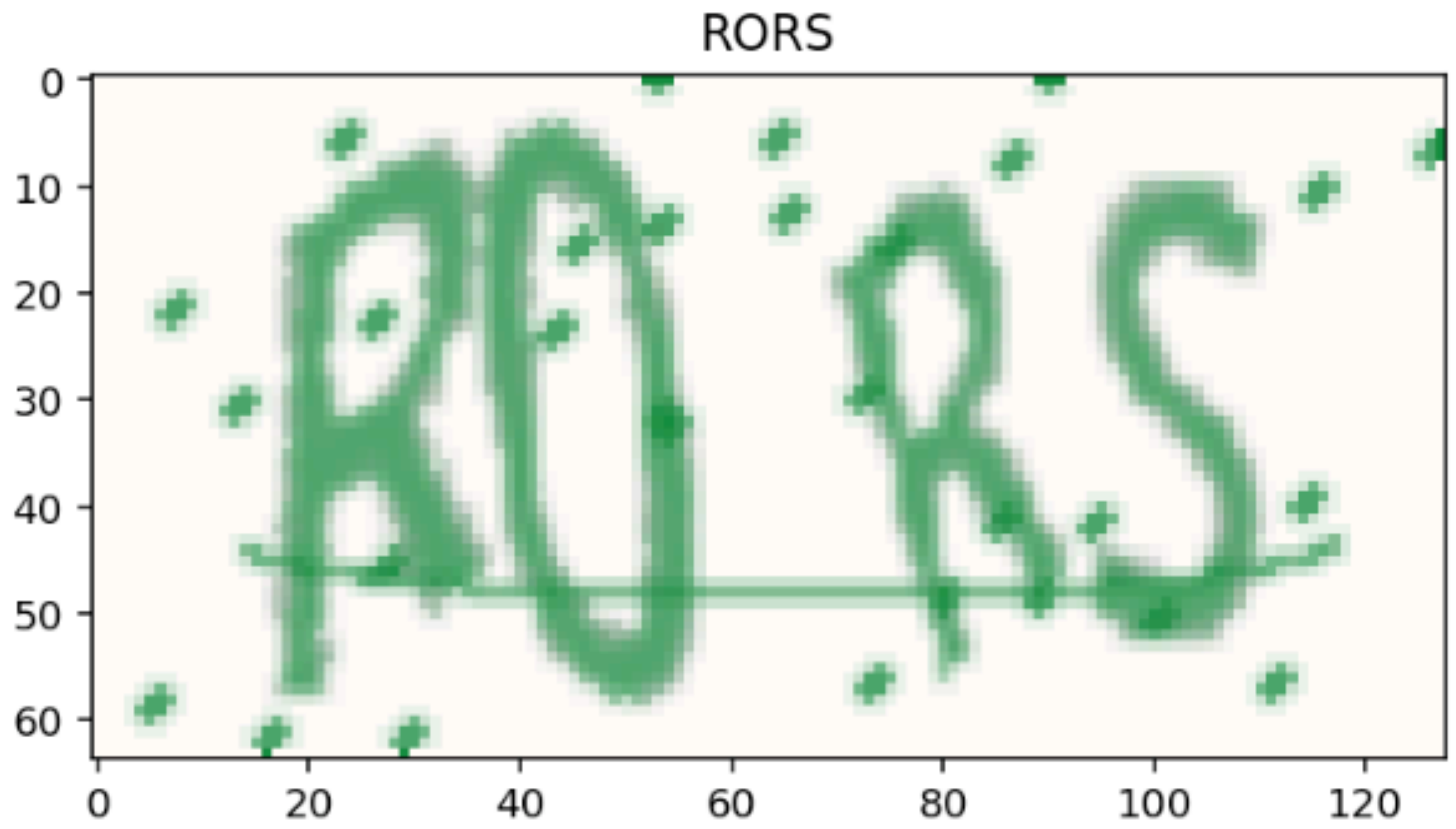
重要知識點

- 了解如何應用 CNN 在驗證碼識別
- 有興趣的學員們可以進一步了解CRNN 這篇論文與 CTC 損失函數

CNN驗證碼識別

想必各位學員到這個階段，對 CNN 都有了一定的了解，今天就讓我們利用 CNN 來實作驗證碼識別，同樣的方法也可以用來做街景門牌識別、車牌識別與文字 OCR。

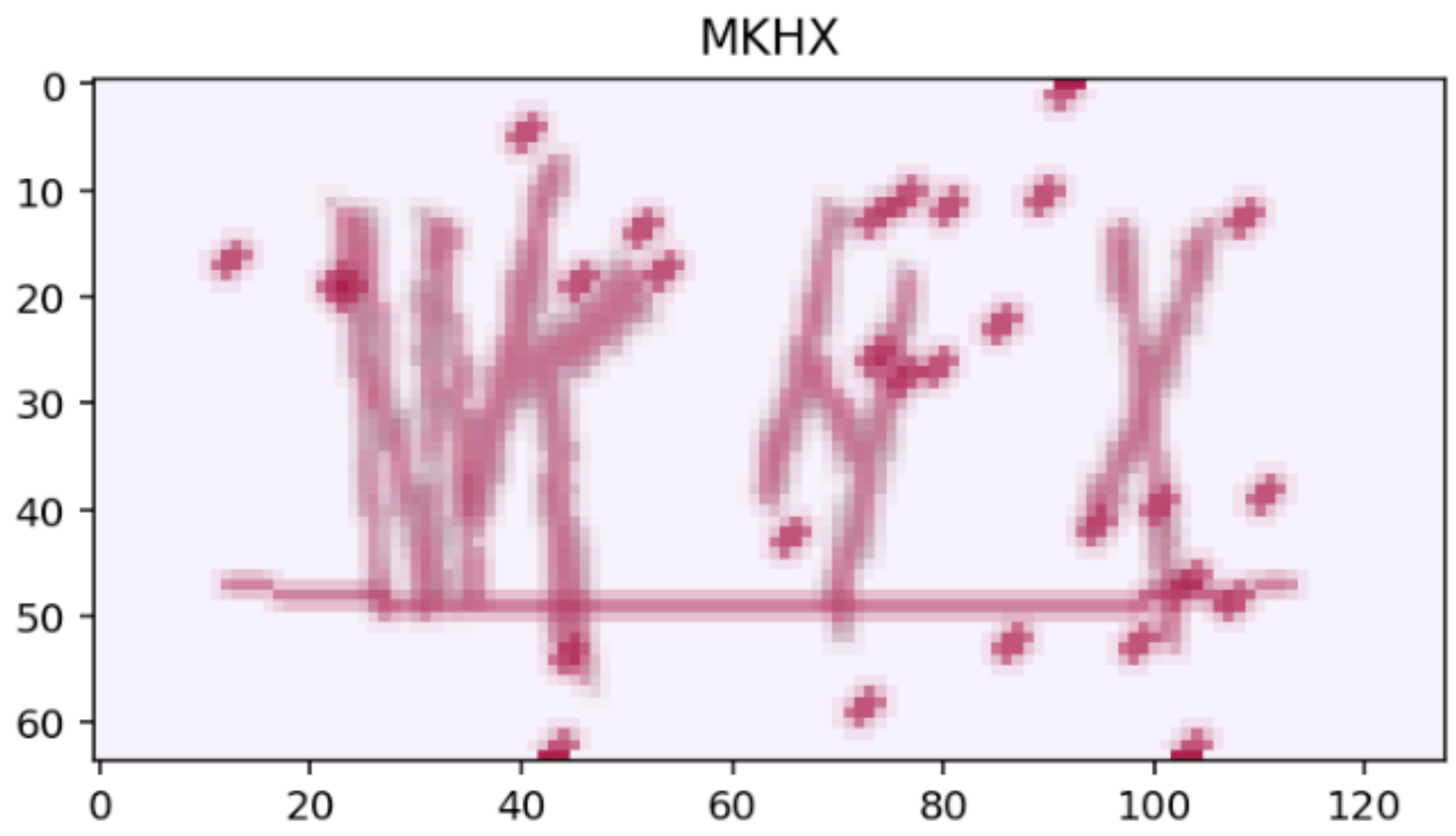


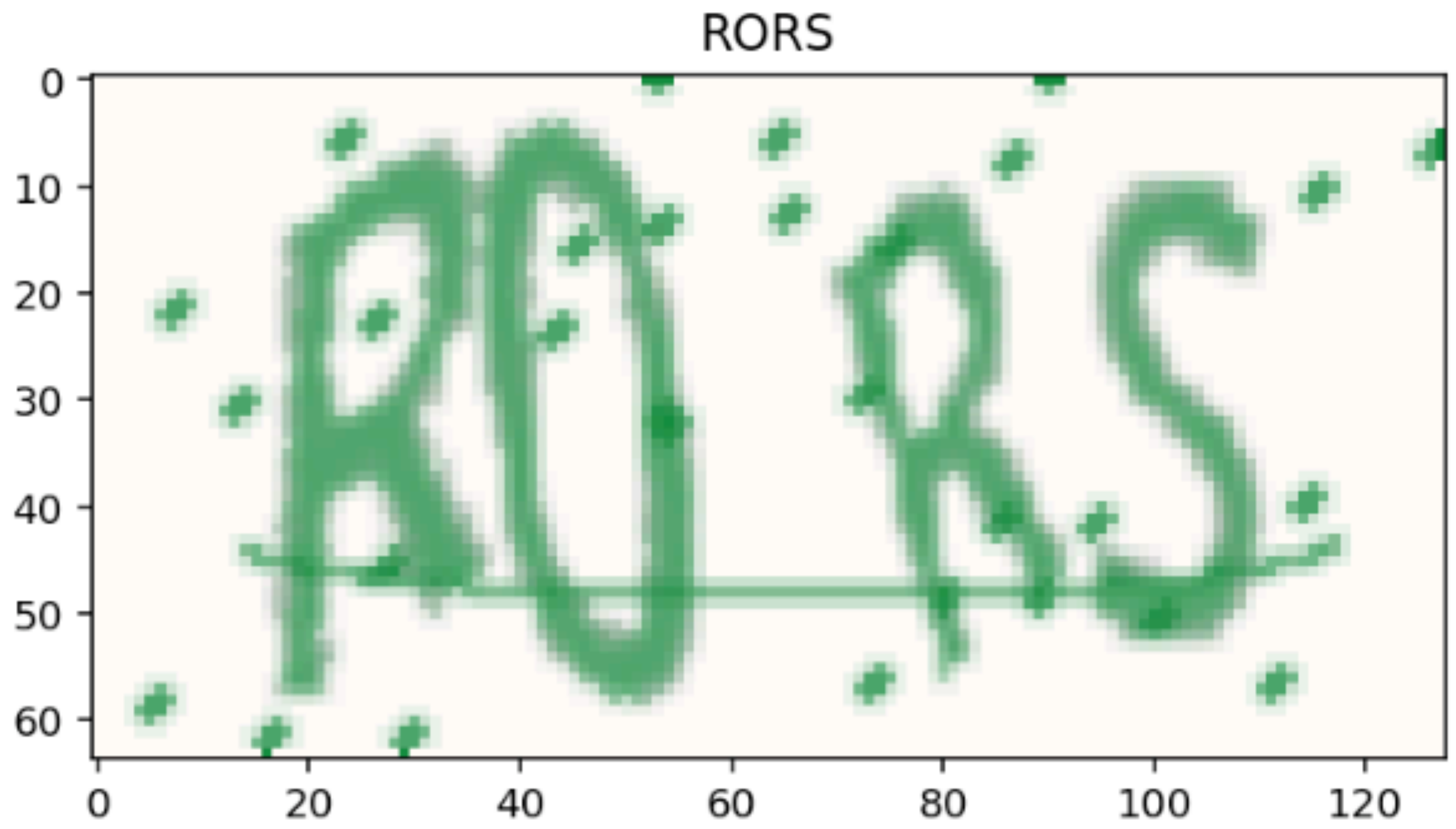


CNN驗證碼識別

