Prior Knowledge and Memory Enriched Transformer for Sign Language Translation

Tao Jin
Zhejiang University
jint_zju@zju.edu.cn

Zhou Zhao†
Zhejiang University
zhaozhou@zju.edu.cn

Meng Zhang
Huawei Noah's Ark Lab
zhangmeng92@huawei.com

Xingshan Zeng
Huawei Noah's Ark Lab
zxshamson@gmail.com

Abstract

This paper attacks the challenging problem of sign language translation (SLT), which involves not only visual and textual understanding but also additional prior knowledge learning (i.e. performing style, syntax). However, the majority of existing methods with vanilla encoder-decoder structures fail to sufficiently explore all of them. Based on this concern, we propose a novel method called ***P***rior knowledge and memory ***E***nriched ***T***ransformer (PET) for SLT, which incorporates the auxiliary information into vanilla transformer. Concretely, we develop gated interactive multi-head attention which associates the multimodal representation and global signing style with adaptive gated functions. One Part-of-Speech (POS) sequence generator relies on the associated information to predict the global syntactic structure, which is thereafter leveraged to guide the sentence generation. Besides, considering that the visual-textual context information, and additional auxiliary knowledge of a word may appear in more than one video, we design a multi-stream memory structure to obtain higher-quality translations, which stores the detailed correspondence between a word and its various relevant information, leading to a more comprehensive understanding for each word. We conduct extensive empirical studies on RWTH-PHOENIX-Weather-2014***T*** dataset with both signer-dependent and signer-independent conditions. The quantitative and qualitative experimental results comprehensively reveal the effectiveness of PET.

1 Introduction

Recently, the combination of vision and language attracts increasing attention. Sign language translation which aims to provide translated natural sentences for sign language videos is a valuable but challenging task in this topic (Camgoz et al., 2018, 2020a,b; Jin and Zhao, 2021). Since the visual and textual modalities are not aligned strictly in a weakly-supervised manner, the difficulties of sign language translation mainly lie in the multimodal representation learning of both modalities and the alignments between them. Besides, additional prior knowledge (i.e. the performing style of different signers, the common syntactic structures of sentences) also has a strong influence on multimodal learning.

Figure 1: An example of sign language translation, where the video frames and the sentence correspond to each other. Besides, each word (red) has its syntactic attribute (green).

Encoder-decoder structures built upon long short-term memory unit (Hochreiter and Schmidhuber, 1997) (LSTM) or transformer (Vaswani et al., 2017) are widely used in end-to-end sign language translation, which directly generates natural sentences without intermediate products like gloss sequences. Generally, the encoder extracts and encodes the sign language information, the decoder makes full use of the encoded results with cross-modal interaction. Camgoz et al. (2018) first proposes the sign language translation task and utilizes LSTMs combined with attention mechanism (e.g. Luong Attention (Luong et al., 2015), Badanau Attention (Bahdanau et al., 2014)) to solve it. Due to the insufficient capacity to capture the long-range temporal correlations, Camgoz et al. (2020b) replaces LSTM with transformer, which could correlate any two-time steps of sequential features. The stacked attention blocks improve most of the metrics by a large margin. Camgoz et al. (2020a) combines multiple articulatory channels with an-

† corresponding author

Findings of the Association for Computational Linguistics: ACL 2022, pages 3766 - 3775

choring losses and proposes a novel multi-channel transformer architecture for sign language translation. Li et al. (2020) employs video segment representation with multiple temporal granularities to develop a semantic pyramid network. In summary, many endeavors are devoted to the improvement of deep architectures for multimodal representation learning. However, the influences of additional prior knowledge are totally ignored. For example, as shown in Fig. 1, the natural sentence has its unique syntactic structure.

Motivated by the above observations, we propose a new method called prior knowledge and memory enriched transformer for sign language translation. Specifically, we develop gated interactive multi-head attention which associates the multimodal representation and global signing style with adaptive gated functions. Besides, we employ sentence templates that consist of POS tags to represent the syntactic structures of natural sentences, and accordingly, syntax learning is performed by directly inferring POS tags with the style-specific multimodal representation. The natural sentences are generated conditioned on such auxiliaries. Furthermore, we find that the visual and textual context information, and additional auxiliary knowledge of a word may appear in more than one sign language video. For example, a word that comes up with different words may lead to various contextual visual perceptions, and the general gestural tendency of a word could support the decoding process. Therefore, we design a multi-stream memory structure to store the full-spectrum correspondence between a word and its various relevant information in training data. The obtained memory contents are employed to aid in decoding. We conduct extensive empirical studies on the benchmark dataset, RWTH-PHOENIX-Weather-2014***T*** (PHOENIX14**T**) (Camgoz et al., 2018), with both signer-dependent and signer-independent conditions. The quantitative and qualitative results comprehensively reveal the effectiveness and generalization of PET. The main contributions of this paper can be summarized as follows:

- We propose a new method called prior knowledge and memory enriched transformer for sign language translation, which explores not only multimodal understanding but also the influences of additional prior knowledge on multimodal learning.
- We develop gated interactive multi-head attention by associating the multimodal representation and global signing style with adaptive gated functions. The POS sequence generator relies on the style-specific multimodal information to predict the syntactic structure, which is leveraged to guide the natural sentence generation.
- We design a multi-stream memory structure to store the full-spectrum correspondence between a word and its various relevant information in training data, leading to a more comprehensive understanding for each word.
- The quantitative and qualitative results on the challenging dataset, PHOENIX14**T** of both signer-dependent and signer-independent conditions comprehensively reveal the effectiveness and generalization of PET.

