超越硬共享: 使用□家混合□督的高效多任□□音□文本建模

*Hojun Jin**, *Eunsoo Hong**, *Ziwon Hyung*, *Sungjun Lim*, *Seungjin Lee*, *Keunseok Cho*

Samsung Research, Korea

hojun.jin@samsung.com, eunsoo1.hong@samsung.com, ziwon.hyung@samsung.com, sungjunj.lim@samsung.com, sjsr.lee@samsung.com, ks1.cho@samsung.com

Abstract

硬□□共享是在不同任□中共同□□□□模型的一□常□策略。然而 ,□通常□□致任□干□ ,妨碍□□模型性能。□了解□□□□□ ,我□提出了一□□□而有效的□督□家混合(S-MoE)。□□□的□家混合模型不同 ,S-MoE通□利用特殊指□□□□□每□任□路由到其指定□家 ,□而消除了□□□控函□的需要。通□□每□任□分配到一□□□的前□□□ ,S-MoE克服了硬□□共享的限制。我□□一步□S-MoE□用于□音到文本模型 ,使模型能□□理混合□□□入 ,同□□合□行自□□音□□(ASR)和□音□□(ST)。□□□果□明了所提出的S-MoE的有效性 ,□同□□用于□□器和解□器□ ,在字□□率(WER)方面□□了相□提升6.35 %

Index Terms: automatic speech recognition, speech translation, multi-task learning, supervised mixture of experts

1. 介□

□音□文本(STT)模型通常在□□(WB)音□上□行□□ ,因□□能提供更□富的□□特征。然而 ,窄□(NB)音□在移□□境中仍然是必不可少的 ,特□是在□□□用方面。由于采□率差□□致的□□差□ ,NB和WB模型通常是分□□□的。此外 ,□多□音□文本□用需要□行自□□音□□(ASR)和□音□□(ST)。□部署的角度□看 ,由于移□和嵌入式□□上的□源限制 ,□□多□用于每□任□的模型是不切□□的。

多任□□□(MTL)是解□□一挑□的一□有前途的解□方案 ,□使一□模型能□在不同的□入和任□之□□□共享表示 [?, ?, ?] 。□□的使用硬□□共享的MTL方法□常遭遇任□干□ ,其中一□任□的□化□降低其他任□的性能 [?] 。

□了□□□□干□ ,已□提出了□多方法 ,其中一□有效的方法是使用像□家混合(MoE) [?] □□的模□化架□。MoE因其能□□不同任□分配□□的□□而受到□注。然而 ,□□的MoE模型□重依□于□□□□入路由到特定□家的□控函□。

在本文中 ,我□提出了一□□□□督□家混合(S-MoE)的新方法 ,□□高效的多任□STT□□□□。我□的方法借□了MoE原□ ,但□□□入和□出□件均已明□的□景□行了定制。我□□注于□□化的多□件任□ ,在□□情□下 ,□督路由提供了一□□□而有效的解□方案。□□准的MoE模型不同 ,我□的方法通□引入特殊的指□□□□□式地□任□引□到□用的□家□□ ,□而消除了□□□控函□的需要。□□方法□化了路由机制 ,同□□保模型在□□和推理□程中不增加□外的□算□□ ,保持了效率。

本文的□□如下:

*These authors contributed equally.

