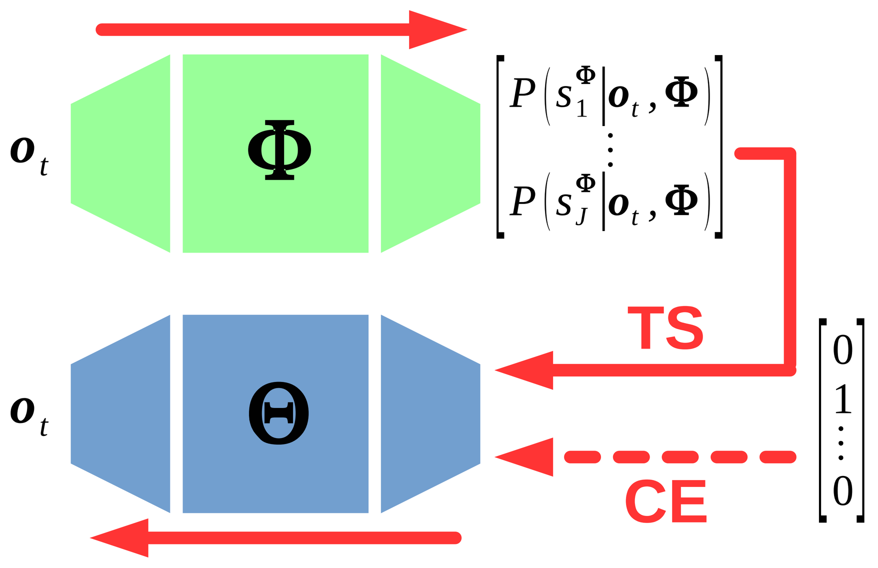
ASR teacher-student training and ensemble target diversity

Jeremy H. M. Wong and Mark J. F. Gales
Department of Engineering, University of Cambridge
jhmw2@cam.ac.uk, mjfg@eng.cam.ac.uk

MOTIVATION

- An ensemble with target diversity can give good combination gains.
- Improve recognition efficiency by training a student to emulate the ensemble.
- How to propagate information across different output targets?

TEACHER-STUDENT TRAINING



- Train a single student model to emulate the ensemble behaviour.
- **Standard CE training:**

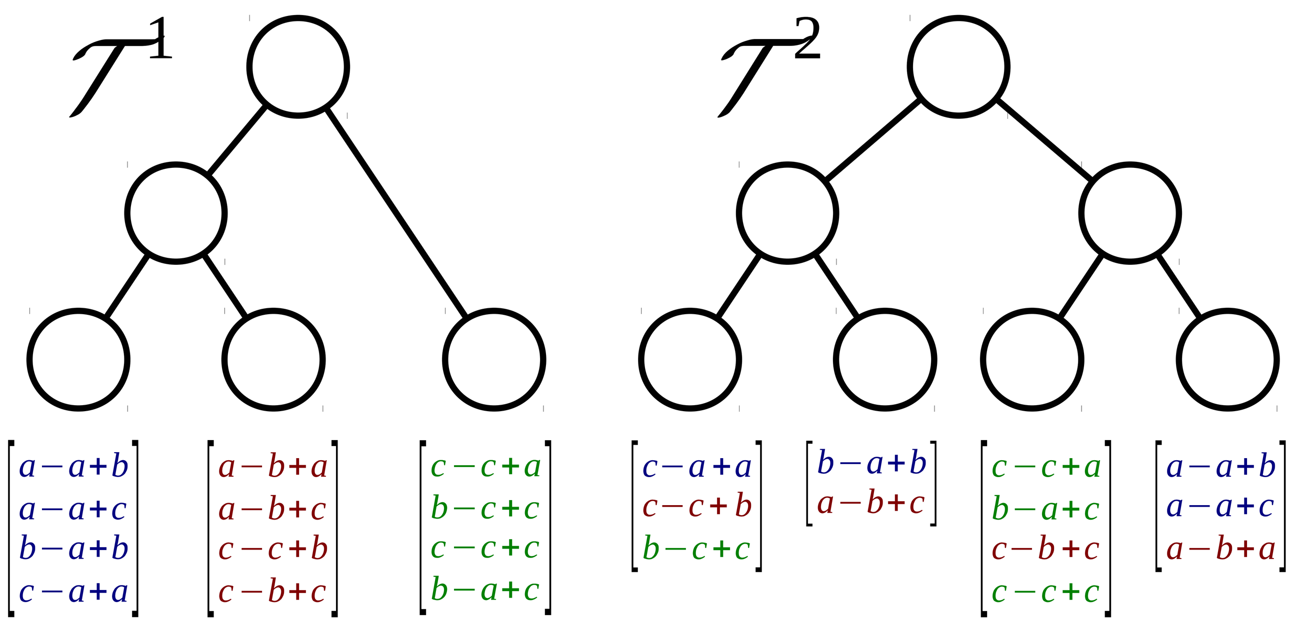
$$\mathcal{F}_{\text{CE}} = -\sum_t \sum_{s \in \mathcal{T}} \delta\left(s, s_t^*\right) \log P\left(s | \boldsymbol{o}_t, \boldsymbol{\Theta}\right)$$

