BRIDGENETS: STUDENT-TEACHER TRANSFER LEARNING BASED ON RECURSIVE NEURAL NETWORKS AND ITS APPLICATION TO DISTANT SPEECH RECOGNITION

Jaeyoung Kim, Mostafa El-Khamy, Jungwon Lee
Samsung Semiconductor Inc., San Diego, California, USA, 92121

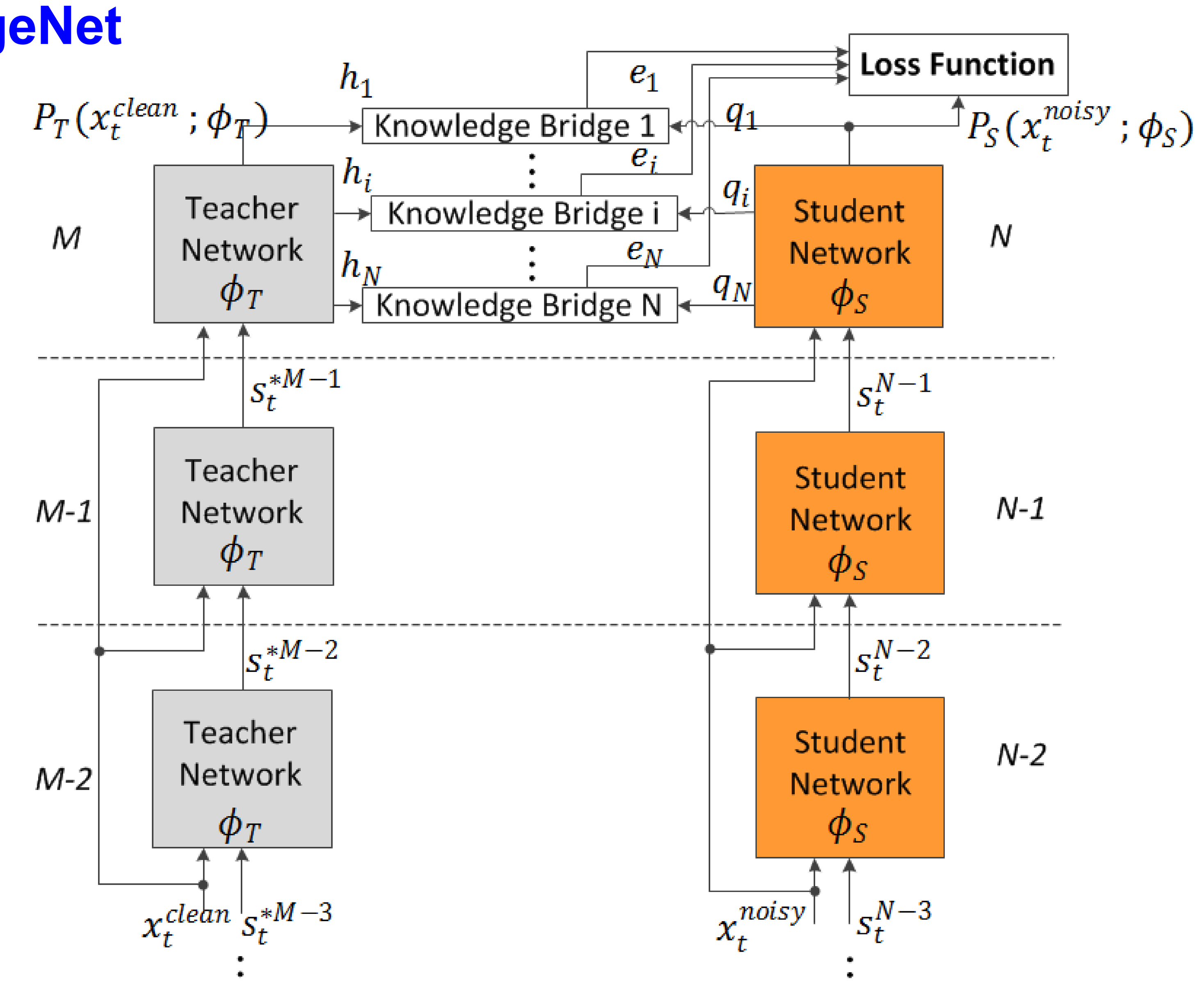
Distant Speech Recognition (DSR)

- DSR is to recognize human speeches in the presence of various noise sources caused by the large distance between speakers and microphones.
- Traditional speech recognizers trained with clean data often fail to recognize due to signal quality mismatch between training and test environment.