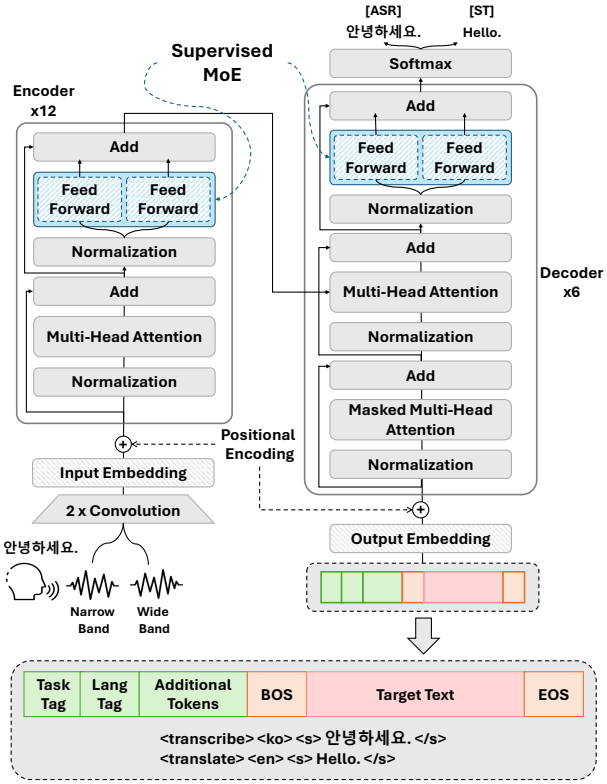

Figure 1: 嵌入在*STT*模型中的*S-MoE*□。□□器中的*S-MoE*根据□音□□(窄□或□□)□□入□向□家□□ ,而解□器中的*S-MoE*□根据任□(自□□音□□或□音□□)□路由□入。

- 我□提出了一□S-MoE方法 ,通□使用指□□□□行任□路由 ,□少□算成本 ,□而□□MTL中的任□干□。
- 我□□一步□S-MoE□□□用于一□STT模型 ,使其能□□理混合□□□入 ,同□□合□行ASR和ST。
- □我□的方法同□□用于□□器和解□器□ ,在WER上取得了6.35 % 的提升。

2. 方法

在本□中 ,我□提出了一□如□ 1 所示的架□。□一步的□□□在以下子□中介□。

2.1. □督□家混合模型 (S-MoE)

通□稍微□整□□的 MoE [?] □□ ,S-MoE 模□的□出 y 表□□方程 1 。□□准 MoE □相似 ,$E_i(x)$ 表示□定□入 x □的第 i □□家□□的□出。S-MoE 的□□□□在于□控机制。□□□ MoE 模型依□可□□□控□□ G 不同 ,我□的方法采用□定□的□控函□ G' ,无需□外的□控□□□□。如□ 1 所示 ,在整□工作□程中 ,□□器和解□器中的□家□□□量 n 均固定□ 2 。

$$y = \sum_{i=0}^{n-1} G'(x)_i E_i(x) \tag{1}$$

2.1.1. □□器*S-MoE*的□控功能

在□理□□(WB)□音□ ,基于窄□(NB)□□的模型□向于表□不佳 [?, ?] 。□了在一□模型中有效□理NB和WB信□ ,我□在□□器中引入了一□S-MoE架□。在□□□□中 ,前□□□(FFN)作□每□□□的□用□家 ,而□□器模□中的其余□件□在NB和WB□入之□共享。□□信息□每□□入□先□□ ,□使得□控函□ $G'(x)$ 可以有□□地激活合适的□家。□□准MoE□似 ,如果 $G'(x)_i = 0$,□不□□算相□的□家 $E_i(x)$ 。因此 ,WB信□由□家 E_0 □理 ,而NB信□□由□家 E_1 □理。□□器S-MoE的□控函□在公式 2 中被正式定□。

$$G'_{enc}(x)_i = \begin{cases} G'_{enc}(x)_0 & = \begin{cases} 0, & \text{if } x \text{ is NB signals} \\ 1, & \text{if } x \text{ is WB signals} \end{cases} \\ G'_{enc}(x)_1 & = \begin{cases} 0, & \text{if } x \text{ is WB signals} \\ 1, & \text{if } x \text{ is NB signals} \end{cases} \end{cases} \tag{2}$$

2.1.2. 解□器 *S-MoE* 的□控函□

□□□器□似 ,我□在解□器中□用了S-MoE架□ ,以有效地□理ASR和ST任□。□了□定合适的解□□家 ,任□□□被添加到文本□入的前面。基于目□任□ ,□控函□□ASR□入□向□家 E_1 ,□ST□入□向□家 E_0 。解□器S-MoE的□控函□在方程 3 中□出。

$$G'_{dec}(x)_i = \begin{cases} G'_{dec}(x)_0 & = \begin{cases} 0, & \text{if task of } x \text{ is ASR} \\ 1, & \text{if task of } x \text{ is ST} \end{cases} \\ G'_{dec}(x)_1 & = \begin{cases} 0, & \text{if task of } x \text{ is ST} \\ 1, & \text{if task of } x \text{ is ASR} \end{cases} \end{cases} \tag{3}$$

