UPRet: Uncertainty-aware Sign Language Video Retrieval with Probability Distribution Modeling (Supplementary Materials)

Xuan Wu*[1], Hongxiang Li*[2], Yuanjiang Luo[3], Xuxin Cheng[2], Xianwei Zhuang[2], Meng Cao[4], and Keren Fu†[1,3]

[1] College of Computer Science, Sichuan University
[2] School of Electronic and Computer Engineering (SECE), Peking University
[3] National Key Lab of Fundamental Science on Synthetic Vision, Sichuan University
[4] Mohamed bin Zayed University of Artificial Intelligence

Overview

In this supplementary material, we present the following.

- Proof of Optimal Transport
- More Visualization Cases
- Detail of Datasets

A Proof of Optimal Transport

The theoretical proof for optimal transport (OT) generally involves showing that the transport plan minimizes the cost of transporting mass from one distribution to another. Mathematically, given a distribution μ representing the video and another distribution ν representing the text. We define μ and ν as two probability distributions over metric spaces X and Y respectively, where X represents the space of video features and Y represents the space of text features. The goal of OT is to find a transport plan $\pi \in \Pi(\mu, \nu)$, which is a joint distribution over $X \times Y$ with marginals μ and ν, that minimizes the total cost of transporting mass from μ and ν. The cost function $c : X \times Y \to \mathbb{R}$ represents the cost of transporting a unit mass from X to Y. The OT problem can be formally stated as finding a transport plan π that minimizes the following cost:

$$\mathrm{OT}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y) d\pi(x, y), \tag{1}$$

where $\Pi(\mu, \nu)$ is the set of all joint distributions π whose marginals are μ and ν. In a discrete setting, μ and ν can be represented as vectors of probabilities that

* Equal contributions.
† Corresponding author (fkrsuper@scu.edu.cn)

sum to 1, and the cost function c as a matrix where c_{ij} is the cost of transporting mass from point i in X to point j in Y.

$$\mu_v = \sum_{i=1}^{N_v} p_i^v \delta\left(v_i\right), \quad \mu_t = \sum_{i=1}^{N_t} p_i^t \delta\left(t_i\right), C(v_i, t_i) = 1 - \frac{v_i^\top t_i}{||\boldsymbol{v_i}|| \cdot ||\boldsymbol{t_i}||}. \tag{2}$$

Then,the OT problem can be reformulated as a linear programming problem:

$$\mathrm{OT}(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^{m} \sum_{j=1}^{n} \pi_{ij} c_{ij}, \tag{3}$$

$$\text{subject to:} \quad \sum_j \pi_{ij} = \mu_i \quad \forall i, \quad \sum_i \pi_{ij} = \nu_j \quad \forall j, \quad \pi_{ij} \geq 0 \quad \forall i, j \tag{4}$$

This problem can be tackled using a fast iterative approach that reformulates the optimisation objective in Eq. 3 into a convex but non-linear configuration by adding an entropic regularisation term E:

$$\min_{\pi} \quad \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} \pi_{ij} + \gamma E(\pi_{ij}), \tag{5}$$

where $E(\pi_{ij}) = \pi_{ij}(\log \pi_{ij} - 1)$ is the negative entropy regularization and $\gamma \geq 0$ is a regularization hyper-parameter. Then we can have a fast optimization solution with a few iterations as:

$$\boldsymbol{T}^* = \mathrm{diag}(\mu_t^i) \exp(-\boldsymbol{E}/\eta) \mathrm{diag}(\mu_v^i), \tag{6}$$

According to Lagrange Multiplier Method, the constraint optimization target in Eq. 5 can be convert to a nonconstraint target:

$$\min_{\pi} \quad \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} \pi_{ij} + \gamma E(\pi_{ij}) + \alpha_j (\sum_{i=1}^{m} \pi_{ij} - d_j) + \beta_i (\sum_{j=1}^{n} \pi_{ij} - s_i), \tag{7}$$

where $\alpha_j (j = 1, 2, \ldots, n)$ and $\beta_i (i = 1, 2, \ldots, m)$ are Lagrange multipiers. By letting the derivatives of the optimization target equal to 0, the optimal plan π^* is resolved as:

$$\pi_{ij}^* = \exp(-\frac{\alpha_j}{\gamma}) \exp(-\frac{c_{ij}}{\gamma}) \exp(-\frac{\beta_i}{\gamma}). \tag{8}$$

