2-2-8-MLP

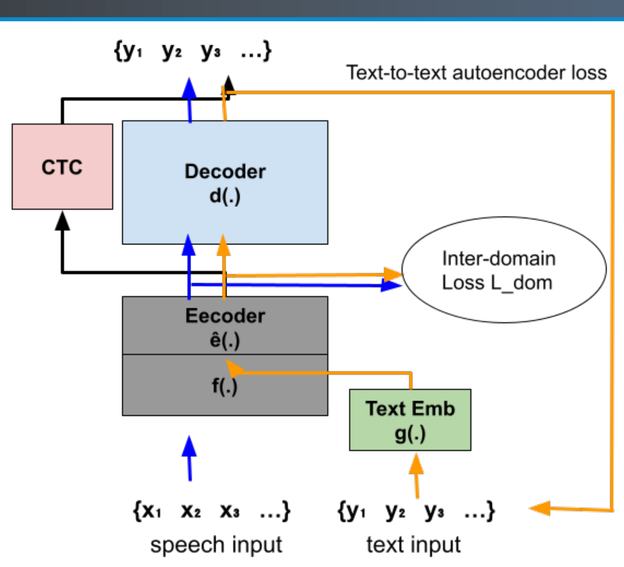
Improving Semi-supervised End-to-End ASR using CycleGAN and Inter-domain Losses

Chia-Yu Li and Ngoc Thang Vu
Institute for Natural Language Processing - University of Stuttgart, Germany

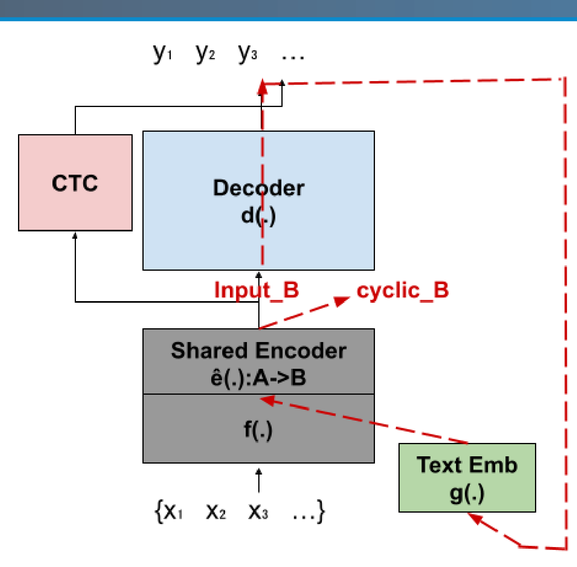
Main Contributions

- We propose a cycle-consistent inter-domain loss, which is dissimilarity between encoded speech and hypothesis, for generating better representation. Besides, we combine the cycle-consistent inter-domain loss and the identity mapping loss from CycleGAN in a single framework for semi-supervised E2E ASR and achieve noticeable performance improvement.
- We provide the analysis on the ASR output and the visualization of inter-domain embedding from speech and text, which explains the reason of performance gain by our proposed method.

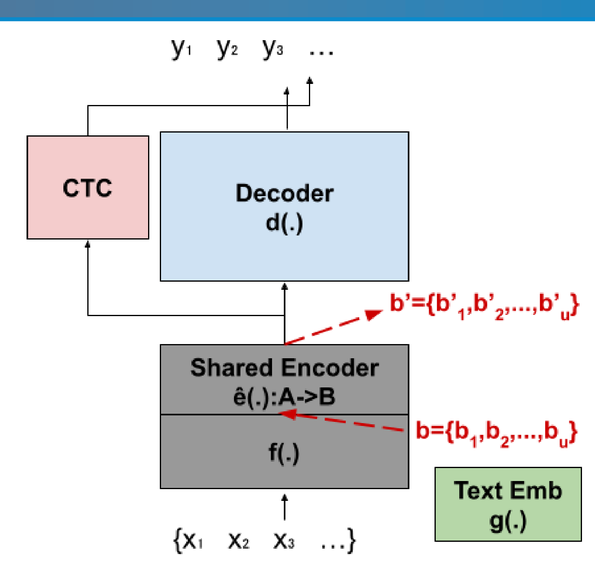
Method



(a) Semi-supervised E2E (Karital, 2018)