如□ 1 所示 ,□□□程包含特殊□□以引□模型。在每□序列的□□ ,□入一□任□□□(`<transcribe>` 或 `<translate>`) ,□□其后的是目□□言□□(`<en>` 或 `<ko>`)。其他特殊□□也可能被包括在□ ,但本□究中未使用。序列然后以 `<beginning_of_sentence>`□□□始 ,后面□□目□文本。□ 2 展示了所提出的基于 S-MoE 的多任□模型的嵌入流程。

2.1.3. □□□段

□入音□ ,无□是NB(8 kHz)□是WB(16 kHz) ,首先被□理□□入到Transformer□□器中。在□□器□ ,基于Encoder S-MoE的□控函□□□□□FFN□。Encoder S-MoE的□控函□根据□□/推理期□□入信□的□□□□定□家 ,如在第 2.1.1 □中所述。□意味着□于NB和WB□入使用不同的FFN□ ,□而使模型能□有效捕捉特定□□的表示。在解□器中 ,Decoder S-MoE利用□□□立的FFN□□□理共享的□□表示。每□□□批次由□一任□的□本□成 ,以□保任□特定的□化。模型通□交□批次□行□□ ,□意味着ASR和ST任□交替□□。

2.1.4. 推理□段

在推理□程中 ,模型可以同□生成ASR和ST□出。通□□批量大小□置□2 ,□相□地分配ASR和ST任□□□ ,可以在一次推理步□中同□□得□□和□□□果。

3. □□□置

3.1. □据集

3.1.1. □□□据集

我□的□□□据集由□自公□可用的AIHub□料□ [?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?] 的40,000小□□□□音□据□成。□于自□□音□□(ASR) ,我□使用提供的□□文本作□目□文本。□于□音□□(ST) ,如果□有相□的英文□□ ,我□□使用□部机器□□(MT)模型生成目□文本。如果提供了原始□□ ,我□□直接使用□。因此 ,每□音□□本都有配□的ASR和ST□据 ,□而□□多任□□□。

□了在不同□□□件下□□模型 ,我□生成了□据的NB和WB□□。WB□据□ 16 kHz ,而NB□据□ 8 kHz 。我□采用□□□□方法:(1)□□地□ 16 kHz 下采□到8 kHz ,(2)使用AMR-WB [?] 、AMR-NB和G.711-NB [?] □解□器□行基于□解□器的□理。□些□□□用于15 % NB/WB

3.1.2. □□□据集

□了□□ASR和ST的□□性能 ,我□使用了□□□部□□集以及公□可用的Fleurs [?] □据集。我□的□部□□集由1000□□自日常□□的男女□音□本□成 ,□由□□□者撰□□考□□。此外 ,□了在不同□□□件下□□模型 ,我□使用□□□中相同的□□方法生成□□集的NB/WB□□。

3.2. 模型

我□的基□模型遵循一□基于Transformer的序列到序列架□ [?] ,包含一□12□的□□器和6□解□器。我□采用正弦位置嵌入和□有GLU激活的□正□化Transformer□□ ,□使用8□注意力□。前□□□的□部□度□□2048 ,嵌入□度□512 ,且使用0.15的置□率。□于分□ ,我□使用□□大小□40,000的字□□字□□□□(BBPE)分□器。□入□音被重采□到 16 kHz ,□□□成80□的□□梅□倒□□波器□(fbank)特征 ,□些特征是使用 25 ms □□□和 10 ms 步□提取的。我□□用最大□音序列□度□3,000和最大文本序列□度□120的限制 ,□□掉□于30秒的□音片段。模型□□10□□次 ,使用大小□3,200的有效批次 ,□采用余弦退火□度器□行□□率衰□。章□ 4.1 的□□器-解□器S-MoE模型通□章□ ?? 的解□器S-MoE模型□行初始化 ,□使用NB/WB微□子集□□5□□次。

3.3. 度量

□于ST ,我□使用BLEU指□ [?] □衡量□□□量。我□使用了 `sacrebleu` □。□于ASR ,我□□告WER ,□是ASR□究中□泛使用的指□。

4. □□□果

□了□□S-MoE□用在解□器模□中的有效性 ,我□□行了□□任□的□□:□□到英□(ko2en)的□音□□(ST)

