Transformer-based TTS embedded implementation

Sian-Yi Chen

Advisors : Tay-Jyi Lin and Chingwei Yeh

Outline

- Action item
  - 使用FPGA實現Transformer-based TTS的embedding system
- Status report
  - **參數量評估與FPGA板選擇**
    - 架構為 x-vector + Transformer，並各別計算參數量
      1. 手算
      2. 透過程式執行結果
         - 目前遇到困難，尚未完成
         - 原因：專案使用pytorch，而pytorch中有一函式(model.parameters())可以印出模型的參數，此函數需新增至神經網路建模處，但專案中建模的程式碼並沒有被使用，而執行的程式使用的是額外載入(import)的程式
         - 解決方案：尋找其他Transformer專案估計參數，與此版本差別為x-vector，但x-vector參數量很小應可以忽略
    - FPGA 老師指示使用 ZedBoard
  - **後續規劃**
    - 先分別找 Transformer 與 x-vector 的 C code，並確認是否可順利執行再將兩者合併
      1. 目前 Transformer 僅找到 C++版本
      2. 尚未找到 x-vector 的 C code 版本
         - 解決方案
           ① 若找不到，考慮參考論文架構自行實作，x-vector 架構主要為多層DNN，但每層輸入增加了上下文訊息，輸入為連續的幾個frame

x-vector 參數估計

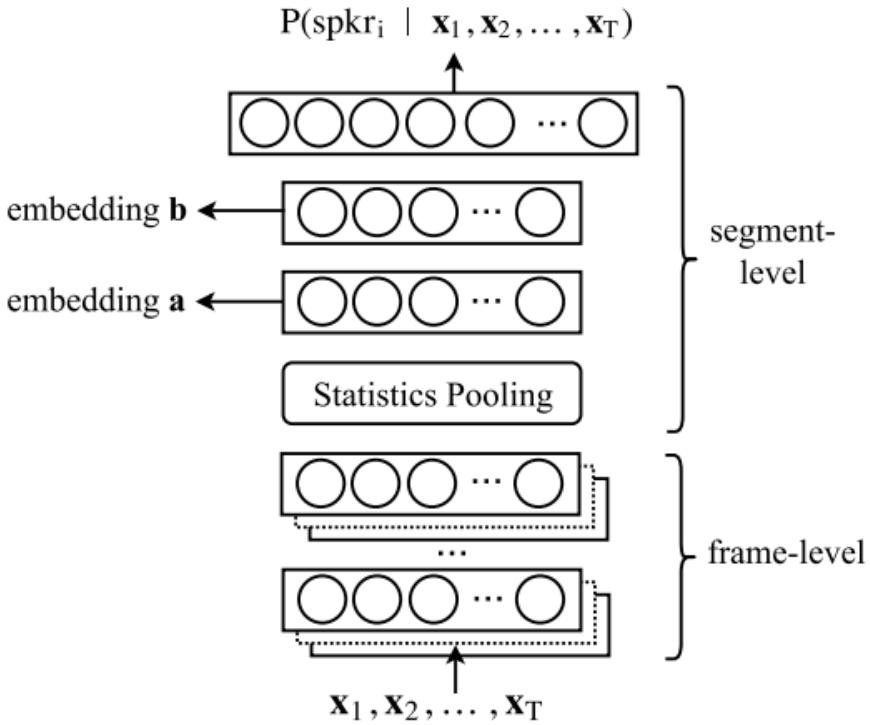
架構

- 包含多層TDNN (結構與DNN相同)
- pooling層
- embedding層
- softmax

x-vextor 參數量計算

- **輸入層**
  - N=1(輸入層特徵為單個frame提取，每一層的time-delay)，J=16(16個單元個數)，總共15個frame
- **第一層**
  - N=2，J=8，總共(15-2)13個frame，參數量8*3*16=384
- **第二層**
  - N=4，J=3，(13-4)9個，參數量3*4*8=120
- **輸出層**
  - N=8，J=3，(9-8)1個，參數量3*9*3=81

- 總參數量為384+120+81=585



X-vector 架構 [1]

[1] D. Snyder, D. G.-Romero, G. Sell, D. Povey, S. Khudanpur, "X-vectors: robust dnn embeddings for speaker recognition." in *Proc. IEEE ICASSP*, 2018.

Transformer 參數估計

Transformer 架構包含

- embedding向量
  - 論文中使用了大小為37000的詞彙表、model為512維
- Multi-heads attention
  - d_{model} = h*d，因此有幾頭並不影響維度變化
  - query與轉置的key做內積，得到512*512的attention，Q、K、V互相獨立、且總共有3塊，最後 N=6
- FeedForward
  - 全連階層公式：$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$
  - W_1與W_2大小相同，為512*(4*512)*總共2層
- （37000+512）* 512 +
- 6 *（512 * 512 * 3 *3 +
- 512 * 512 + 512 * 2048 * 2 * 2) **= 60,100,608**

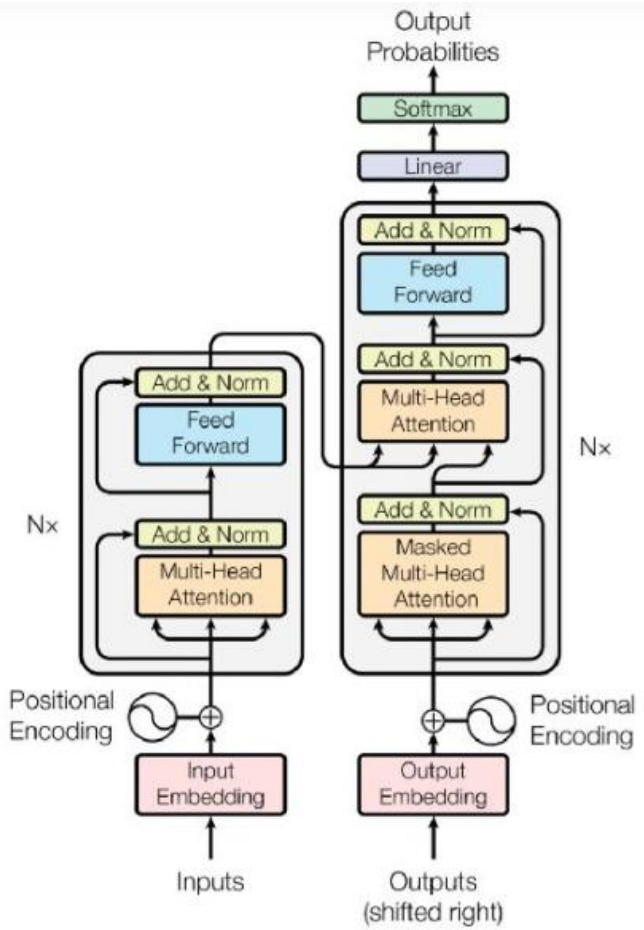


Figure 1: The Transformer - model architecture

	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train steps	PPL (dev)	BLEU (dev)	params $\times 10^6$
base	6	512	2048	8	64	64	0.1	0.1	100K	4.92	25.8	65

Transformer 參數量大小 [2]

[2] A. Vaswani, et al., "Attention is all you need." in *Proc. NIPS*, 2017.