Main Contribution

- Proposed a new student-teacher paradigm for DSR: **BridgeNet**
- BridgeNet provides teacher's intermediate features as additional hints, which can properly regularize a student network.
- Proposed a new recursive architecture that can iteratively improve signal denoising and recognition in BridgeNet

BridgeNet



- Knowledge bridges (hints) provide an error measure to guide intermediate feature representation of a student network:

$$e_i(\phi_S) = \sum_{t=1}^{L} \left\| h_i(x_t^{clean}) - q_i(x_t^{noisy}; \phi_S) \right\|^2 \; for\ i = 2, \dots, N$$

$$e_1(\phi_S) = \sum_{t=1}^{L} \left(P_T(x_t^{clean}; \phi_T)\right)^T \log P_S(x_t^{noisy}; \phi_S) \quad for\ i = 1$$

- Loss function is a weighted sum of all error measures:

$$L(\phi_S) = \sum_{i=1}^{N} \alpha_i e_i(\phi_S) + \sum_{t=1}^{L} \left(P_T(y_t^{label})\right)^T \log P_S(x_t^{noisy}; \phi_S)$$

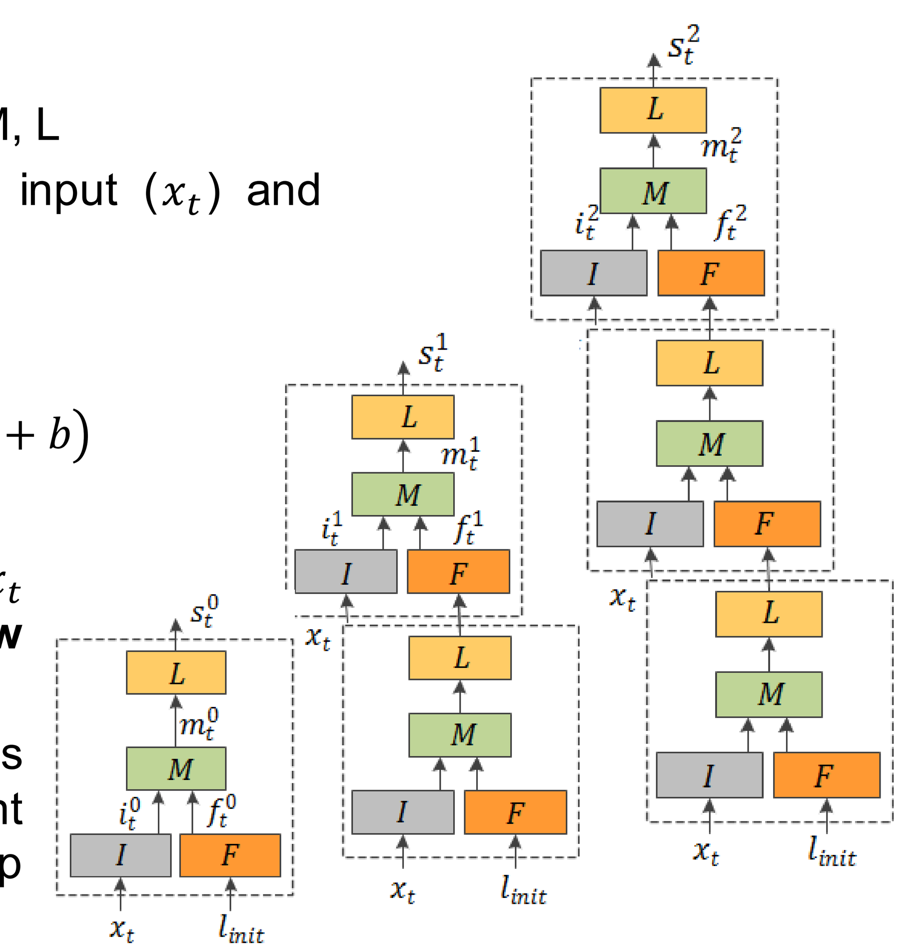
where the second term is the cross-entropy between label and student softmax output.

