Integrating Source-channel and Attention-based Sequence-to-sequence Models for Speech Recognition

Qiujia Li, Chao Zhang, Phil Woodland
Cambridge University Engineering Department, Cambridge, UK

Introduction

Motivation:

- Two major classes of models for speech recognition:
  - **Source-channel models** (SC), *e.g.* DNN-HMM, CTC;
  - **Attention-based sequence-to-sequence models**, *e.g.* Listen Attend & Spell, Transformer.
- SC & attention-based models are:
  - different assumptions, architecture & training objectives;
  - potentially very complementary.
- Existing model combination methods are not directly applicable.

Contribution:

- Proposed **Integrated Source-Channel & Attention** (ISCA) framework that rescores SC model hypotheses with attention-based model;
- Demonstrated effectiveness of ISCA framework, both for multi-task training and separate training of two models;
- ISCA gives relative WER reduction up to 21% over individual system, and 13% over joint CTC/attention training & decoding.

Background

Noisy source-channel model

- Definition: $\hat{\mathcal{W}} = \arg\max_{\mathcal{W}} P(\mathcal{W}|\mathcal{O}) = \arg\max_{\mathcal{W}} p(\mathcal{O}|\mathcal{W})P(\mathcal{W})$.
  - $\mathcal{W}$: word sequence, *i.e.* hypotheses;
  - $\mathcal{O}$: the observation sequence, *i.e.* acoustic features.
- $p(\mathcal{O}|\mathcal{W})$ – acoustic model; $p(\mathcal{W})$ – language model.
  - HMM-based acoustic models, *e.g.* GMM-HMM, DNN-HMM;
  - CTC is a special case of 2-state HMMs without prior & transition probabilities.
- Lexicon or decision tree may be required.

Attention-based model

- Definition: $P(\mathcal{C}|\mathcal{O}) = P(c_1|\mathcal{O})P(c_2|c_1,\mathcal{O})\dots P(c_L|c_1,\dots,c_{L-1},\mathcal{O})$.
  - $\mathcal{C}$: sequence of subword units of length L, $c_1,\dots,c_L$, *e.g.* characters, byte pair encoding;
  - Word sequence $\mathcal{W}$ can be directly decomposed into $\mathcal{C}$.
- Model:
$$\mathcal{E} = \textsc{Encoder}(\mathcal{O});$$
$$\mathbf{h}_i = \textsc{Attention}(\mathcal{E}, \mathbf{d}_i);$$
$$c_i, \mathbf{d}_i = \textsc{Decoder}(\mathbf{h}_i, c_{i-1}, \mathbf{d}_{i-1}).$$
- Can be viewed as RNNLM conditioned on acoustics.

Multi-task training

- Joint training of acoustic model in SC and attention-based model.
- Parameter sharing between acoustic model & neural encoder.
- Losses from SC and attention-based model are interpolated.

ISCA Framework

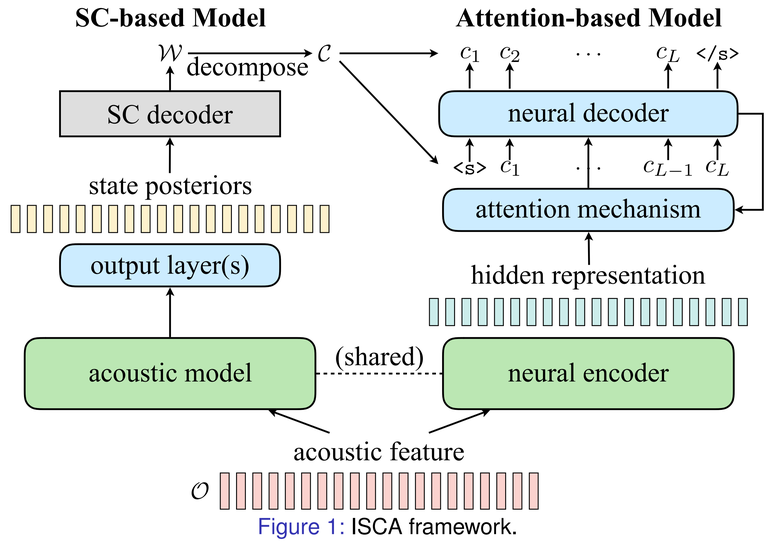

Figure 1: ISCA framework.

