

GENERAL SKELETON SEMANTICS LEARNING WITH PROBABILISTIC MASKED CONTEXT RECONSTRUCTION FOR SKELETON-BASED PERSON RE-IDENTIFICATION – APPENDIX II ASSUMPTIONS AND ANALYSES

Anonymous authors

Paper under double-blind review

APPENDIX OUTLINE

The overview for this appendix is presented as follows.

- In Sec. A, we offer a general computing formula for occurrence probabilities of different sub-tasks contained in Prompter.
- In Sec. B, we provide simplified theoretical assumptions and analyses of Prompter on potential model regularization.

A SUB-TASK TRANSFORMATION OF PROMPTER

With IID masks sampled from the Bernoulli distribution, the probability of occurrence for spatially *retaining* n_S body-joint locations can be computed by

$$\mathcal{P}_S(n_S) = \binom{J}{n_S} (p_S)^{J-n_S} (1 - p_S)^{n_S}, \quad (1)$$

where $\binom{J}{n_S} = \frac{J!}{n_S!(J-n_S)!}$ and $\mathcal{P}_S(n_S)$ denotes the probability of keeping spatial locations of arbitrary n_S joints. The spatial masked skeleton context representation \tilde{v}_t (see Eq. (7) of the paper) is essentially a random combination of different numbers ($1 \leq n_S \leq J$) of body-joint representations with the probability $\mathcal{P}_S(n_S)$, which keeps each joint with an *independent probability* $(1 - p_S)$ and leverages partial spatial context of skeletons to perform location reconstruction and inference.

Likewise, the masked temporal context representation \bar{w}^i (see Eq. (8) of the paper) can be viewed as randomly sampling different sub-trajectories of the same joint, where each temporal position is *discarded* with the same independent probability p_T and the probability for keeping n_T temporal positions can be computed as

$$\mathcal{P}_T(n_T) = \binom{f}{n_T} (p_T)^{f-n_T} (1 - p_T)^{n_T}. \quad (2)$$

The proposed Prompter with task transformability (TT) can be viewed as a general probabilistic form of existing reconstruction or masked reconstruction based SSL tasks (Rao et al., 2021; Rao & Miao, 2023): (1) Prompter contains a sub-task, direct spatial reconstruction (*i.e.*, all body-joint locations are unmasked), and the probability of performing this sub-task is $\mathcal{P}_S(J) = (1 - p_S)^J$ (see Eq. (1)); (2) The second contained sub-task is the masked spatial skeleton reconstruction, with the performing probability of $\mathcal{P}_S(n_S) = \binom{J}{n_S} (p_S)^{J-n_S} (1 - p_S)^{n_S}$ when n_S locations of joints are unmasked; (3) Prompter can also be transformed to the masked temporal skeleton reconstruction under the probability of $\mathcal{P}_T(n_T) = \binom{f}{n_T} (p_T)^{f-n_T} (1 - p_T)^{n_T}$ in the case of n_T trajectory positions of a joint are unmasked. The task transformability enables Prompter to jointly optimize different SSL sub-tasks and achieve better semantics learning performance (see Sec. 4.2 and 4.4 in the paper).

Intuitively, Prompter introduces more possible spatial-temporal reconstruction cases (*i.e.*, under varying partial spatial and temporal contexts) than both direct reconstruction and masked reconstruction Rao & Miao (2023) that employs a *fixed* number of masks, thereby potentially improving the reconstruction diversity and reducing model over-fitting. It can also be viewed as a special

representation-level Dropout (Baldi & Sadowski, 2014) to randomly drop joint representations over spatial and temporal dimensions. We provide a simplified case to show that both Prompter and Dropout can introduce random perturbations and similar model regularization in Sec. B.

B MODEL REGULARIZATION VIA PROMPTER

The Prompter objective ($\mathcal{L}_{\text{SSCR}}$) can be modeled as an equivalent objective containing the model regularization (*i.e.*, ℓ_2 weight regularization). We discuss and analyze the case of performing probabilistic spatial skeleton context reconstruction with Prompter under single linear units (which can be generalized to the probabilistic temporal skeleton context reconstruction and non-linear units).