Recursive Architecture

- Composed of four sub-blocks: I, F, M, L
- I and F take two inputs: acoustic input (x_t) and output (s_t^{n-1}) from prior recursion.
- M merges two independent paths.

$$m_t^n = g\left(W_1 i_t^n(x_t) + W_2 f_t^n(s_t^{n-1}) + b\right)$$

- For each new recursion, the same x_t is fed into I, which acts as a **new global shortcut path**.
- The global shortcut paths act as highway paths that facilitate gradient flows. → helps to have deep recursive architecture.



(a) R=0 (b) R=1 (c) R=2

Unrolling of a Recursive Network

Existing Approaches for DSR

- **Multi-task denoising (MTD):**
  - Jointly optimize denoising (DE) and recognition (RE) subnetworks integrated within the unified neural network.
  - Minimizing MSE between raw acoustic data and high-level abstracted features in MTD is often unsuccessful.

→ BridgeNet provides the similar high-level features to guide a student network.

- **Knowledge distillation (KD):**
  - Transfer the generalization ability of a bigger teacher network to a typically much smaller student network.

- **Generalization distillation (GD):**
  - Extend KD by training a teacher network with parallel clean data in order to apply it to signal denoising.
  - GD improved ASR. However, utilization of parallel data is too limited.

→ BridgeNet provides multiple hints from teacher's intermediate layers.

Main Result

- BridgeNet presented 5.29% accuracy improvements over the baseline CNN-LSTM model on AMI corpus.
- Compared with KD, it showed 2.72% relative WER reduction.
- Recursive architecture further improved BridgeNet: 13.24% improvement of relative WER over CNN-LSTM, 10.88% over KD.

Experiments

Multi-Task Denoising on AMI SDM corpus: CNN-LSTM* is trained with clean alignment. Rest of them used noisy alignment

Acoustic Model	WER(all)	WER (main)
DNN	59.1%	50.5%
DNN, denoised	58.7%	50.2%
CNN-LSTM	50.4%	41.6%
CNN-LSTM, denoised	50.1%	41.4%
CNN-LSTM*	46.5%	37.7%
CNN-LSTM*, denoised	46.9%	38.2%

- CNN-LSTM is our baseline model: two layers of CNN layers are stacked with 3 layers of LSTM. DNN model has 8 layers.
- Multi-task denoising showed marginal improvement for DNN and CNN-LSTM.
- CNN-LSTM using clean alignment showed degradation with MTD.

BridgeNet: single channel SDM corpus is used for training a student network

Acoustic Model	WER(all)	WER (main)
CNN-LSTM(baseline), R0	46.5%	37.7%
KD, R0	44.8%	35.7%
KD+DR, R0	44.1%	35.3%
KD+DR+LSTM3, R0	44.0%	35.1%
CNN-LSTM(baseline), R2	45.8%	36.9%
KD, R1	43.7%	34.7%
KD+DR, R1	43.4%	34.7%
KD+DR+LSTM3, R1	42.6%	33.8%

BridgeNet: 8-channel beamformed MDM corpus is used for training a student network

Acoustic Model	WER(all)	WER (main)
CNN-LSTM(baseline), R0	43.4%	34.0%
KD, R0	42.8%	33.1%
KD+DR, R0	42.3%	32.5%
KD+DR+LSTM3, R0	41.8%	32.2%
CNN-LSTM(baseline), R2	43.0%	33.3%
KD, R1	40.4%	30.8%
KD+DR, R1	39.5%	29.9%
KD+DR+LSTM3, R1	39.3%	29.5%

- KD, DR and LSTM3 are knowledge bridges between student and teacher networks.
- Each added bridge incrementally improves BridgeNet: KD+DR+LSTM3 provided 6.9% gain over CNN-LSTM and 1.6% gain over KD.
- BridgeNet with recursion presented huge gain: 13.24% and 10.88% WER reduction over CNN-LSTM and KD.