Mutual Knowledge Distillation Between Heterogeneous Architectures for Continuous Sign Language Recognition

Michał Kalinowski and Bożena Kostek

Gdańsk University of Technology, Gdańsk, Faculty of Electronics, Telecommunications and Informatics, Multimedia Systems Department and Audio Acoustics Laboratory, Poland

Objectives

- Test whether **mutual knowledge distillation** between CorrNet [1] and SlowFastSign [2] lowers WER on PHOENIX14.
- Find what makes fine-tuning pretrained CSLR models **stable**.

Introduction

Continuous sign language recognition (CSLR) aims to transcribe sign language videos into sequences of glosses. The task remains challenging, partly due to the limited availability of large and diverse annotated datasets, making it difficult to train new architectures from scratch. This motivates approaches that leverage knowledge already acquired by well-performing models.

In this work, we investigate mutual knowledge distillation between two structurally distinct CSLR architectures, CorrNet [1] and SlowFastSign [2], both selected for their strong performance on the PHOENIX14 dataset. We hypothesize that these heterogeneous architectures learn complementary representations from the same visual input and can therefore benefit one another through bidirectional knowledge transfer. To explore this, both networks are optimized simultaneously within a shared training framework. Alongside their original objectives, each model receives a distillation signal derived from the other model's output probability distribution. In this way, the soft predictions of each architecture guide the other without requiring identical internal representations or additional annotated data.

GDAŃSK UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRONICS, TELECOMMUNICATIONS AND INFORMATICS

Challenges in CSLR

- **Weak supervision**: only sentence-level labels are available [1].
- Signs combine **manual** (hand/arm) and **non-manual** (face, head, body) cues [1].
- Recognition requires **spatially** meaningful gestures (hand direction, finger and mouth shape) and **dynamically** meaningful gestures (arm and body movement) [2].
- Common limitation: **frame-by-frame processing** misses local temporal patterns [1, 2].

Architectures

- **CorrNet** [1]: Combines 2D CNN, 1D CNN and BiLSTM components. **Correlation** and **identification** modules link adjacent frames to track **hands and face** while suppressing background information.
- **SlowFastSign** [2]: Employs **Slow** (spatial detail) and **Fast** (motion) pathways with bidirectional feature fusion, followed by a BiLSTM.
- Both models use **CTC** and VAC-style **visual supervision** to produce gloss sequences.

Results

- Both models were fine-tuned on the PHOENIX14 **train** split and evaluated on the **test** split (WER, beam decoder).
- Each distillation run is compared with a control trained identically, without distillation.
- **Ensemble:** the authors' two checkpoints, softmax probabilities averaged per timestep with equal weights, decoded once; no training.

Test WER (%)	CorrNet	SlowFastSign
Authors' checkpoints (pretrained)	19.4	18.3
Without distillation (control)	19.08	17.91
Mutual distillation	**19.01**	**17.71**
Ensemble of both (no training)	17.78	

The Proposed Pipeline

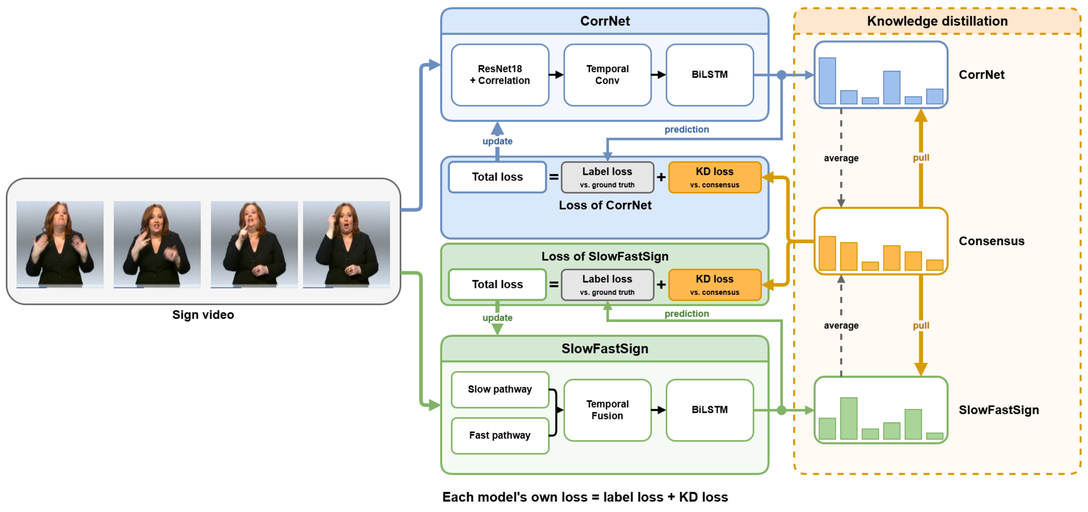

Figure 1: Mutual knowledge distillation between two continuous sign language recognition models, CorrNet [1] and SlowFastSign [2]. Both models process the same input video and produce frame-wise gloss probability distributions. These are averaged into a consensus distribution, which serves as a shared teacher. Each model is trained with its own loss, the sum of a label loss (CTC against the ground-truth glosses) and a distillation loss that pulls its predictions toward the consensus.

The Proposed Method

- **Setup:** PHOENIX14; both models start from the **authors' checkpoints**, with a shared gloss dictionary and decoder.
- **Joint training:** both models receive the **same batch** in one loop; each has its own optimizer and own loss (CTC with visual supervision).
- **Distillation:** at every timestep, each model also learns from the **average prediction of both models**, treated as a fixed target.
- **Combined loss:** $\mathcal{L} = (1-\alpha)\,\mathcal{L}_{\text{own}} + \alpha\,\mathcal{L}_{\text{KD}}$, with $\alpha = 0.1$; $\alpha = 0$ is the control without distillation.

Stabilizing the Training

- **Frozen BatchNorm** (main fix): training on single samples corrupted the pretrained normalization statistics.
- **Gradient accumulation and clipping:** one update per 8 samples, with capped size, so no single sample disturbs the pretrained weights.
- **Separate mixed-precision scalers:** each model's half-precision training is scaled independently.
- **Normalized distillation loss:** keeps it on the same scale as the CTC losses.

Conclusion

- **Mutual distillation:** Improves WER for **both** models (CorrNet [1]: 19.01, SlowFastSign [2]: 17.71), outperforming controls and pretrained checkpoints.
- **Training stability:** Freezing BatchNorm was essential; without it, both models degraded even without distillation.
- **Limited headroom:** The pretrained ensemble achieves 17.78 WER, suggesting limited model complementarity. Distilled SlowFastSign [2] matches it using a single model.
- **Limitations:** Gains over controls are modest (<0.2 WER) and close to seed-to-seed variation.

Future directions:

- Evaluate more seeds and report mean ± standard deviation.
- Explore selective distillation via agreement- or confidence-based timestep gating.

References

[1] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network, 2023.

[2] Junseok Ahn, Youngjoon Jang, and Joon Son Chung. Slowfast network for continuous sign language recognition, 2023.