一般來說，類似的問題最困難的點在於：

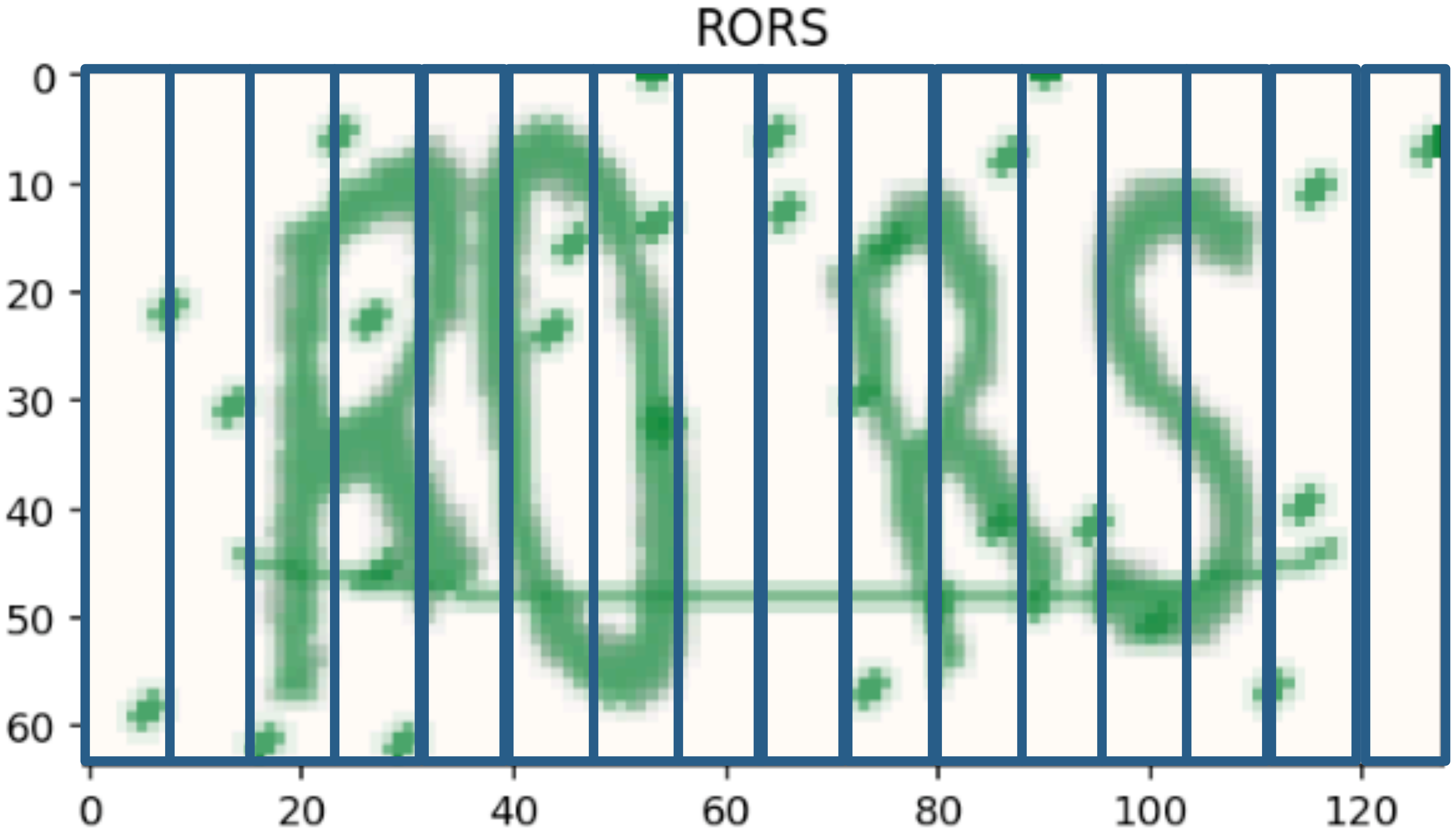
1. 字串長度不固定
2. 字與字之間寬度不同，無法切開成單一字符辨識





CNN驗證碼識別

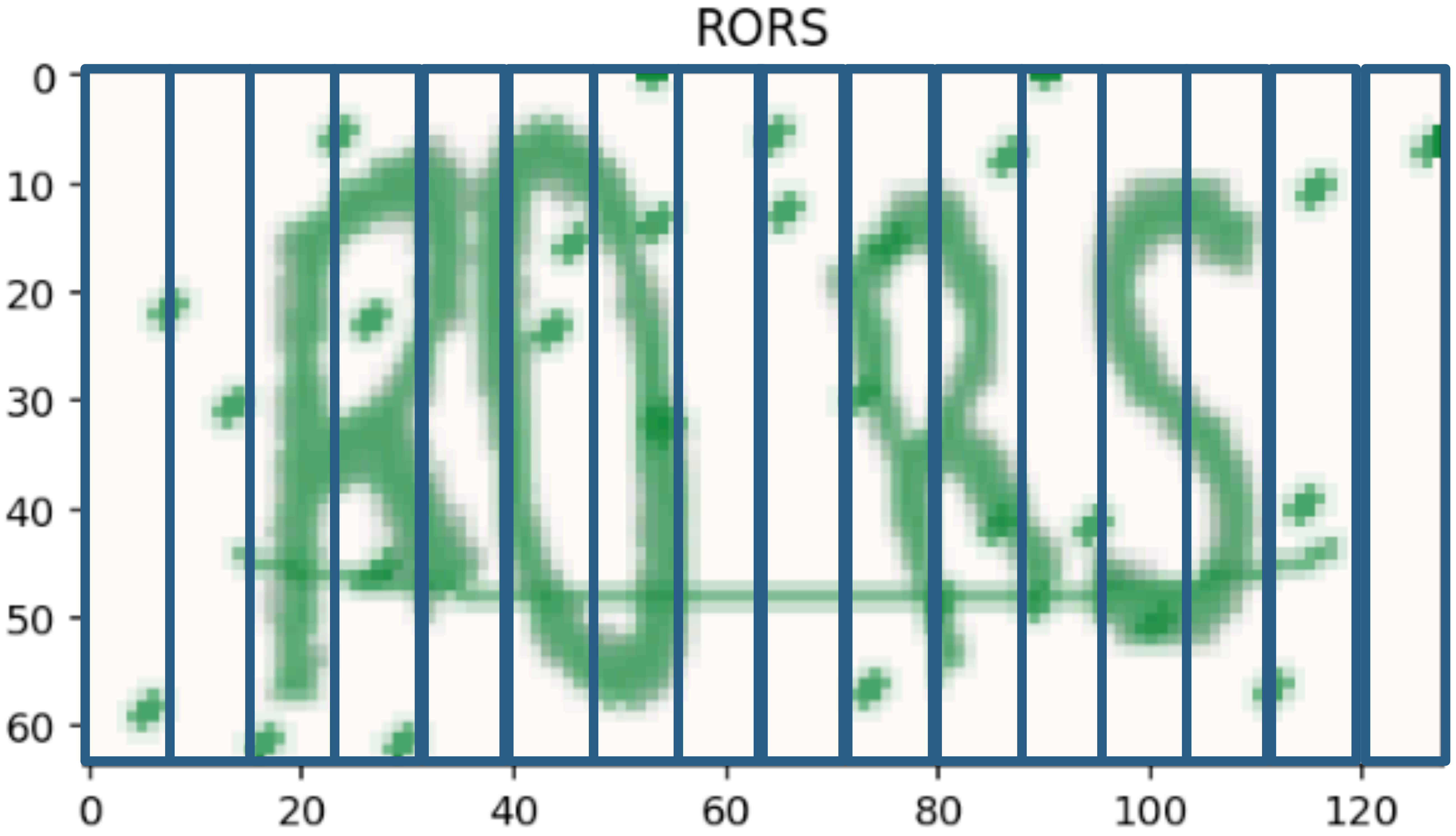
因此我們設計的理念就是，**強迫 CNN 沿著 X 軸來學習資訊**，舉個例子來說，如下方寬高為 128,64 的影像，假設輸出寬度為 16 (看CNN設計)，我們可以想像成在寬度上，**每 8 個 pixels的資訊照順序被壓再一起**，因此我們就是預測這 16 個影像區，每個影像區內應該是什麼字符。