Preliminaries. Different spatial context masking strategies (*i.e.*, combinations of unmasked joints) are assumed to correspond to different *sub-model* learning and error optimization. For example, only masking the left hand body joint or only masking knee joints to construct the masked skeleton context representation for training can lead to different learned models, which can be viewed as different *sub-models* of the original model trained with all joints unmasked. For clarity and convenience, we adopt a more general notation here, which is different from that used in the paper. We use Err_{Ens} to denote the error function of the ensemble of all possible sub-models corresponding to different spatial context masking strategies, and use Err_{Pro} to represent the error function of the model using Prompter. With the mean square error (MSE) as the error metric for reconstruction, we can define them as:

$$Err_{\text{Ens}} = \frac{1}{2} (t - O_{\text{Ens}})^2 = \frac{1}{2} \left(t - \sum_{i=1}^n w_i I_i \right)^2, \quad (1)$$

where

$$I_i = \sum_{k=1}^J p_k v_i^k. \quad (2)$$

In Eq. (1) and Eq. (2), I_i denotes the i^{th} element of the input vector $I \in \mathbb{R}^n$, w_i represents the i^{th} element of the learnable weight vector $w \in \mathbb{R}^n$, t denotes the target value (*i.e.*, single value corresponding to a specific dimension of ground-truth joint positions), O_{Ens} is the expected output value of all possible sub-models with the probability p_k for masking the k^{th} body-joint location representation $v^k \in \mathbb{R}^n$, v_i^k denotes the i^{th} element of v^k , and J is the total number of body joints. In Eq. (1), we compute the single-position error using a single training input I , while the error of each training example can be combined additively for sequence reconstruction. Likewise, the error function for the model applying Prompter can be formulated as:

$$Err_{\text{Pro}} = \frac{1}{2} (t - O_{\text{Pro}})^2 = \frac{1}{2} \left(t - \sum_{i=1}^n w_i \bar{I}_i \right)^2, \quad (3)$$

where

$$\bar{I}_i = \sum_{k=1}^J \delta_k v_i^k. \quad (4)$$

Here O_{Pro} denotes the output value of model when using Prompter, \bar{I}_i denotes the i^{th} element of the input vector $\bar{I} \in \mathbb{R}^n$, which is generated by masking body-joint location representation v^k with the gating 0-1 Bernoulli variable δ_k and $P(\delta_k = 1) = p_k$ in Eq. (4). The variable δ_k is assumed to be independent of each other, independent of the weights, and independent of the model optimization. By taking over all possible gating variables in different sub-models, the expectation (*i.e.*, ensemble average) of input vector can be computed by $I_i = \sum_{k=1}^J p_k v_i^k$, as shown in Eq. (2).

Based on Eq. (1) and Eq. (3), the learning gradient of the errors with regard to w_i are computed as:

$$\begin{aligned} \frac{\partial Err_{\text{Ens}}}{\partial w_i} &= -(t - O_{\text{Ens}}) \frac{\partial O_{\text{Ens}}}{\partial w_i} = -(t - O_{\text{Ens}}) I_i = -t I_i + w_i I_i^2 + \sum_{j \neq i} w_j I_i I_j \\ &= -t \sum_{k=1}^J p_k v_i^k + w_i \left(\sum_{k=1}^J p_k v_i^k \right)^2 + \sum_{j \neq i} w_j \left(\sum_{k=1}^J v_i^k v_j^k (p_k)^2 + \sum_{a=1, b \neq a}^J 2 v_i^a v_j^b p_a p_b \right), \quad (5) \end{aligned}$$

$$\frac{\partial Err_{\text{Pro}}}{\partial w_i} = -(t - O_{\text{Pro}}) \frac{\partial O_{\text{Pro}}}{\partial w_i} = -(t - O_D) \bar{I}_i = -t\bar{I}_i + w_i \bar{I}_i^2 + \sum_{j \neq i} w_j \bar{I}_i \bar{I}_j. \quad (6)$$

As the gradient of model using Prompter is a random variable, we take its expectation with:

$$\mathbb{E} \left[\frac{\partial Err_{\text{Pro}}}{\partial w_i} \right] = -t \mathbb{E} [\bar{I}_i] + w_i \mathbb{E} [\bar{I}_i^2] + \sum_{j \neq i} w_j \mathbb{E} [\bar{I}_i \bar{I}_j], \quad (7)$$

where

$$\mathbb{E} [\bar{I}_i] = \sum_{k=1}^J p_k v_i^k, \quad (8)$$

$$\text{D} [\bar{I}_i] = \sum_{k=1}^J (v_i^k)^2 \text{Var}(\delta_k) = \sum_{k=1}^J (v_i^k)^2 p_k (1 - p_k), \quad (9)$$

$$\mathbb{E} [\bar{I}_i^2] = \text{D} [\bar{I}_i] + [\mathbb{E}(\bar{I}_i)]^2 = \sum_{k=1}^J (v_i^k)^2 \text{Var}(\delta_k) + \left(\sum_{k=1}^J p_k v_i^k \right)^2, \quad (10)$$

$$\begin{aligned} \mathbb{E} [\bar{I}_i \bar{I}_j] &= \mathbb{E} \left[\left(\sum_{k=1}^J \delta_k v_i^k \right) \left(\sum_{k=1}^J \delta_k v_j^k \right) \right] = \mathbb{E} \left[\sum_{k=1}^J (\delta_k)^2 v_i^k v_j^k + \sum_{a=1, b \neq a}^J 2\delta_a \delta_b v_i^a v_j^b \right] \\ &= \sum_{k=1}^J v_i^k v_j^k \mathbb{E} [(\delta_k)^2] + \sum_{a=1, b \neq a}^J 2v_i^a v_j^b \mathbb{E} [\delta_a \delta_b] = \sum_{k=1}^J v_i^k v_j^k p_k + \sum_{a=1, b \neq a}^J 2v_i^a v_j^b p_a p_b. \end{aligned} \quad (11)$$