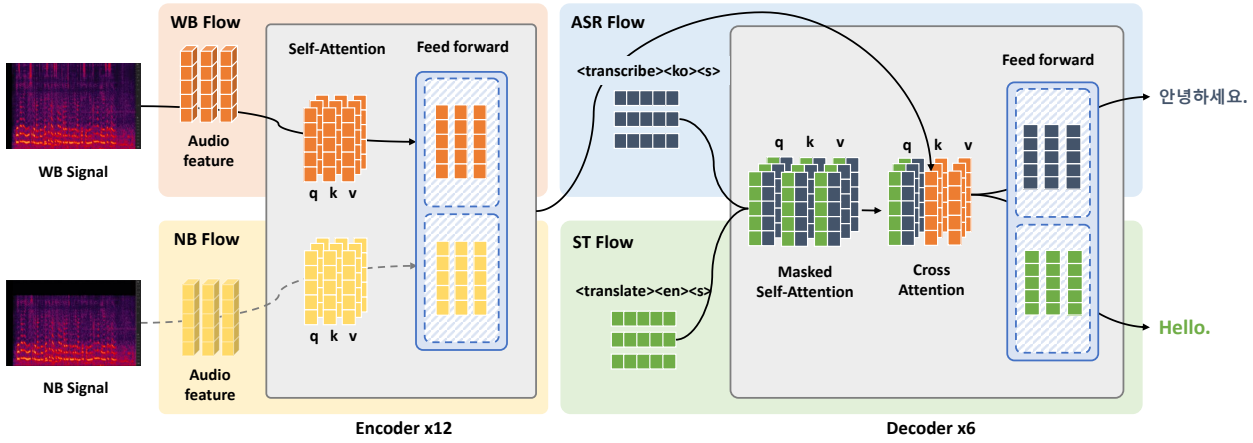

Figure 2: *S-MoE Transformer* 的嵌入流。在□□/推理期□ ,□□器□□□一□前□□ ,而解□器□的□□前□□都被利用。

Table 1: 解□器*S-MoE*□*ASR/ST*的影□

Model	Param		Male TC		Female TC		Fleurs	
	Trainable	Active	WER	BLEU	WER	BLEU	WER	BLEU
Whisper-base	74M	74M	8.73	13.6	13.40	11.9	18.03	3.3
Whisper-small	244M	244M	5.15	26.5	7.40	25.8	10.27	12.6
Whisper-large-v3	1,550M	1,550M	4.46	33.7	6.87	32.8	5.32	20.1
Ours (Base-Model)	107M	107M	3.93	34.3	5.36	33.2	5.46	21.1
Ours (DecFFNx2)	119M	119M	3.83	34.1	5.36	33.5	5.46	21.5
Ours (DecS-MoE)	119M	107M	3.53	34.8	4.89	33.9	5.23	21.5

Table 2: □□器-解□器 *S-MoE* □ *ASR/ST* 的影□

Model	Codec	Male TC		Female TC		Fleurs	
		WER	BLEU	WER	BLEU	WER	BLEU
Base-Model	NB	4.11	33.8	6.60	32.5	5.91	19.9
	G.711 NB	4.18	33.9	6.77	32.4	6.50	19.9
	AMR NB	4.78	32.8	11.30	29.8	13.69	17.9
	WB	4.51	33.8	6.16	32.0	8.71	19.1
	AMR WB	3.98	33.9	5.46	32.9	5.84	19.8
EncDecS-MoE	NB	4.08	34.1	6.22	32.9	5.66	20.4
	G.711 NB	3.92	34.2	5.96	32.8	6.15	20.0
	AMR NB	4.65	34.0	10.43	30.5	12.96	18.2
	WB	4.17	34.2	6.98	32.3	8.30	19.1
	AMR WB	3.88	34.2	5.18	33.5	5.34	20.2

和□□自□□音□□(ko-ASR)。□果□□在表 1 中。第 3.2 □中描述的模型作□□比的主要基□(Base-Model)。

□着□家□量的增加 ,S-MoE的可□□□□□性增□。根据 [?] ,□管□□略有增加 ,但 MoE □□仍保留了□算□□ ,因□□基于路由机制□激活了部分□家□□。在我□的□□中 ,我□□□家□量□置□2 ,以支持 ASR 和 ST 的多任□□□。在表格中 ,□S-MoE □用于解□器的模型□□DecS-MoE。□了匹配DecS-MoE中的□□增加 ,我□□□□了一□□□ ,□□□的解□器FFN□度加倍(DecFFNx2)。