拆成16個影像區來預測

CNN驗證碼識別

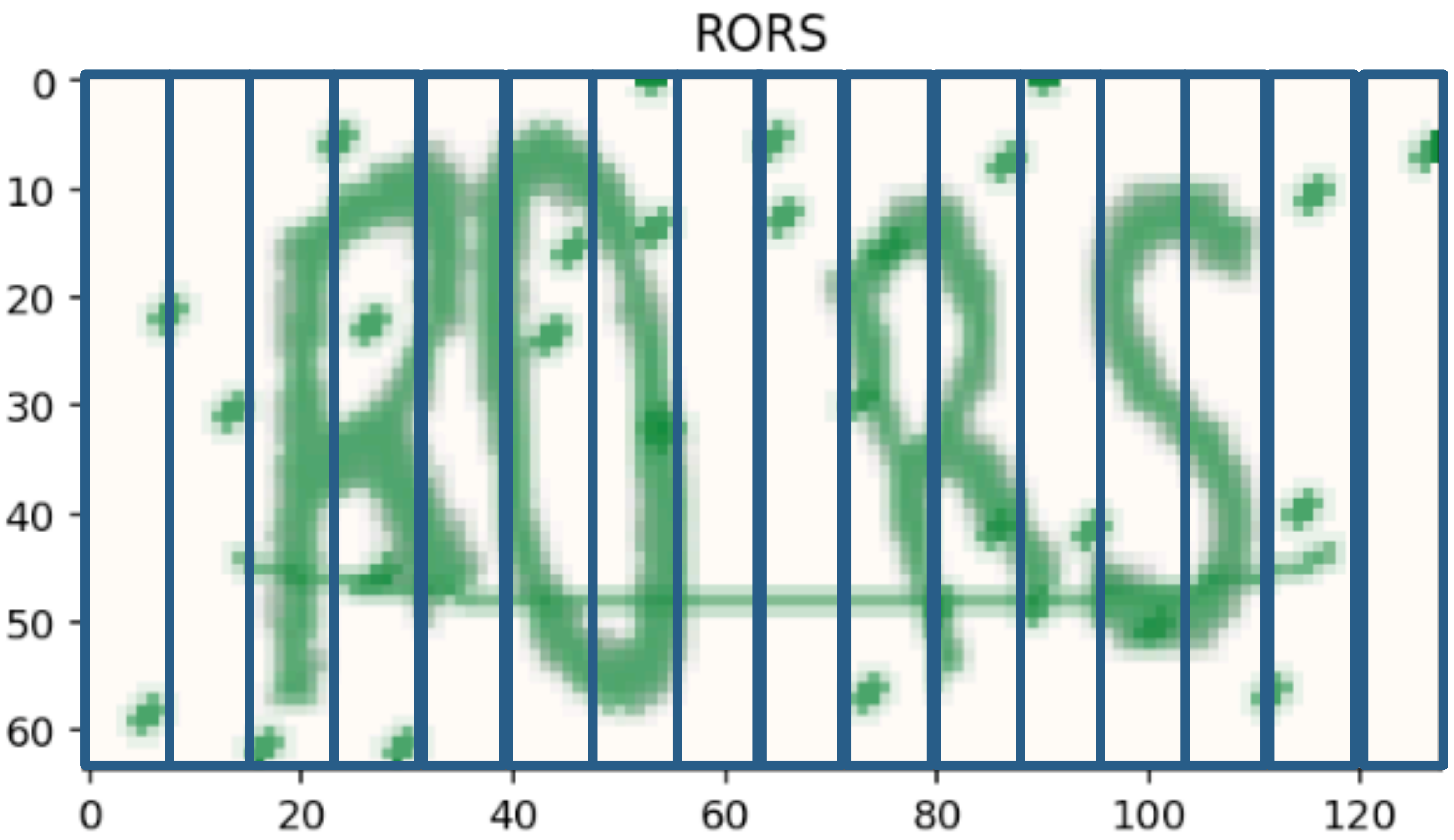
雖然各位學員可能還沒接觸時間序列模型(RNN)，但其實這類問題可以看成時間序列的問題來解，我們把 **16 個影像區看成是 16 個時間點**，將 CNN 的輸出**特徵圖**當作時間序模型(RNN)的輸入，這就構成了經典的 CRNN 模型。



拆成16個影像區來預測

CNN驗證碼識別

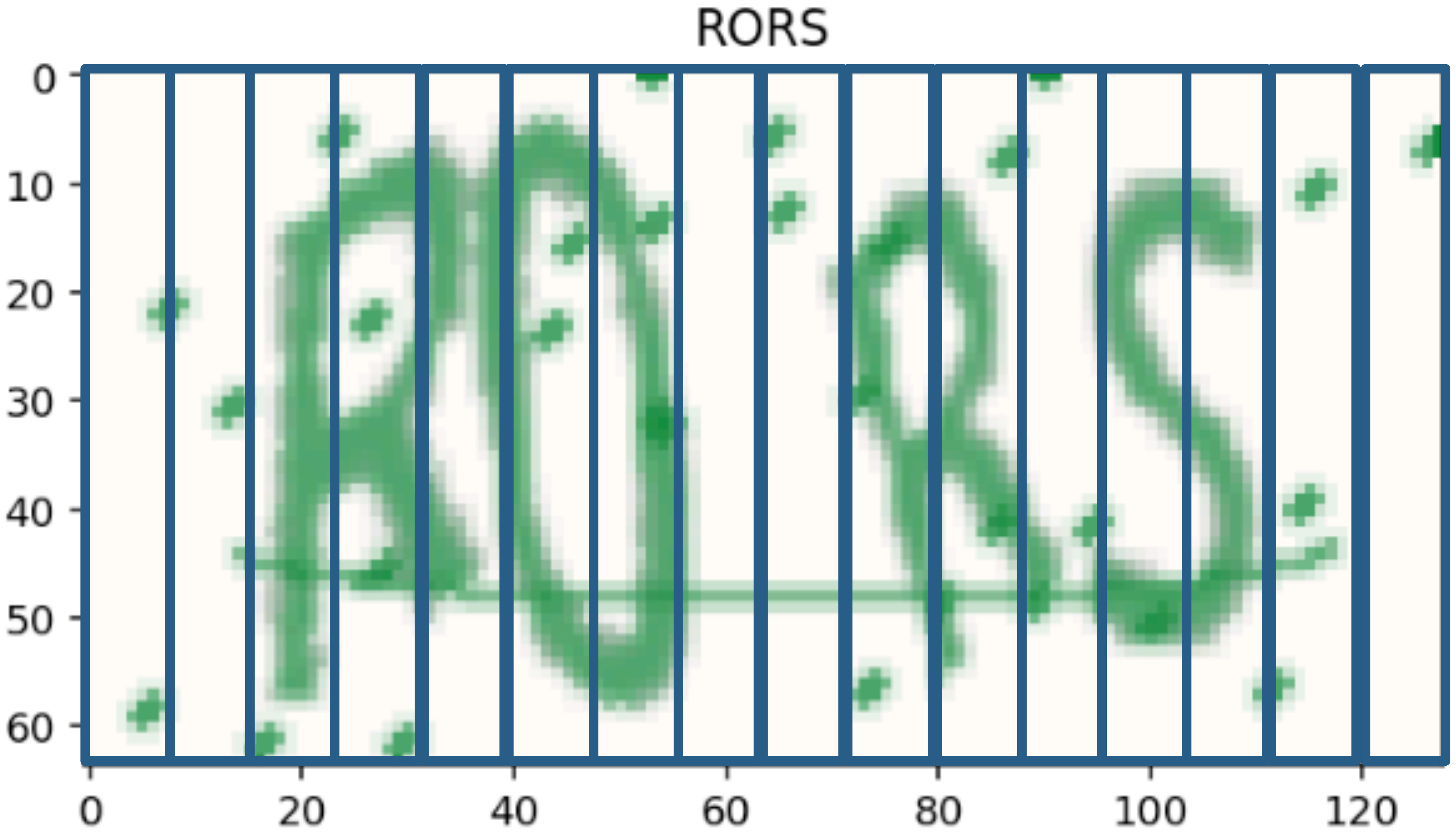
- 有興趣的學員們可參考原文: CRNN
- 本次實作重點會放在CNN模型，但在程式碼附錄中也附上了CNN+RNN代碼，各位學員可以比較看看純**CNN與CNN+RNN模型的表現**。



拆成16個影像區來預測

CTC Loss (Connectionist temporal classification)