(b) CycleGAN inter-domain loss



(c) Identity mapping loss

- Figure (a) is model architecture for supervised training (oracle and initial model).
- Figure (b) is the cycle-consistent inter-domain loss, which minimizes the distance between the inter-domain embedding from speech and its hypothesis. $L_{cyc,dom} = \mathcal{D}(input\_B, cycle\_B) = \mathcal{D}(e(x), \hat{e}(g(d(e(x)))))$ Where $\mathcal{D}(.)$ is a method to measure the distance between distributions. In this work, we use Maximum Mean Discrepancy (MMD).
- Figure (c) is the identity mapping loss that encourages to preserves important features after translation. $L_{idt} = \|\hat{e}(b) - b\|_1$.

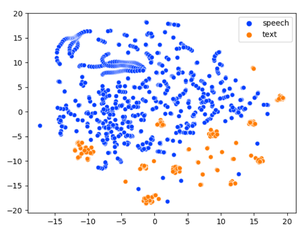
Resource

We conducted the experiments on WSJ, Librispeech 100+360 Task and Voxforge (it,nl,de,fr). The oracle and initial models in result table are trained by Espnet [3] and our method is implemented under Espnet [a].

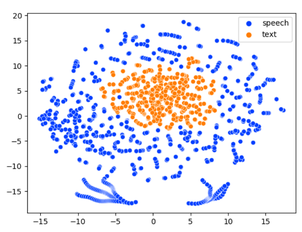
[a] https://github.com/chiayuli/semi-supervised-E2E-using-CycleGAN

Visualization

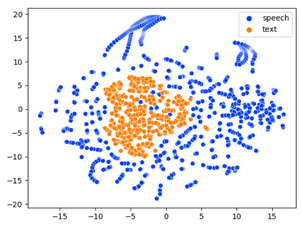
t-SNE visualization of inter-domain embedding from speech and text. Our retrained models show better regularization.



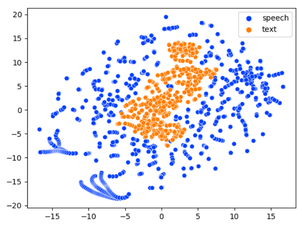
(a) Baseline [2] ●



(b) Retrain-idt ●



(c) Retrain-cyc ●



(d) Retrain-cyc+idt ●

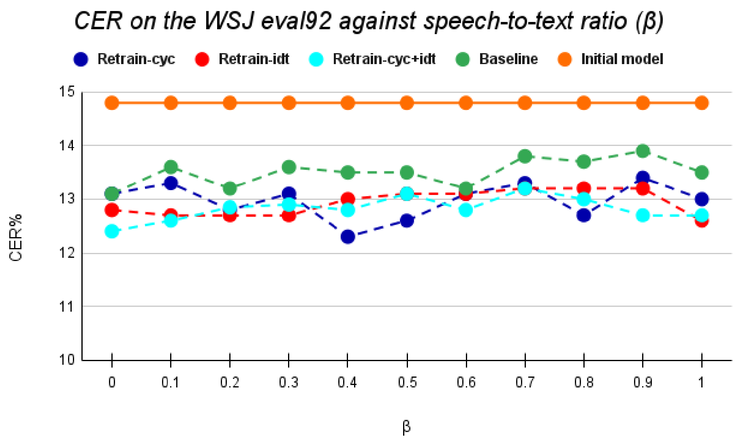
References

[1] J.-Y. Zhu et al., "Unpaired image-to-image translation using cycle-consistent ad- versarial networks," in Proc. of ICCV, 2017.
[2] S. Karita1 et al., "Semi- supervised end-to-end speech recognition," in Proc. of Interspeech, 2018.
[3] S. Kim et al., "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in Proc. of ICASSP, 2017.