□了更□泛的比□ ,我□在□□中包括了OpenAI的Whisper [?] 模型。Whisper是一□□泛使用的多□言□音模型 ,使用多□□言□□行□□。我□的模型□□□一□□方向□行了□化 ,所以直接比□可能不完全公平 ,但Whisper□□□ko-ASR和ko2en ST性能提供了一□強有力的□照点。

如表 1 所示 ,DecS-MoE在所有□□集上的ST和ASR方面均持□□于基□模型。我□□察到 ,在男性□□集上 ,BLEU提高了0.5点 ,WER降低了0.4 % BLEU 0.7 WER 0.47 % Fleurs DecS-MoE BLEU 0.4 WER 0.2 % DecS-MoE WER 0.37 % BLEU 0.53 WER 8.06 % BLEU 1.77 % S-MoE

在可□□□□□量相同的情□下 ,DecS-MoE甚至比DecFFNx2表□更佳。在男性□□集上 ,BLEU提高了0.7分 ,而WER降低了0.3 % DecS-MoE BLEU 0.4 WER 0.47 % Fleurs ST WER 0.23 % WER 0.33 % BLEU 0.37 WER 7.33 % BLEU

1.22 % FFN
DecS-MoE

□使使用Whisper模型,DecS-MoE在所有ASR/ST□□集上也取得了卓越的性能。□管其可□□□□□□Whisper-large(1,550M)的十分之一,DecS-MoE(119M)的平均BLEU提升了1.2,同□WER□少了1 %。□表明WER相□提升了21.98 %,BLEU提升了3.99 %。

4.1. □□器-解□器 S-MoE

我□提出了一□□□器 S-MoE □□理□□不同的□音信□。具□而言,□□器由□□□家□成,一□用于NB信□,一□用于WB信□。在 **??** □中,我□□明了解□器 S-MoE在□合□□ ASR 和 ST 任□□效果□著。因此,我□□ S-MoE □用于□□器和解□器(EncDecS-MoE)。□共四□□家,EncDecS-MoE 模型有144M□可□□□□,而基□模型有107M。然而,在推理□程中,□□□模型使用的主□□□□量相同(107M)。例如,在□理用于ST的NB□音□入□,EncDecS-MoE 模型□□□器路由到NB□家,□解□器路由到ST□家,只激活相□的□家以□□高效□算。

表 2 中的所有模型都在第 3.1.1 □描述的NB/WB□据上□行了微□。由于□WB信□相比,NB□音信□的使用更□有限,因此NB不需要太多的□□□据。□了在NB和WB□境中都能有效□□,我□□□□□的基□模型□行了微□,□模型最初□在WB□据上□行□□。

如表 2 所示,EncDecS-MoE 在□乎所有情□下都表□出一致的性能提升。在 NB □境中,EncDecS-MoE 的平均 BLEU 得分□ 28.57,而基准模型□ 28.10,平均 WER □ 6.67 % 7.09 % WER 相□提高了 6.35 % BLEU 得分提高了 1.63 % WB □境下,我□提出的模型在 WER 上相□于平均性能有 2.39 % 的相□提升,在 BLEU 上有 1.15 % 的相□提升。NB 信□□捕□□低□率,□可能在□理 WB 和 NB □据的□一模型□□□□□生□面影□。然而,□果表明在□□器□□施□立的 FFNs 可以解□□一□□。□然使用□立的□家□略微增加模型大小,但活□□□的□量□基准模型相同。因此,在□□器和解□器中□用 S-MoE 提升了性能,而不影□推理速度。

5. □□

我□提出了S-MoE架□,旨在□解多任□□□中□STT□用的任□干□。通□使用指□性□□代替□□□控功能,S-MoE在提高各□任□性能的同□,□保了高效的□□和推理。我□的方法使得□一模型能□□理具有混合□□□入的ASR和ST任□,□使其成□□源受限□境的理想□□。未□的工作□集中在□展S-MoE以支持更多□音任□和多□言功能,□而□一步增強其在□□世界中的适用性。

6. References