- **Standard Teacher-Student (TS) training:**

$$\mathcal{F}_{\text{TS}} = -\sum_t \sum_{s \in \mathcal{T}} \sum_{m=1}^{M} \lambda_m P\left(s | \boldsymbol{o}_t, \boldsymbol{\Phi}^m\right) \log P\left(s | \boldsymbol{o}_t, \boldsymbol{\Theta}\right)$$

- Use only student model during recognition.
- Requires student's and teacher's outputs to have the same interpretations.

OUTPUT TARGET DIVERSITY



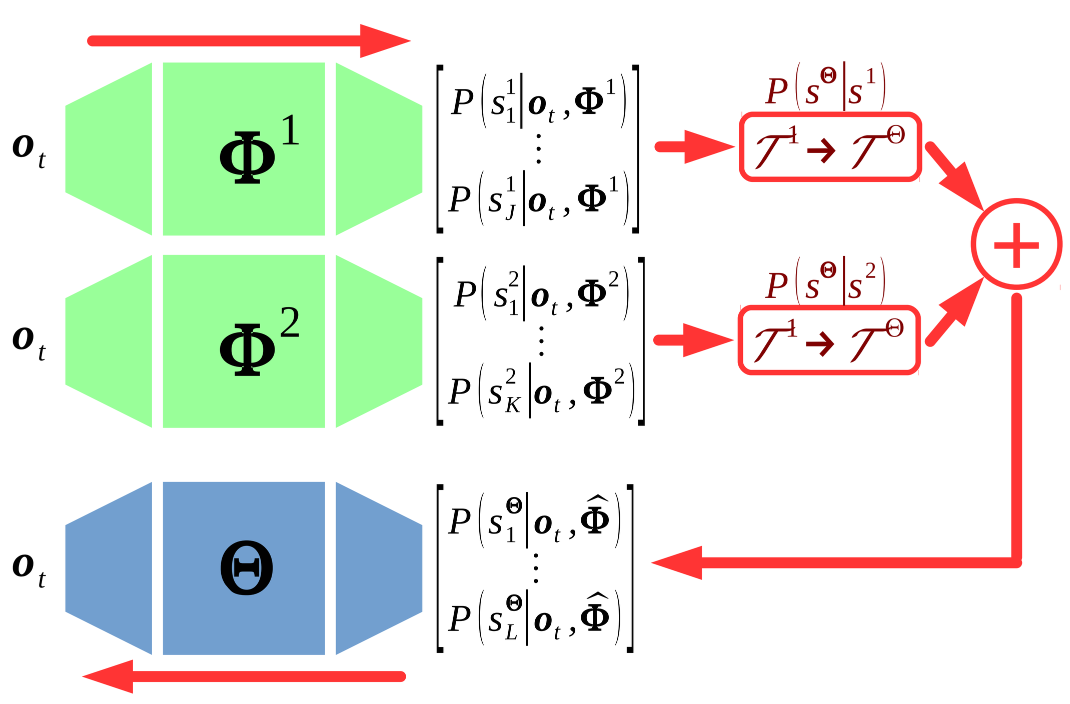
- Output targets are defined by a Phonetic Decision Tree (PDT)