Impact of using CycleGAN losses

Model	Objective	paired data	unpaired data
Initial model ●	$L = L_{pair} = (1 - mtl) * L_{ctc} + mtl * L_{att}$	v	
Semi-supervised models	$L = \alpha * L_{pair} + (1 - \alpha) * L_{unpair}$		
-Baseline [2] ●	$L_{unpair} = \beta L_{dom} + (1 - \beta) L_{text}$	v	v
-Retrain-idt ●	$L_{unpair} = L_{idt}$	v	v
-Retrain-cyc ●	$L_{unpair} = \beta L_{cyc,dom} + (1 - \beta) L_{text}$	v	v
-Retrain-cyc+idt ●	$L_{unpair} = \beta(L_{cyc,dom} + L_{idt}(x)) + (1 - \beta)(L_{text} + L_{idt}(y))$	v	v

- Retrain-idt (red dots) has better CER than Baseline, and its performance does not fluctuate over speech-to-text ratio. Retrain-cyc (blue dots) achieves the best CER at $\beta = 0.4$ and it also performs better than the Baseline (green dots) all the time except at $\beta = 0$.
- Retrain-cyc outperforms Baseline at $\beta = 1$, which implies that the encoder using our proposed $L_{cyc,dom}$ generates better embedding than the one using L_{dom}.
- Retrain-cyc+idt (cyan dots), which combines L_{idt}, $L_{cyc,dom}$ and L_{text}, have advantages from the both losses and achieves good performance while β varies.



CER/WER across English corpus

WSJ (15h paired+80h unpaired data)

Model	Type	LM	CER(%)	WER(%)
Oracle	-	N	4.3	14.1
Initial model	-	N	14.8	42.6
Baseline	Text	N	13.1	38.3
Retrain-cyc+idt	Text	N	**12.4**	**36.9**
Baseline	Both	N	13.5	39.6
Retrain-cyc+idt	Both	N	**12.5**	**36.9**
Oracle	-	Y	2.3	4.9
Initial model	-	Y	8.3	17.6
Baseline	Text	Y	7.3	15.8
Retrain-cyc+idt	Text	Y	**7.1**	**15.4**
Baseline	Both	Y	7.4	15.8
Retrain-cyc+idt	Both	Y	**6.9**	**15.2**

Librispeech (100h paired+360h unpaired data)

Model	Type	LM	CER(%)	WER(%)
Oracle	-	N	3.9	11.0
Initial model	-	N	8.7	22.7
Baseline	Text	N	8.5	22.4
Retrain-cyc+idt	Text	N	**8.3**	**21.7**
Baseline	Both	N	8.5	22.4
Retrain-cyc+idt	Both	N	**8.1**	**21.4**
Oracle	-	Y	3.5	8.9
Initial model	-	Y	7.0	16.1
Baseline	Text	Y	6.8	15.8
Retrain-cyc+idt	Text	Y	**6.7**	**15.6**
Baseline	Both	Y	6.8	15.6
Retrain-cyc+idt	Both	Y	**6.6**	**15.2**

CER on non-English data

Models paired data (hr.)	it 5	nl 5	de 10	fr 5
Oracle	12.9	25.2	5.6	30.8
Initial model	29.4	35	20.3	53.3
Baseline	22.1	33.7	20.2	47.9
Retrain-cyc+idt	**19.7**	**32.8**	**19.4**	**41.4**

Analysis

Our methods show better insertion and substitution.

REF	Baseline	Retrain-idt	Retrain-cyc
departed	the parted	departed	the parted
commodore	commod or	commodare	commodare
/sil/	a	/sil/	/sil/
making	make at	making	makean