Letting $u_j = \exp(-\frac{\alpha_j}{\gamma})$, $v_i = \exp(-\frac{\beta_i}{\gamma})$, $M_{ij} = \exp(-\frac{c_{ij}}{\gamma})$, the following constraints can be enforced:

$$\sum_i \pi_{ij} = u_j (\sum_i M_{ij} v_i) = d_j, \quad \sum_j \pi_{ij} = (u_j \sum_i M_{ij}) v_i = s_i. \tag{9}$$

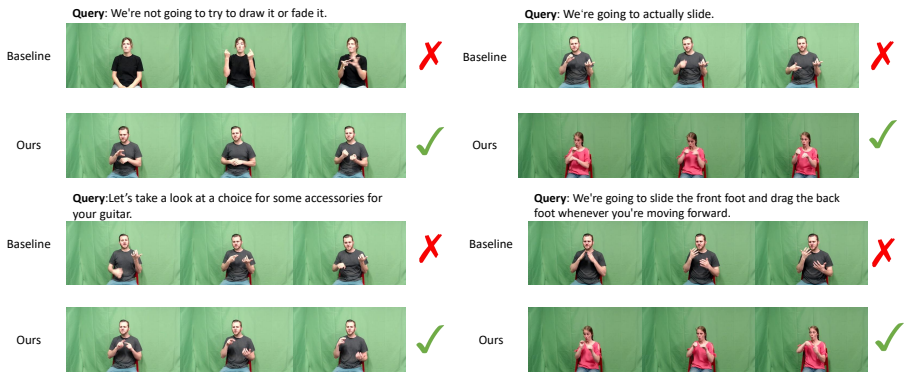

Fig. 1: Visualization of the text-sign video output on the How2Sign. Red: incorrect results of the baseline model. Green: correct results of our method.

To fulfil these two equations simultaneously, one approach is to iteratively update v_i and v_j using the following update rules for a sufficient number of iterations:

$$u_j^{t+1} = \frac{d_j}{\sum_i M_{ij} v_i^t}, \quad v_i^{t+1} = \frac{s_i}{\sum_j M_{ij} u_j^{t+1}}. \tag{10}$$

The update mechanism described in Eq. 10 is also referred to as the Sinkhorn-Knopp Iteration. By executing this iterative process for T iterations, one can derive an approximate solution for the optimal plan π^*:

$$\boldsymbol{\pi}^* = \mathrm{diag}(\mu_t^i) \exp(-\boldsymbol{E}/\eta) \mathrm{diag}(\mu_v^i), \tag{11}$$

B More Visualization Cases

To enhance the understanding of the performance of our model, we present further visualization cases in the supplementary materials. These visualizations provide an insightful complement to the quantitative results discussed in the main text and exemplify the practical application of UPRet in video-text retrieval scenarios, as shown in Figure 1.

C Details of Datasets

How2Sign [2], a comprehensive multimodal library of American Sign Language (ASL), encompasses over 80 hours of videos across 10 categories reflecting daily life scenarios. It includes 31,164 training, 1,740 validation, and 2,356 test videos. **PHOENIX-2014T** [1] features sign language videos from German Public Television's weather forecasts, offering video-to-text translations and annotated timelines. It comprises 7,096 training, 519 validation, and 642 test video-text pairs.

CSL-Daily [3] centered on Chinese Sign Language (CSL) for daily communication, features diverse expressions and phrases across various everyday topics. It includes 18,401 training, 1,077 validation, and 1,176 test sample pairs. These datasets each have unique features and difficulties that make them ideal for measuring the state-of-the-art performance of sign language video recognition and translation techniques.

References

1. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7784–7793 (2018)
2. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., Giro-i Nieto, X.: How2sign: a large-scale multimodal dataset for continuous american sign language. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2735–2744 (2021)
3. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: Improving sign language translation with monolingual data by sign back-translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1316–1325 (2021)