Guiding CTC Posterior Spike Timings for Improved Posterior Fusion and Knowledge Distillation

Gakuto Kurata (IBM Research – Tokyo)
Kartik Audhkhasi (IBM T. J. Watson Research Center)

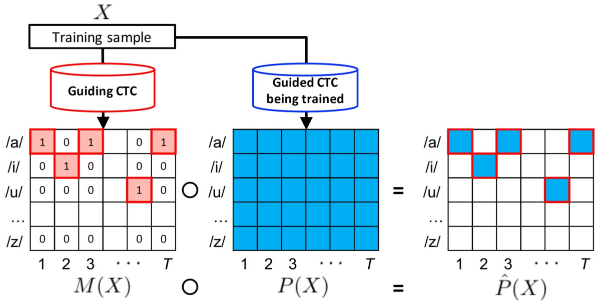
IBM Research AI

Summary

- Sparse and arbitrary posterior spike timings from CTC models pose a new set of challenges in posterior fusion and knowledge distillation from multiple CTC models.
- **We propose a method to train a CTC model so that its spike timings are guided to align with those of a pre-trained *guiding* CTC model.**
- We demonstrate the advantage of our method in various scenarios including posterior fusion of CTC models and knowledge distillation between CTC models with different architectures.

Guided CTC Training

1. Feed a training sample *X* to a pre-trained *guiding* CTC model and obtain posteriors for each time index.
2. Convert the posteriors to a mask *M(X)* by setting 1 at the output symbol with the highest posteriors and 0 at other symbols at each time index.
3. Feed the same training sample to the *guided* CTC model being trained and obtain posteriors *P(X)*.
4. Maximize *M(X)*○*P(X)* jointly with minimizing the CTC loss to train the *guided* CTC model.

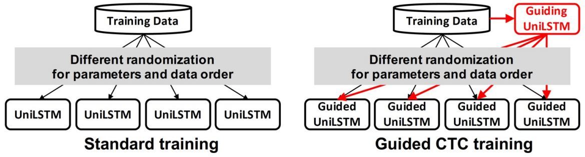


Equivalent with minimizing the frame-level cross entropy where the target is a sequence of the output symbols with the highest posterior from the guiding model over non-blank time indices.

Experiments

Posterior fusion of multiple UniLSTM phone CTC models guided by UniLSTM:

Guided training itself improved accuracy (1A and 1D).



	SWB	CH
1A UniLSTM	15.3	27.6
1B 4× posterior fusion of 1A	15.4	28.8
1C 4× ROVER of 1A	14.1	26.1
1D UniLSTM guided by UniLSTM	14.4	26.2
1E 4× posterior fusion of 1D	12.9	24.2
1F 4× ROVER of 1D	13.7	24.5

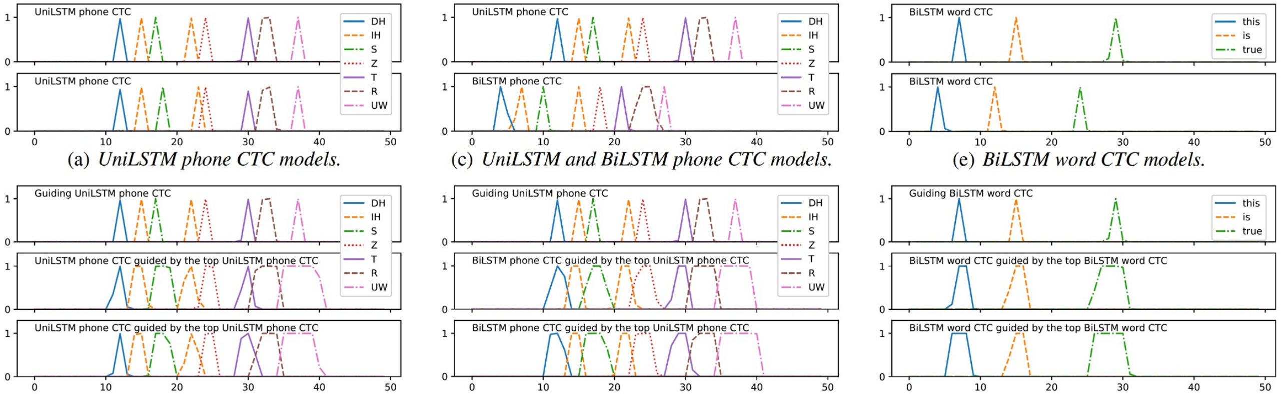
Knowledge distillation between BiLSTM phone CTC and UniLSTM phone CTC:

BiLSTM guided by UniLSTM was useful to train UniLSTM with knowledge distillation.

	SWB	CH
2A UniLSTM	15.3	27.6
2B BiLSTM	11.8	21.8
2C UniLSTM distilled from		
1× BiLSTM (2B)	17.1	29.9
4× BiLSTMs (2B)	29.4	32.7
2D BiLSTM guided by UniLSTM	12.4	22.6
2E UniLSTM distilled from		
1× BiLSTM guided by UniLSTM (2D)	13.4	25.4
4× BiLSTMs guided by UniLSTM (2D)	12.9	24.8
8× BiLSTMs guided by UniLSTM (2D)	12.9	24.7

Knowledge distillation and posterior fusion for BiLSTM word CTC models

	SWB	CH	RT02	RT03	RT04	Avg.
3A BiLSTM	14.9	24.1	23.7	24.1	22.6	21.9
3B 4×posterior fusion of 3A	48.2	57.7	57.7	58.9	59.3	56.4
3C 4×ROVER of 3A	16.0	23.2	24.8	26.1	26.7	23.3
3D BiLSTM guided by BiLSTM	14.3	23.3	23.1	23.8	22.0	21.3
3E 4×posterior fusion of 3D	11.7	20.2	19.2	19.7	18.5	17.9
3F 4×ROVER of 3D	13.0	20.6	20.9	21.2	19.9	19.1
3G BiLSTM distilled from 4×posterior fusion (3E)	13.7	23.1	22.4	22.9	21.7	20.8



(a) *UniLSTM phone CTC models.*

(c) *UniLSTM and BiLSTM phone CTC models.*

(e) *BiLSTM word CTC models.*

(b) *UniLSTM phone CTC models. Bottom two models are guided by the top model.*

(d) *UniLSTM and BiLSTM phone CTC models. Bottom two BiLSTM models are guided by the top UniLSTM model.*

(f) *BiLSTM word CTC models. Bottom two models are guided by the top model.*