- Model combination at word hypothesis level.
- First pass decoding done by SC, then rescored by attention-based model during second pass.
- SC output units can be phones/graphemes with/without context.
- MAP decoding rule:
$$\hat{\mathcal{W}} \approx \arg\max_{\mathcal{W}} \left\{ \log p(\mathcal{O}|\mathcal{W}) + \alpha \log P(\mathcal{W}) + \beta \log P(\mathcal{C}|\mathcal{O}) \right\}$$
  - $\mathcal{C}$ is decomposed from word hypothesis $\mathcal{W}$ from SC;
  - α: language model scale; β: attention-based model scale.

Experimental Setup

- AMI meeting corpus (IHM), 80-D filter bank acoustic features.
- Train 80 hrs, dev 8 hrs, eval 8 hrs.
- Models trained using ESPnet, decoded using HTK.

Experimental Results

Improvements on CTC models

- CTC as a special type of HMM-based model.
- Include structured information and state prior during decoding.

vanilla CTC	47.6
+ graphemic lexicon	43.9
+ trigram LM	38.5
+ prior	33.2
+ multi-task training	32.1

Table 1: AMI dev set WER of the CTC model and several improvements.

Experimental Results (cont.)

SC subword unit & loss functions

- SC decoding uses a lexicon and a trigram LM.
- CTC decoding includes prior probability.
- No RNNLM, 20-best hypotheses used for ISCA.

subword units	loss	SC	Att.	ISCA
grapheme	CTC	32.1	31.7	28.6
monophone	CTC	28.8	31.6	26.7
	CE	28.8	31.8	26.5
triphone	CTC	28.3	31.4	26.2
	CE	**26.6**	31.6	**25.3**

Table 2: AMI dev set WERs of multi-task trained systems.

Effect of the length of *n*-best lists & RNNLM

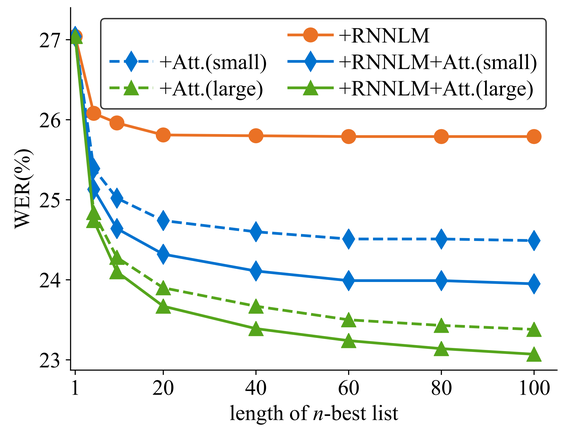

Figure 2: ISCA between a triphone/CE system and an attention-based system.

Multi-task *vs.* separate models

		#params.	dev	eval
baseline	joint CTC/att.	16.1M	28.1	29.2
separate	SC (triphone/CE)	16.0M	25.8	26.8
	Att.(small)	15.9M	30.2	31.0
	Att.(large)	84.0M	25.8	25.9
ISCA	multi-task	17.3M	24.4	25.4
	SC + Att.(small)	31.9M	24.0	24.5
	SC + Att.(large)	100.0M	**23.1**	**23.8**

Table 3: Results on dev & eval sets with an RNNLM using 100-best.

Conclusions

- SC and attention-based systems are highly complementary.
- ISCA is a simple, flexible & efficient framework for combination.

E-Mail: {ql264,cz277,pcw}@eng.cam.ac.uk