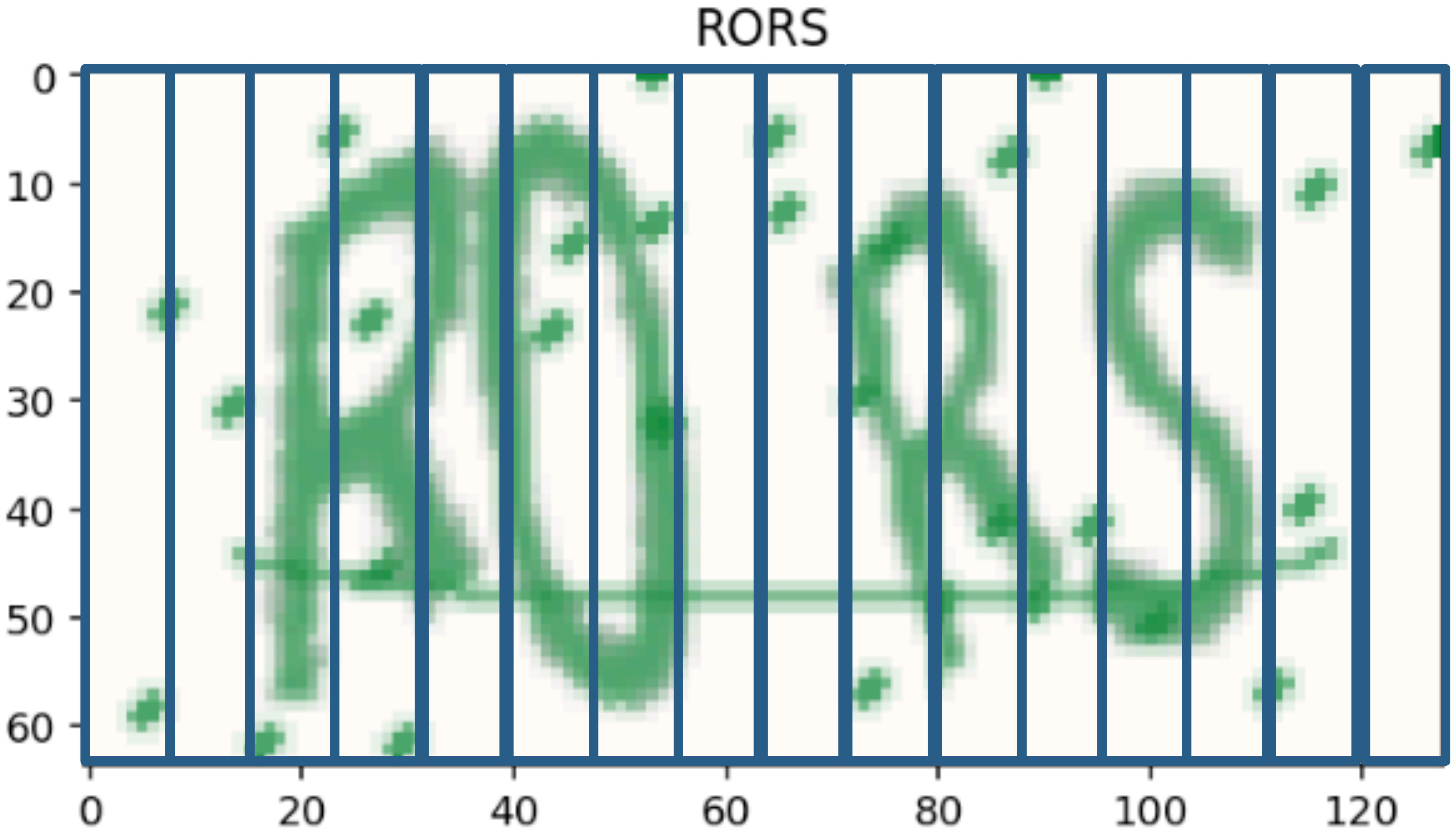
- 然而全部字串可能不到16個字，這個時候就是使用CTC Loss的好時機。
- CTC Loss一開始是用於預測語音資料，在CRNN 這篇文獻中正式被用來處理文字序列，其主要適用於『**不定長度序列**』的問題。



拆成16個影像區來預測

CTC Loss

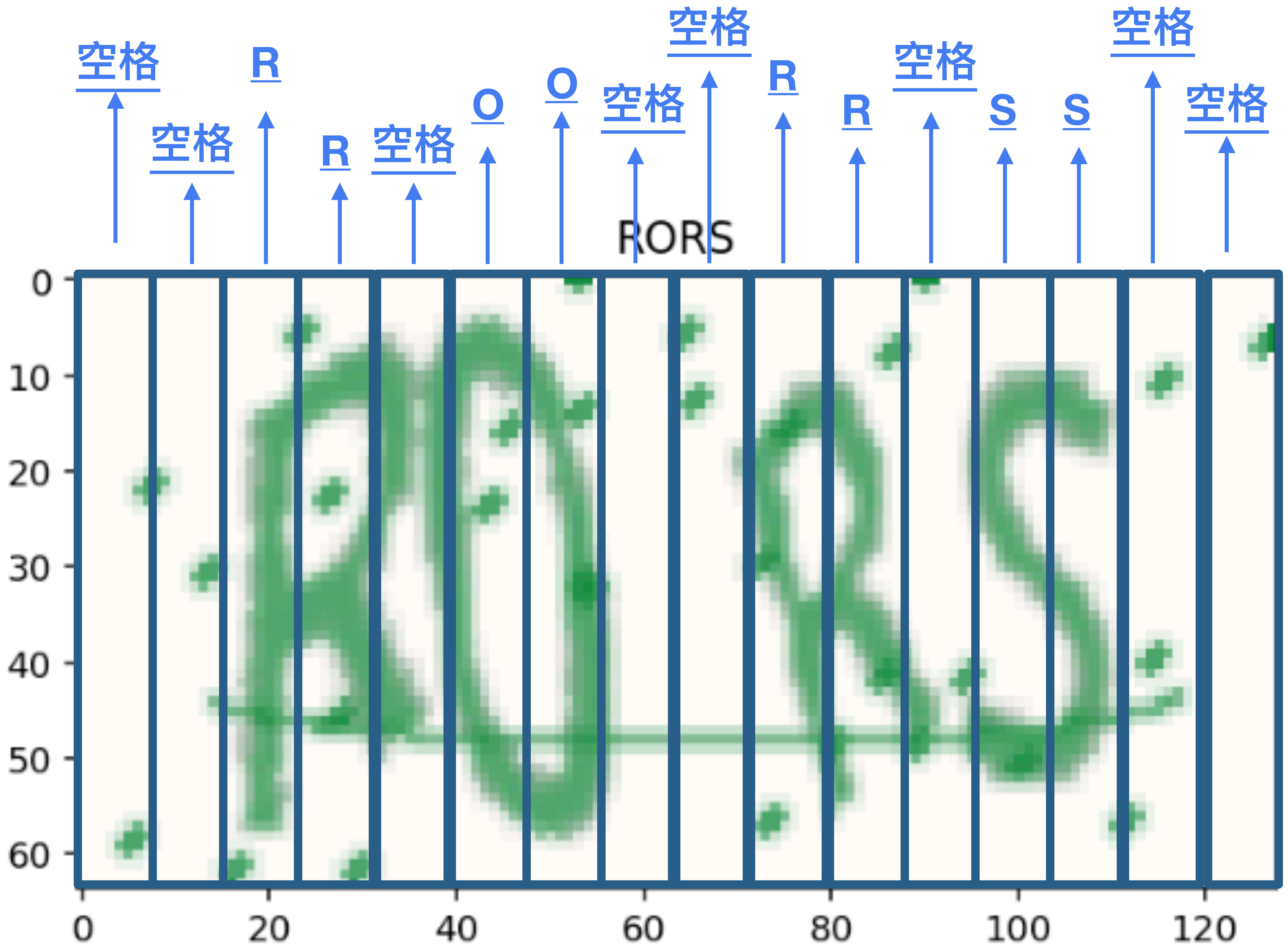
- 如語音訊息中，一段時間內的語音訊號可能包含不定長度的文字內容，又像是文字影像辨識中，相同寬度的影像也可能包含不同的字符數量。
- CTC相較於一般的損失函數針對個別案例計算誤差，其宗旨在於『**優化整串序列**』



