模型性能比較

語音辨識模型比較

從表一可以得知，在我們找到的可行模型中，ReazonSpeech\_v2做語音辨識任務時，單字錯誤率最低，最為穩健。

在我們測試時，也發現ReazonSpeech\_v2的速度快於Whisper一倍，正確率也遠高於ESPnet，因此我們選用ReazonSpeech\_v2模型。

表一、使用淡江日文系學生錄製的日文音檔進行模型評估

|  |  |
| --- | --- |
| Model | WER(word error rate),% |
| Whisper | 29.33 |
| ESPnet | 80.00 |
| **ReazonSpeech\_v2** | **28.89** |

評分模型比較

如表二所示，我們可以得到有關於模型架構對其在驗證集的表現，以下會先對核心層說明，再討論加入GAP層和LINEAR層後性能提升的情況及其可能原因。

核心層：

1. Linear Layer：

僅使用Linear層的模型表現最差，PPC為0.258。這表明單純的線性層無法有效捕捉到語音上下文中的複雜特徵，因此不適合用於模型的核心。

1. BGRU Layer：

只使用BGRU層的模型，PPC為0.533。相較於Linear層，BGRU能夠更好地捕捉序列中的時序資訊，但其表現仍然有限。

1. BLSTM Layer：

只使用BLSTM層的模型，PPC為0.606。BLSTM在捕捉雙向時序資訊方面比BGRU更優秀，因此其性能優於BGRU模型。

GAP層的影響：

將GAP層加入到LINEAR層後，PPC從0.258提高到0.504。

將GAP層加入到BGRU層後，PPC從0.533提高到0.546。

將GAP層加入到BLSTM層後，PPC從0.606提高到0.612。

GAP層可以將時序特徵進行全局平均池化，減少了模型的複雜度，同時保留了重要的全局信息，這有助於提高模型的泛化能力和性能。

LINEAR層的影響：

將LINEAR層加入到BGRU層後，PPC從0.533提高到0.594。

將LINEAR層加入到BLSTM層後，PPC從0.606提高到0.633。

LINEAR層能夠進一步提取高層次特徵，並加強特徵之間的相關性，這使得模型能夠更好地理解和處理複雜的語音特徵。

GAP和LINEAR層：

將GAP和LINEAR層都加入到BGRU層後，PPC達到0.679，這是BGRU模型中的最佳表現。

將GAP和LINEAR層都加入到BLSTM層後，PPC達到0.701，這是BGRU模型中的最佳表現，同時也是所有模型中的最佳表現。

GAP和LINEAR層同時加入能夠充分提取和利用時序特徵中的重要信息，先將核心層提取的時序特徵用Linear層做整合，再利用GAP層做時間維度的平均化，最大化的提升模型性能。

在模型的性能比較中，單純的Linear層表現最差，無法有效捕捉語音中的複雜特徵。相比之下，BGRU層能夠捕捉序列中的時序信息，但其性能次於BLSTM層。BLSTM層在捕捉雙向時序信息方面最為有效，性能優於BGRU層。

加入GAP層能夠整合全局信息，顯著提升模型性能。Linear層在特徵空間內進行變換，進一步提升模型性能。最佳配置為BLSTM層結合GAP層和Linear層，能夠最大程度地提升模型的自動發音評估性能。

表二、使用fine-tuned HuBERT搭配不同架構評分模型預測與實際標籤的相關性

|  |  |
| --- | --- |
| Model Architecture | PPC |
| Linear  Linear + GAP | 0.258  0.504 |
| BGRU  BGRU + GAP  BGRU + Linear  BGRU + Linear + GAP | 0.533  0.546  0.594  0.679 |
| BLSTM  BLSTM + GAP  BLSTM + Linear  **BLSTM + Linear + GAP** | 0.606  0.612  0.633  **0.701** |