$$s_c = \mathcal{T}(c)$$

- Generate ensemble by using a different PDT for each model.
- Models learn to discriminate between different sets of state clusters.
- Computational cost of ensemble combination:

	NN forward	lattice decode
hypothesis combine	M	M
frame combine	M	1
teacher-student	1	1

POSTERIOR MAPPING



- When PDTs differ, train student by minimising logical context KL-divergence:

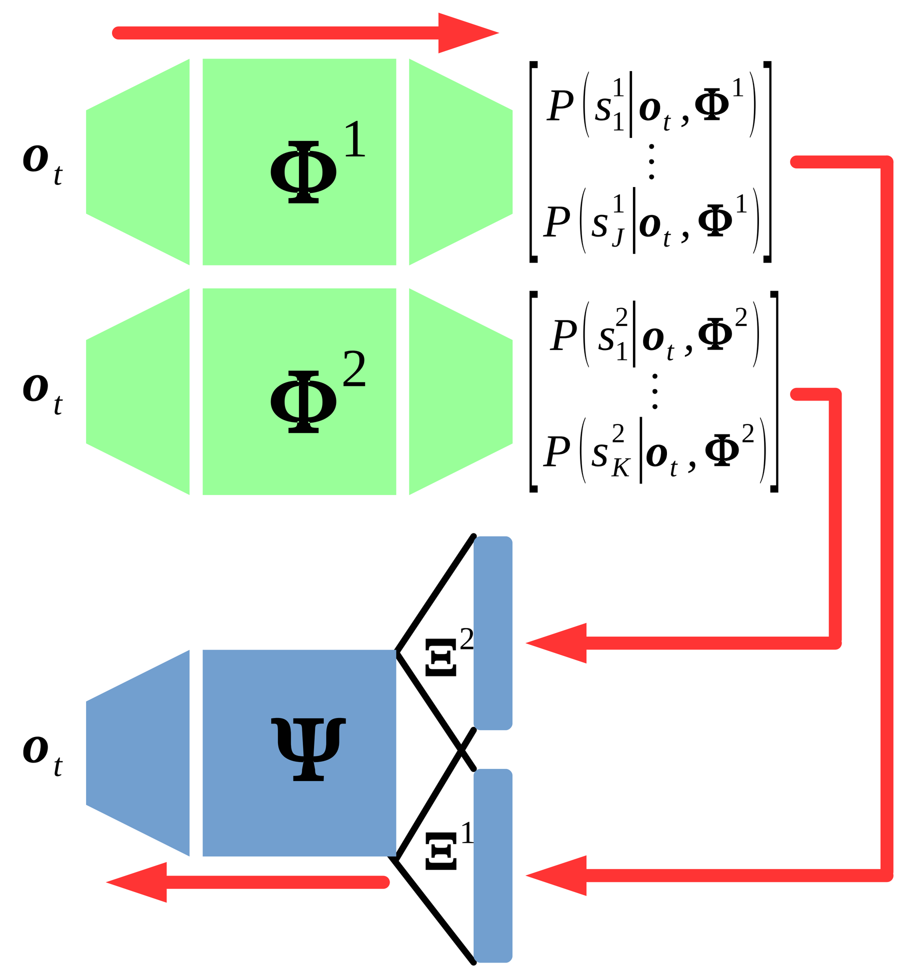
$$\mathcal{F}_{\text{RF-TS}} = -\sum_t \sum_{c \in \mathcal{C}} \sum_{m=1}^{M} \lambda_m P\left(c | \boldsymbol{o}_t, \boldsymbol{\Phi}^m\right) \log P\left(c | \boldsymbol{o}_t, \boldsymbol{\Theta}\right)$$