拆成16個影像區來預測

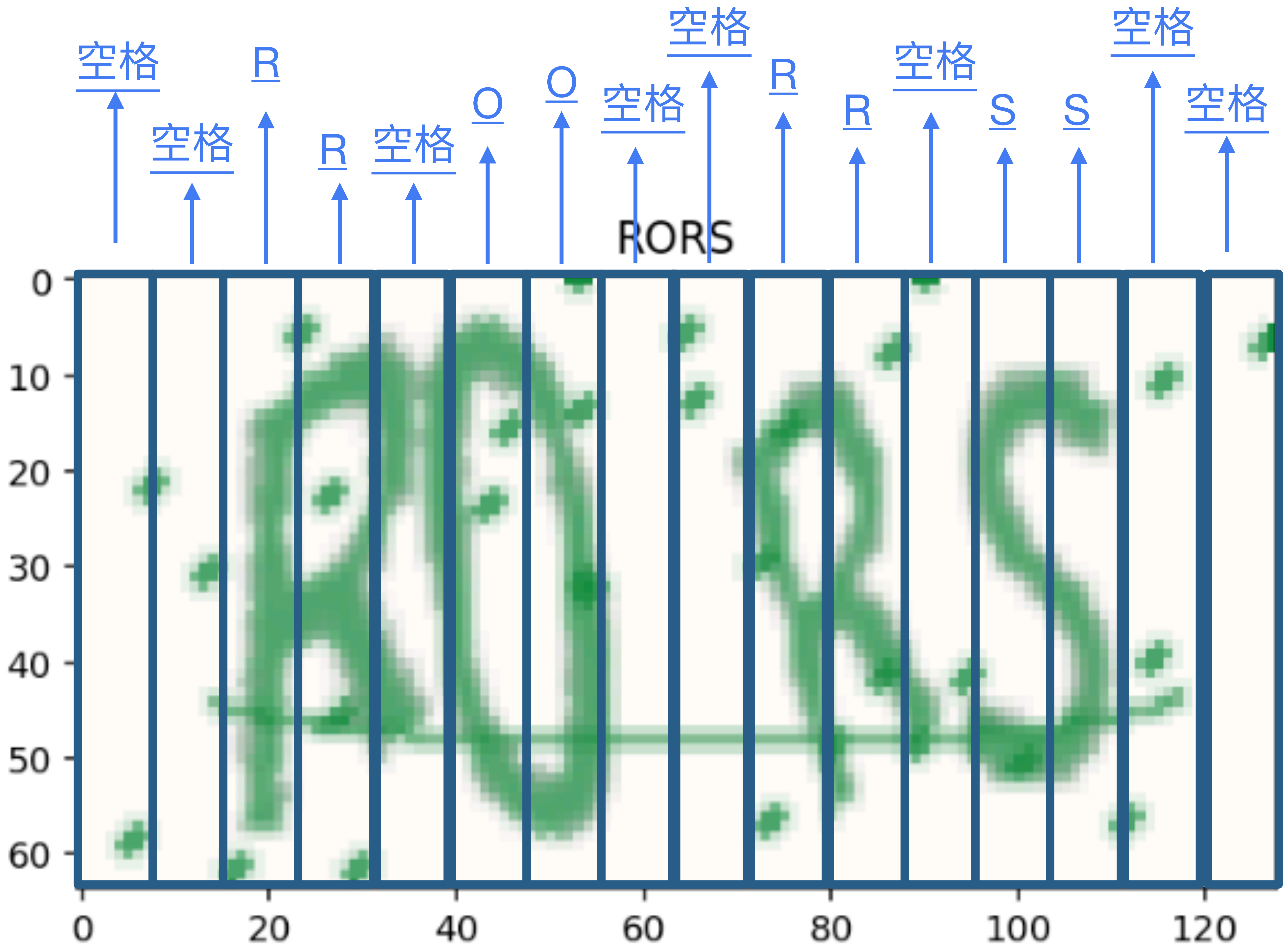
CTC Loss

- 其主要是透過『填空』來區分不同字符，CTC Loss設計的假設是，不同字符間一定有一個空格，各位學員可以觀看下方圖的示意。



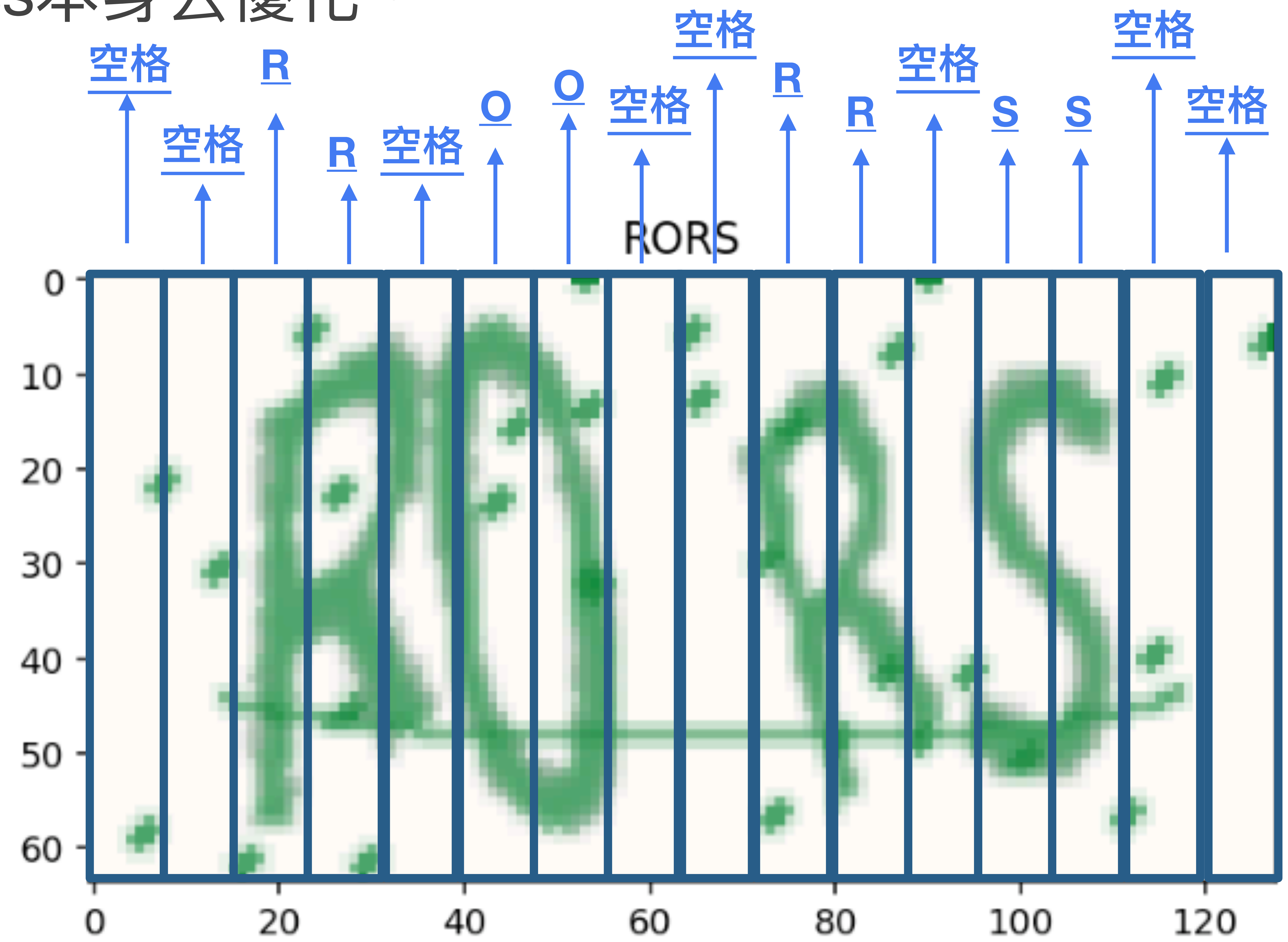
CTC Loss

- 如圖CTC Loss會建一個『空格』類別，並透過**『空格』與『重複字符』**來處理不定字串的問題。



CTC Loss

- 當**『重複字符』**之間沒有空格時，代表該區只有一個字符，因此下方真正的結果應該是RORS， **『空格』與『重複字符』**數沒有一定，主要靠CTC Loss本身去優化。



CTC Loss

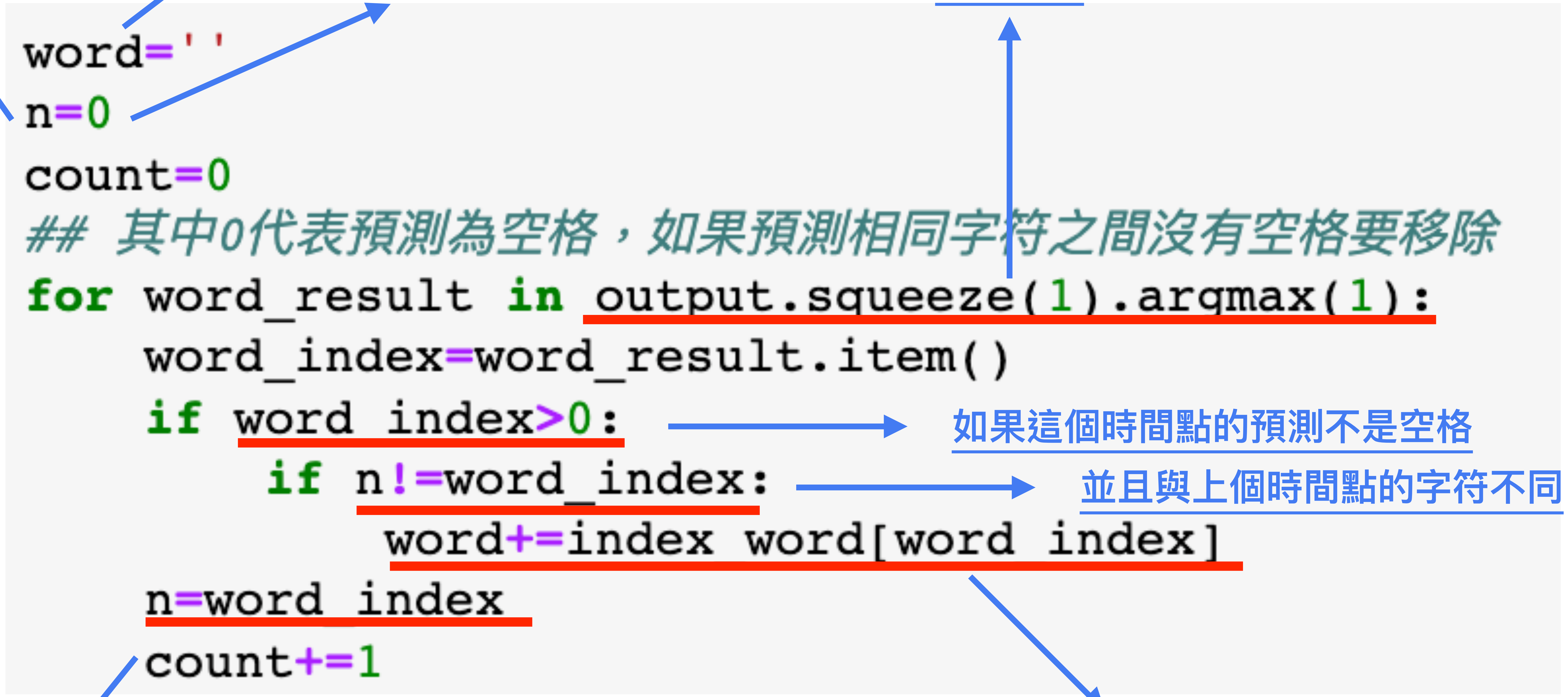
所以當我們拿到模型輸出結果時，還要進一步過濾，代碼如下：

記錄每個點的字符

初始化字串

初始是空格

預測結果



```
word=''
n=0
count=0
## 其中0代表預測為空格，如果預測相同字符之間沒有空格要移除
for word_result in output.squeeze(1).argmax(1):
    word_index=word_result.item()
    if word_index>0:
        if n!=word_index:
            word+=index_word[word_index]
    n=word_index
    count+=1
```