By substituting Eq. (5), (8), (9), (10), (11) into Eq. (7), we have:

$$\begin{aligned} \mathbb{E} \left[\frac{\partial Err_{\text{Pro}}}{\partial w_i} \right] &= -t \sum_{k=1}^J p_k v_i^k + w_i \sum_{k=1}^J (v_i^k)^2 \text{Var}(\delta_k) + w_i \left(\sum_{k=1}^J p_k v_i^k \right)^2 \\ &\quad + \sum_{j \neq i} w_j \left(\sum_{k=1}^J v_i^k v_j^k p_k + \sum_{a=1, b \neq a}^J 2v_i^a v_j^b p_a p_b \right) \\ &= \frac{\partial Err_{\text{Ens}}}{\partial w_i} + w_i \sum_{k=1}^J (v_i^k)^2 \text{Var}(\delta_k) + \sum_{j \neq i} w_j \sum_{k=1}^J v_i^k v_j^k \text{Var}(\delta_k) \end{aligned} \quad (12)$$

Therefore, the gradient expectation $\mathbb{E} \left[\frac{\partial Err_{\text{Pro}}}{\partial w_i} \right]$ when using Prompter is the gradient of the ensemble error Err_{Ens} adding the ℓ_2 regularization of weights and a cross-weight multiplication item as

$$\mathbb{E} [Err_{\text{Pro}}] = Err_{\text{Ens}} + \frac{1}{2} \sum_{i=1}^n (w_i)^2 \sum_{k=1}^J (v_i^k)^2 \text{Var}(\delta_k) + \sum_{i=1}^n \sum_{j \neq i}^n w_i w_j \sum_{k=1}^J v_i^k v_j^k \text{Var}(\delta_k). \quad (13)$$

The magnitude of the ℓ_2 regularization term $\frac{1}{2} \sum_{i=1}^n (w_i)^2 \sum_{k=1}^J (v_i^k)^2 \text{Var}(\delta_k)$ is adaptively scaled by both the input features of body joints and the variance of the Bernoulli variables which is controlled by the context masking probability p_k . When we set $p_k = 0.5$, $\mathbb{E} [Err_{\text{Pro}}]$ possesses the maximal level of the ℓ_2 regularization, and our empirical results also show that the probability value around 0.5 can achieve slightly better performance.

Relations to Dropout Algorithm Hinton et al. (2012); Baldi & Sadowski (2013; 2014): As analyzed in our paper, the proposed Prompter can be viewed as a special *representation-level* Dropout

algorithm performed on the body-joint representations over spatial and temporal dimensions. It randomly and independently masks body-joint representations (can be viewed as randomly and independently dropping joints) in each *skeleton* or in each *trajectory* to construct a new skeleton representation for skeleton semantics learning (*e.g.*, reconstruction in our work), which can (1) introduce random perturbations into skeleton semantics learning tasks (shown in our paper) to enhance model training (*e.g.*, improve robustness against perturbations); (2) fully exploit different random subsets of body-joint representations as more diverse contexts for semantics learning tasks (*e.g.*, reconstruction, prediction) to capture richer key semantic features; (3) may reduce the feature co-adaptation (Hinton et al., 2012) of body joints (*e.g.*, avoid only utilizing a few highly-correlated joints for skeleton prediction) to force the model to learn more effective representations for each body joint; (4) could provide a regularized error function (see Eq. (13)) to potentially reduce the model over-fitting. These properties allows Prompter and the idea of probabilistic spatial-temporal context masking to be broadly applied to more models and tasks (see Sec. D of Appendix I).

REFERENCES

- Pierre Baldi and Peter Sadowski. The dropout learning algorithm. *Artificial intelligence*, 210:78–122, 2014.
- Pierre Baldi and Peter J Sadowski. Understanding dropout. *Advances in Neural Information Processing Systems (NeurIPS)*, 26, 2013.
- Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.
- Haocong Rao and Chunyan Miao. TranSG: Transformer-based skeleton graph prototype contrastive learning with structure-trajectory prompted reconstruction for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- Haocong Rao, Xiping Hu, Jun Cheng, and Bin Hu. SM-SGE: A self-supervised multi-scale skeleton graph encoding framework for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1812–1820, 2021.