- Under mild assumptions, the criterion reduces to:

$$\mathcal{F}_{\text{RF-TS}} = -\sum_t \sum_{s^{\Theta} \in \mathcal{T}^{\Theta}} \sum_{m=1}^{M} \lambda_m \sum_{s^m \in \mathcal{T}^m} P\left(s^{\Theta} | s^m\right) P\left(s^{\Theta} | \boldsymbol{o}_t, \boldsymbol{\Phi}^m\right) \log P\left(s^{\Theta} | \boldsymbol{o}_t, \boldsymbol{\Theta}\right)$$

- $P\left(s^{\Theta} | s^m\right)$ maps posteriors between PDTs.
- Can estimate $P\left(s^{\Theta} | s^m\right)$ from forced alignments.
- Student PDT size can be chosen independently of teacher PDTs.

MULTI-TASK ARCHITECTURE



- Avoid mapping by using Multi-Task (MT) student.
- **Multi-task CE training:**

$$\mathcal{F}_{\text{MT}} = -\sum_t \sum_{m=1}^{M} \sum_{s^m \in \mathcal{T}^m} \delta\left(s^m, s_t^{m*}\right) \log P\left(s^m \middle| \boldsymbol{o}_t, \boldsymbol{\Psi}, \hat{\boldsymbol{\Xi}}\right)$$

- **Multi-task teacher-student training:**

$$\mathcal{F}_{\text{MT-TS}} = -\sum_t \sum_{m=1}^{M} \sum_{s^m \in \mathcal{T}^m} P\left(s^m | \boldsymbol{o}_t, \boldsymbol{\Phi}^m\right) \log P\left(s^m \middle| \boldsymbol{o}_t, \boldsymbol{\Psi}, \hat{\boldsymbol{\Xi}}\right)$$

EXPERIMENTS

- Datasets:
  - **207V:** IARPA Babel Tok Pisin
    - 3 hours VLLP training set, 1000 PDT states
  - **AMI:** Augmented multi-party interaction
    - 81 hours IHM training set, 4000 PDT states
  - **HUB4:** English broadcast news
    - 144 hours training set, 6000 PDT states
- **Ensemble size =** 4
- Student and teachers have the same architecture.

SINGLE MODEL PERFORMANCE

	Single model WER (%)				
Dataset	mean	best	worst	std dev	cross-WER (%)
207V	48.3	48.0	48.4	0.17	28.4
AMI	26.0	25.9	26.2	0.13	15.2
HUB4	9.3	9.2	9.4	0.10	7.0

- Measure diversity using cross-WER:

$$\text{cross-WER} = \frac{1}{M(M-1)} \sum_{m=1}^{M} \sum_{n \neq m} \text{WER}\left(\boldsymbol{\mathcal{H}}^m, \boldsymbol{\mathcal{H}}^n\right)$$

ENSEMBLE PERFORMANCE

		Combined WER (%)		
Dataset	ensemble	hypothesis	frame	student
207V	separate	45.8	46.0	46.6
	MT	47.7	47.8	47.3
	MT-TS	45.7	45.7	46.3
AMI	separate	24.5	24.6	24.6
	MT	25.4	25.5	25.1
	MT-TS	24.3	24.4	24.6
HUB4	separate	8.7	8.7	9.0
	MT	9.1	9.1	8.8
	MT-TS	8.8	8.7	8.9

- Single-output student can learn from teachers with different PDTs.
- Multi-task student is able to match the ensemble performance.

CONCLUSIONS

- Proposed teacher-student method when output targets differ.
- Proposed multi-task teacher-student method.

REFERENCES

[1] J. Wong and M. Gales, "Student-teacher training with diverse decision tree ensembles", *Interspeech*, Sep 2017

[2] J. Wong and M. Gales, "Multi-task ensembles with teacher-student training", *ASRU*, Dec 2017