如果這個時間點的預測不是空格

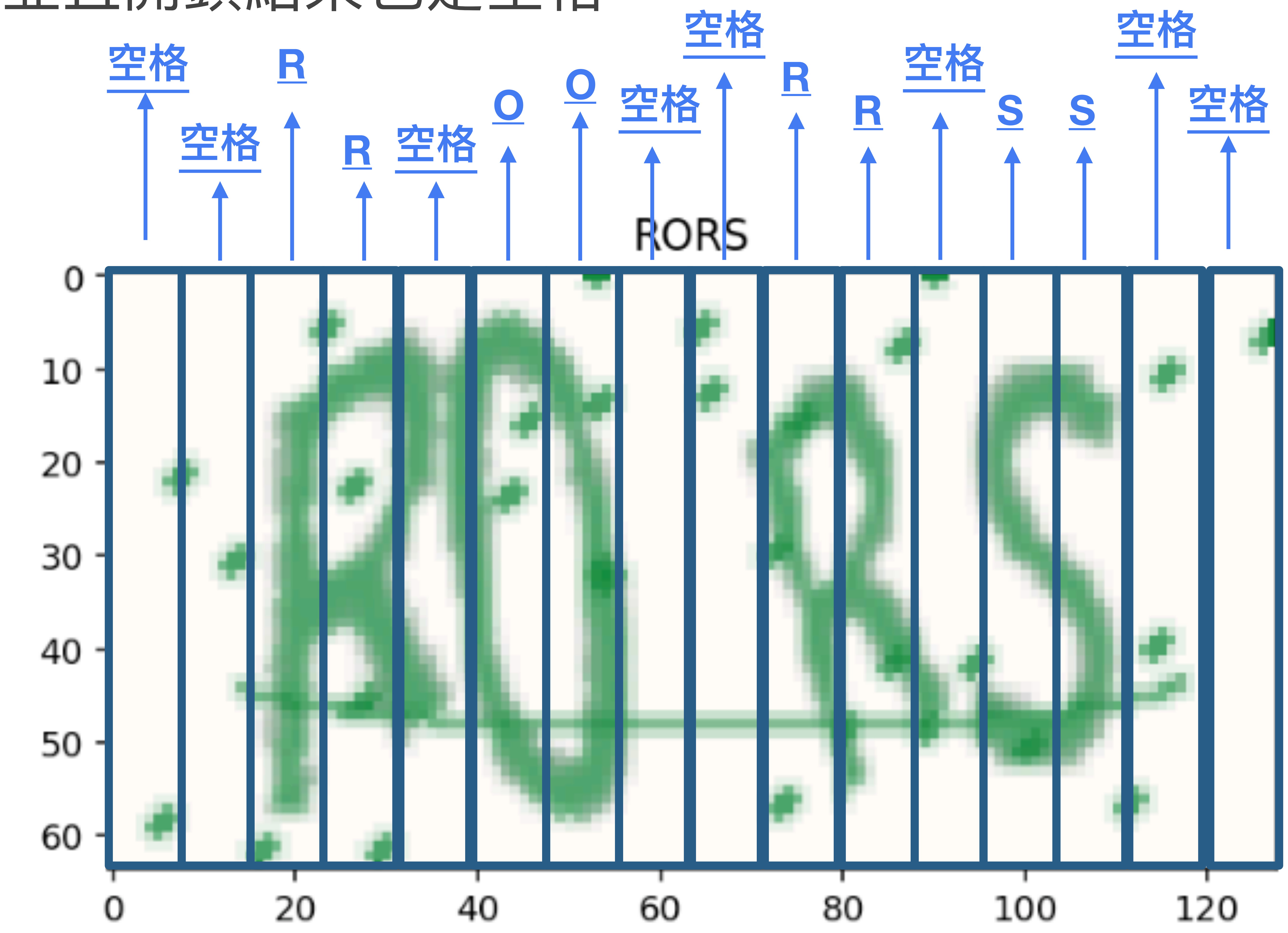
並且與上個時間點的字符不同

紀錄這個時間點的字符

我們就加入這個字符到字串中，其中Index_word是個字典。

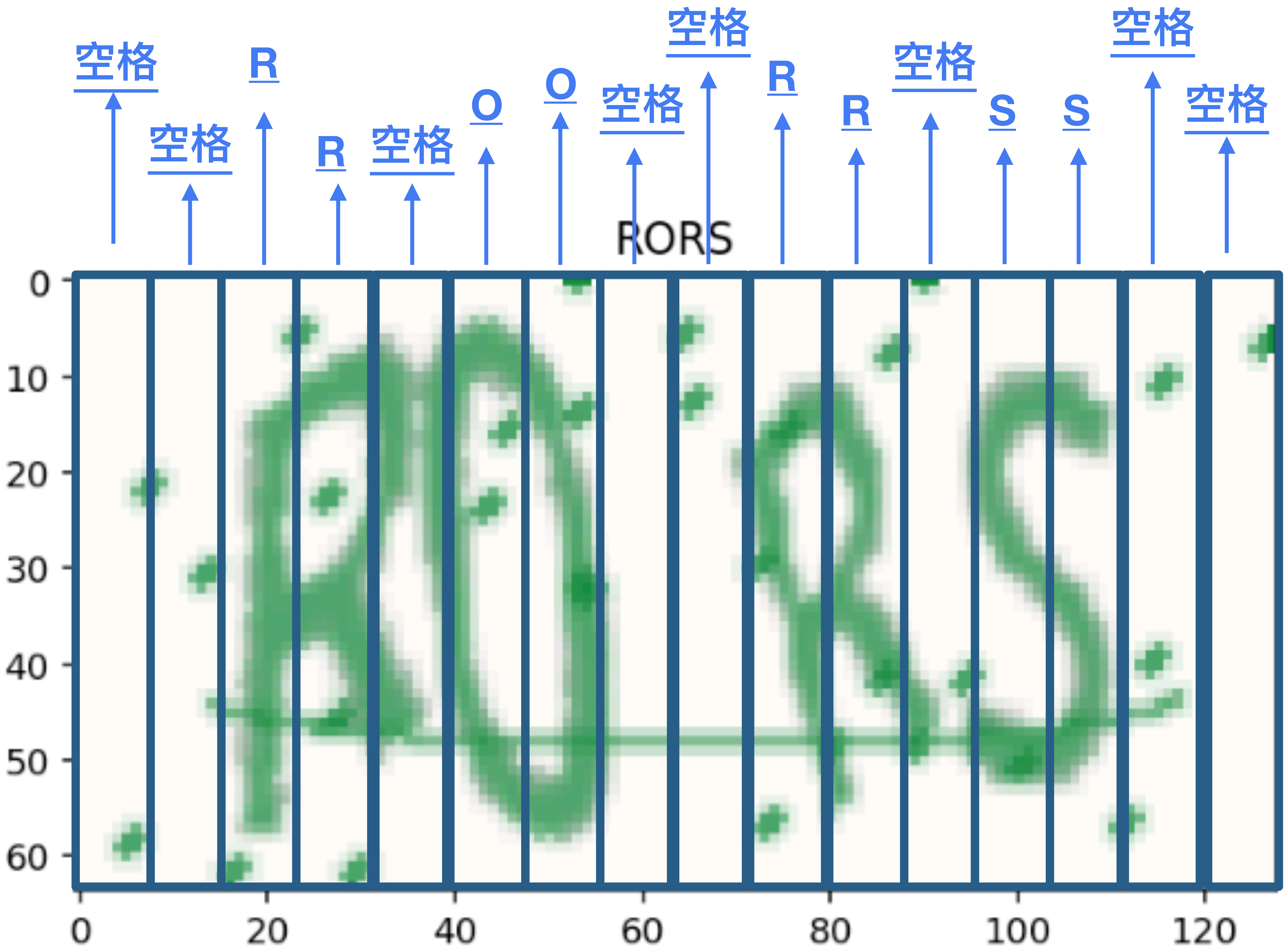
CTC Loss

- 這裡還有一個重點要強調，其實原圖切成幾格並沒有一定，但最好大於**『2倍的字串長＋1」**，這主要是由於 CTC Loss 的假設是不同字符間至少有一個空格，並且開頭結束也是空格。



CTC Loss

所以假如我們的驗證碼有長度 4,5,6，那影像過完 CNN 後寬度最好大於 **6*2+1= 13**。



CTC Loss

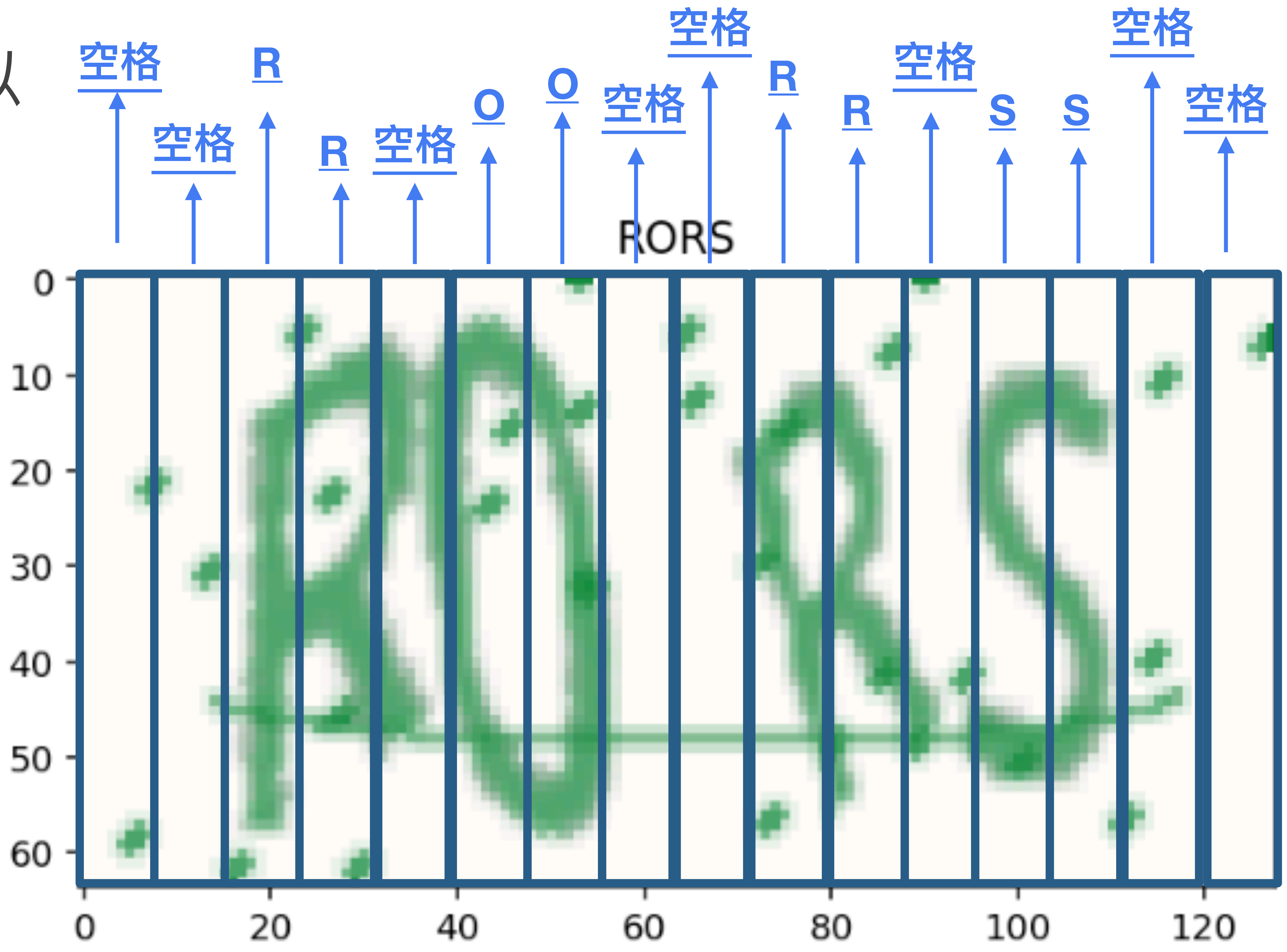
想要深入研究CTC Loss的學員們可參考以下兩個連結：

CTC Loss 原理

參考連結

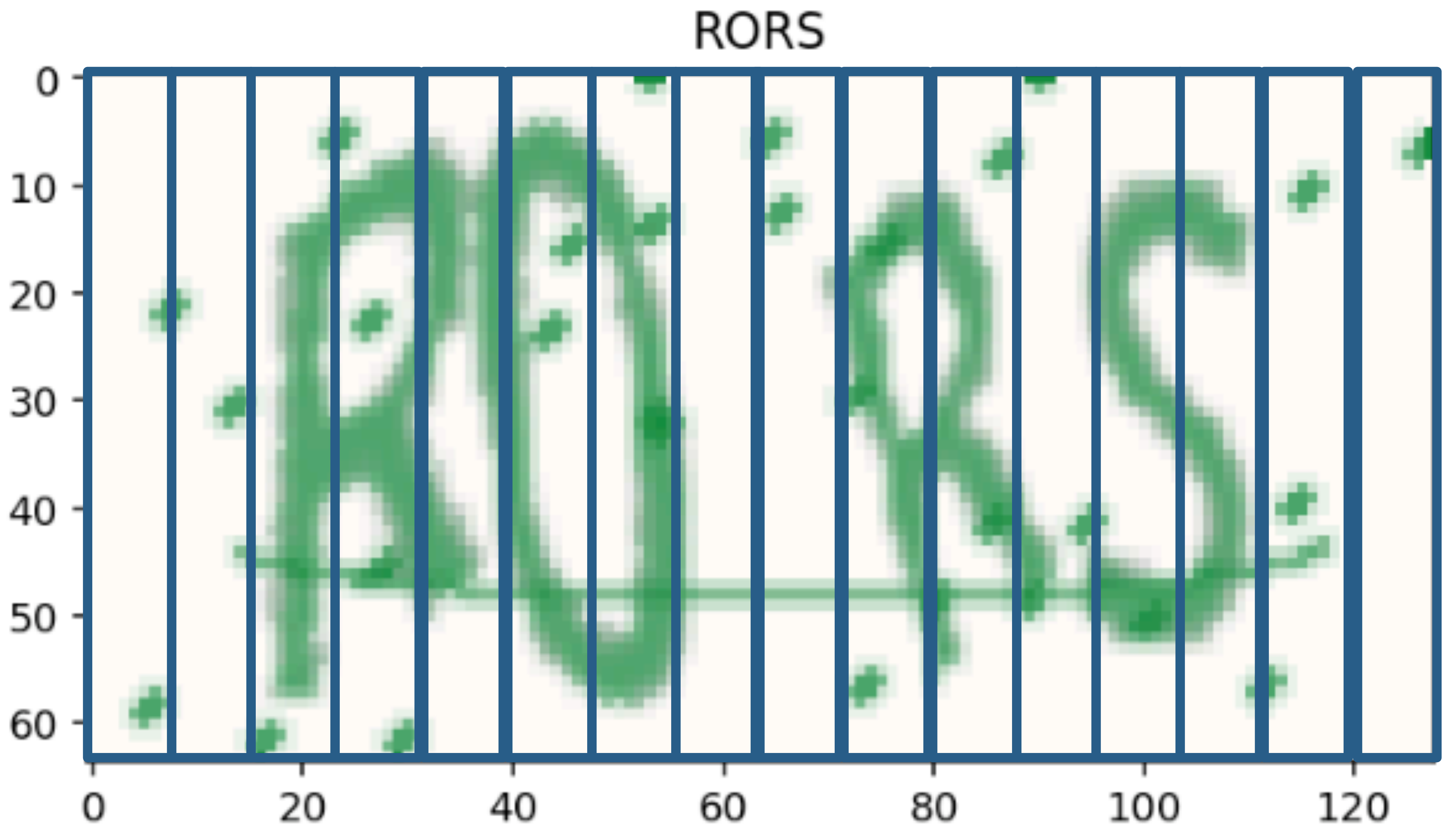
CRNN+CTC Loss 原理

參考連結



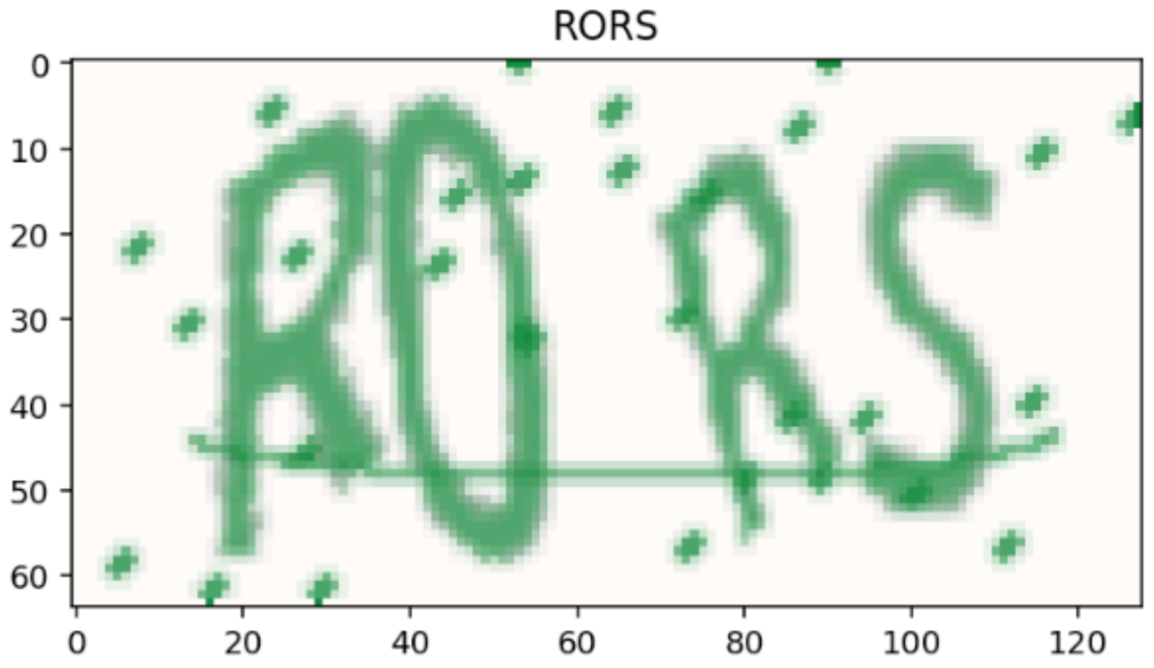
CTC Loss

- 除了上述步驟外，CRNN作者還將圖像高度壓成1，可以想像成將x軸一定範圍內的資訊都壓在一起，舉個例子來說，假如輸入影像是(Batch size,64,128,3)，其在輸出時就變為(Batch size ,1,16,channel深度)，其中16(寬度)跟channel深度都是我們可以控制的。



CTC Loss

有了概念後，大家就可以了解這次模型的主要架構：

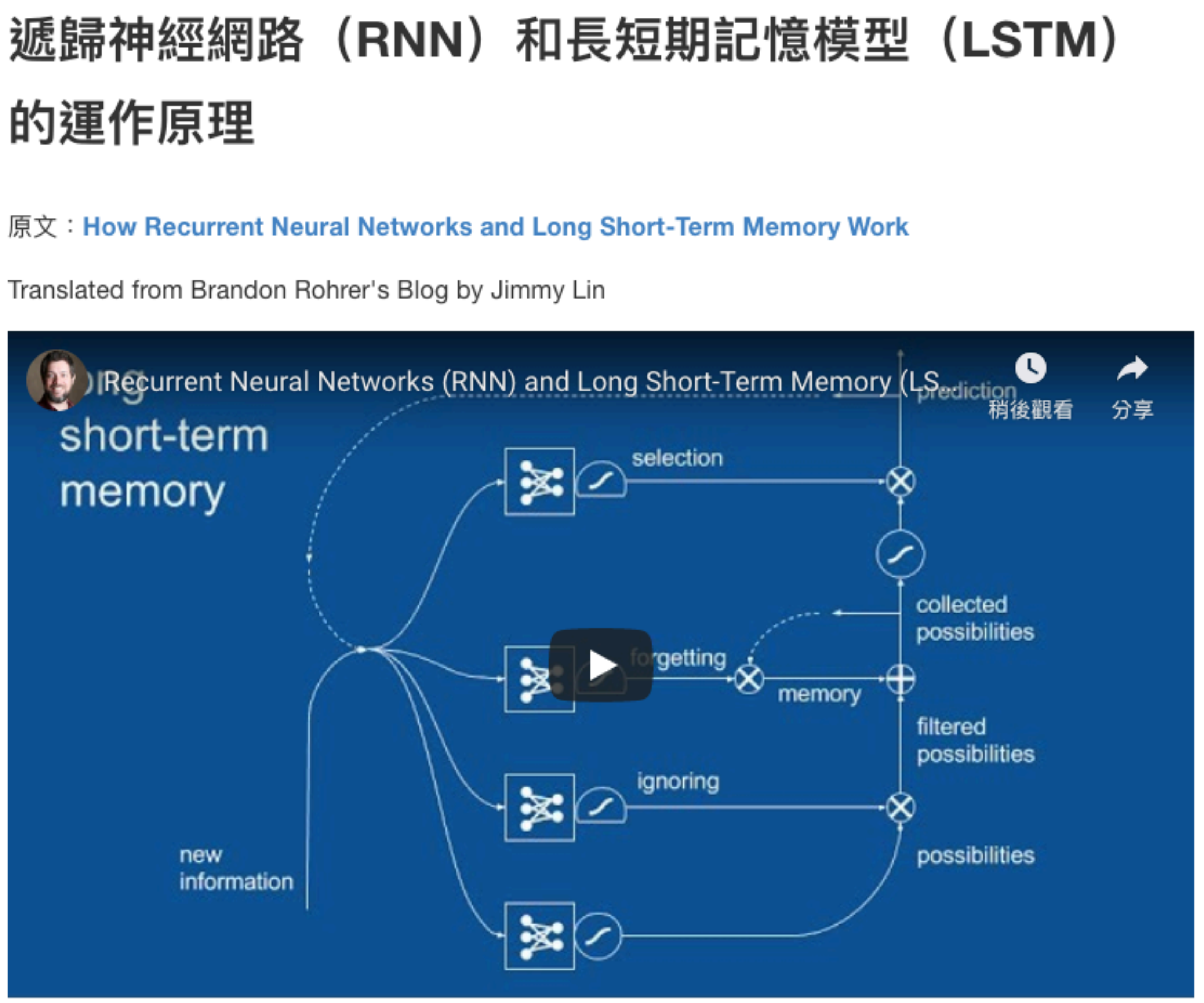


任何CNN模型，將高度壓成1，輸出寬度大於最長字串*2+1

RNN (選擇性)

CTC Loss

推薦延伸閱讀



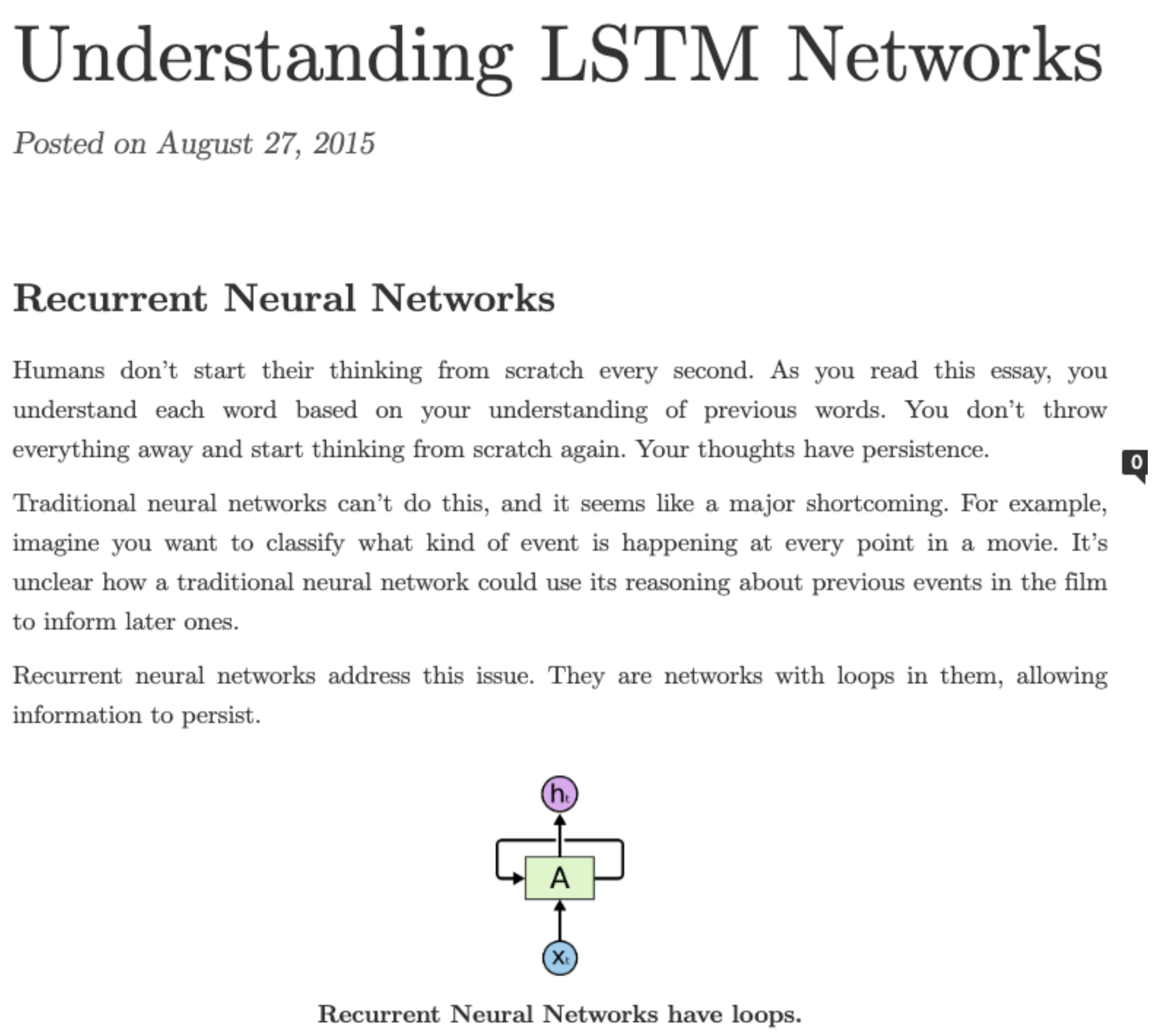
遞歸神經網路（RNN）和長短期記憶模型（LSTM）的運作原理

原文：How Recurrent Neural Networks and Long Short-Term Memory Work

Translated from Brandon Rohrer's Blog by Jimmy Lin

認識RNN

連結



Understanding LSTM Networks

Posted on August 27, 2015

Recurrent Neural Networks

Humans don't start their thinking from scratch every second. As you read this essay, you understand each word based on your understanding of previous words. You don't throw everything away and start thinking from scratch again. Your thoughts have persistence.

Traditional neural networks can't do this, and it seems like a major shortcoming. For example, imagine you want to classify what kind of event is happening at every point in a movie. It's unclear how a traditional neural network could use its reasoning about previous events in the film to inform later ones.

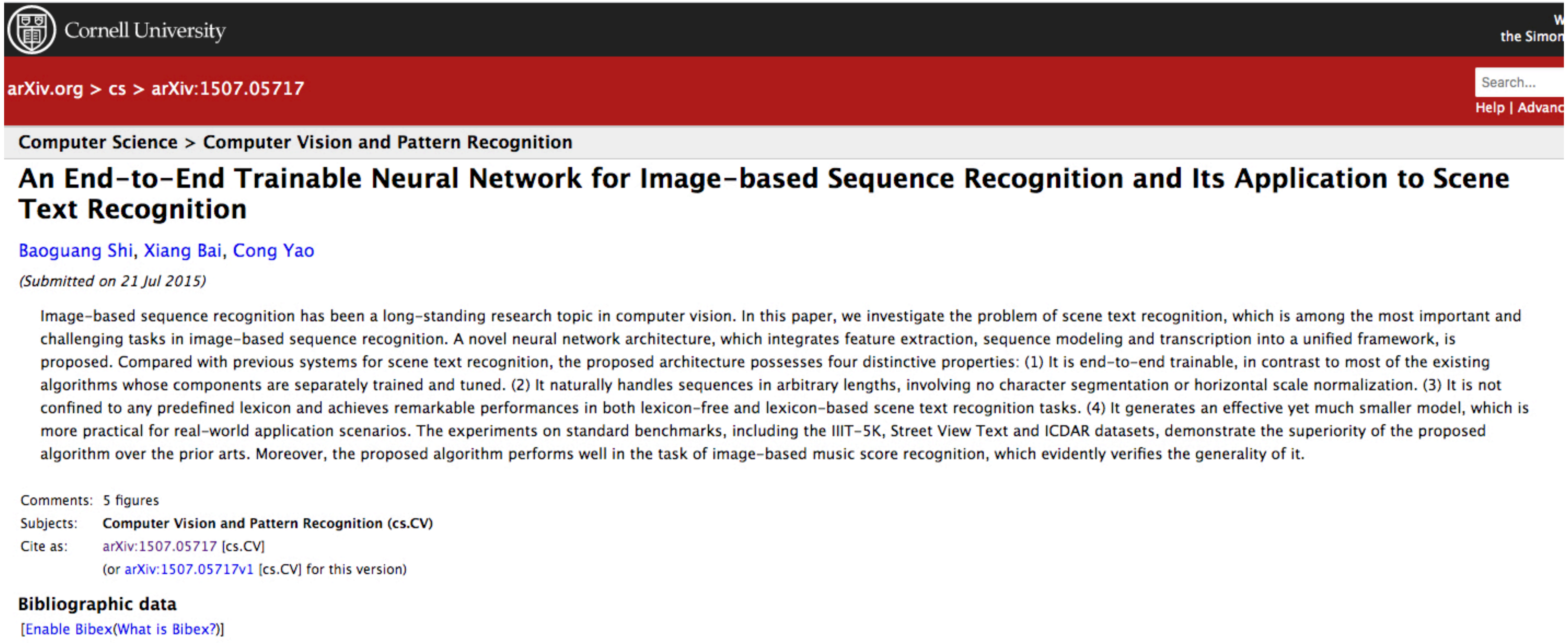
Recurrent neural networks address this issue. They are networks with loops in them, allowing information to persist.

Recurrent Neural Networks have loops.

認識LSTM

連結

推薦延伸閱讀



arXiv.org > cs > arXiv:1507.05717

Computer Science > Computer Vision and Pattern Recognition

An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition

Baoguang Shi, Xiang Bai, Cong Yao

(Submitted on 21 Jul 2015)

Image-based sequence recognition has been a long-standing research topic in computer vision. In this paper, we investigate the problem of scene text recognition, which is among the most important and challenging tasks in image-based sequence recognition. A novel neural network architecture, which integrates feature extraction, sequence modeling and transcription into a unified framework, is proposed. Compared with previous systems for scene text recognition, the proposed architecture possesses four distinctive properties: (1) It is end-to-end trainable, in contrast to most of the existing algorithms whose components are separately trained and tuned. (2) It naturally handles sequences in arbitrary lengths, involving no character segmentation or horizontal scale normalization. (3) It is not confined to any predefined lexicon and achieves remarkable performances in both lexicon-free and lexicon-based scene text recognition tasks. (4) It generates an effective yet much smaller model, which is more practical for real-world application scenarios. The experiments on standard benchmarks, including the IIIT-5K, Street View Text and ICDAR datasets, demonstrate the superiority of the proposed algorithm over the prior arts. Moreover, the proposed algorithm performs well in the task of image-based music score recognition, which evidently verifies the generality of it.

Comments:	5 figures
Subjects:	**Computer Vision and Pattern Recognition (cs.CV)**
Cite as:	arXiv:1507.05717 [cs.CV]
	(or arXiv:1507.05717v1 [cs.CV] for this version)

Bibliographic data

[Enable Bibex(What is Bibex?)]

CRNN paper

連結

解題時間 Let’s Crack It

請跳出 PDF 至官網 Sample Code & 作業開始解題