2 Related Work

2.1 Sign Language Translation

Sign language recognition (SLR) aims to recognize single gestures from an input video clip. Many endeavors are devoted to SLR (Camgoz et al., 2016, 2017; Cui et al., 2019; Graves et al., 2006; Wang et al., 2018; Cui et al., 2017). Sign language translation is the final goal of recognition, which aims to directly translate the sign language videos into natural sentences. SLT is similar to video captioning (Jin et al., 2019a, 2020, 2019b; Pei et al., 2019), to some extent. Existing methods are categorized into two-stage and end-to-end methods. Two-stage methods first transform the videos into gloss (gesture) sequences and then rearrange them to generate natural sentences. To guarantee the fluency of sentences, some words that do not carry visual information are added (Camgoz et al., 2018). End-to-end sign language translation aims to directly translate the original sign language videos into natural sentences without intermediate products. Camgoz et al. (2018) first proposes the sign language translation task and utilizes both two-stage and end-to-end methods to solve it. Camgoz et al. (2018) adopts vanilla LSTM-based encoder-decoder structure. Due to the insufficient capacity to capture the long-range temporal correlations. Camgoz et al. (2020b) replaces LSTM with transformer, which could correlate any two-time steps of sequential features. The stacked attention blocks improve most of the metrics with a large margin. Li et al. (2020)

employs video segment representation with multiple temporal granularities to develop a semantic pyramid network.

However, the methods mentioned above fail to explore the multimodal understanding and additional prior knowledge learning sufficiently. In this paper, we propose PET to solve this problem.

3 Approach

Fig. 2 shows the overall framework of prior knowledge and memory enriched transformer based on encoder-decoder structure. We develop gated interactive multi-head attention in all the attention blocks with adaptive gated control of signing style embeddings. In the decoder, we treat the sentence templates which consist of POS tags as the syntax-aware auxiliary for natural sentence generation. Practically, two consecutive decoding blocks (syntactic and textual blocks) rely on the style-specific multimodal representation to predict the target words. Furthermore, we design a multi-stream memory structure to enhance the comprehensive understanding for each word.

3.1 Style-Aware Gated Interactive Encoder

Following (Camgoz et al., 2020b), we utilize the 2D-CNN (Tan and Le, 2019) pre-trained with recognition task (Koller et al., 2019) to extract visual features of sign language videos. Concretely, we first sample video frames and then send them to 2D-CNN. For convenience, we use $I \in \mathbb{R}^{T_i \times d}$ to denote the extracted features, where T_i is the number of video frames. As shown in Fig. 2, the encoder consists of stacked attention blocks. Considering the fact that different signers have corresponding performing styles (i.e. body, pose), we perform adaptive gated interaction for the self-attention mechanism, which associates the visual representation and signing style with adaptive gated functions. Note that, for each specific signer, we obtain the performing style embedding g by simply mean-pooling all the visual features of the corresponding signer (both videos and frames) in the dataset. Specifically, the self-attention layer is formulated as:

$$\text{GI\_Self}(I) = \text{GI\_MH}(I, I|g) \tag{1}$$

where "GI", "Self", "MH" denote gated interactive, self attention, and multi-head attention, respectively. The first "I" in GI_MH(.) denotes query, the second "I" denotes key and value. Further, the calculation of each head is expressed as:

$$\begin{aligned} \text{GI\_MH}(I, I|g) &= [hd_1, ..., hd_h]W_1 \\ hd_i &= \text{GI\_AT}(IW_i^Q, IW_i^K, IW_i^V|gW_i^G) \end{aligned} \tag{2}$$

where $[.]$ denotes concatenation operation, hd_i denotes the output of i-th head, $W_1 \in \mathbb{R}^{d \times d}$, $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times \frac{d}{h}}$ are trainable variables. "GI_AT" takes the signing style embedding into consideration and the process is as below:

$$\text{GI\_AT}(Q, K, V|s) = \text{softmax}(\frac{Q^{'}K^{'\mathbf{T}}}{\sqrt{d_k}})V \tag{3}$$

where we utilize Q, K, V, and s to denote IW_i^Q, IW_i^K, IW_i^V, and gW_i^G to save space. $Q^{'}$ and $K^{'}$ are the results of style-specific interaction with adaptive gated functions:

$$\begin{aligned} Q^{'} &= (1+G_q) \odot Q, \quad G_q = \sigma([s, Q_{\text{M}}, s \odot Q_{\text{M}}]W_q) \\ K^{'} &= (1+G_k) \odot K, \quad G_k = \sigma([s, K_{\text{M}}, s \odot K_{\text{M}}]W_k) \end{aligned} \tag{4}$$

where $\odot$ denotes element-wise multiplication, $\sigma(.)$ denotes sigmoid gated function, the subscript of $K_{\text{M}} \in \mathbb{R}^{\frac{d}{h}}$ and $Q_{\text{M}} \in \mathbb{R}^{\frac{d}{h}}$ denotes mean-pooling, $W_q, W_k \in \mathbb{R}^{\frac{3d}{h} \times \frac{d}{h}}$ are trainable variables. We employ residual connection and layer normalization following the self-attention layer:

$$I^{'} = \text{LN}(I + \text{GI\_Self}(I)) \tag{5}$$

where "LN" denotes layer normalization, followed by a feed-forward layer (FFN) to introduce non-linear transformation:

$$\begin{aligned} \text{FFN}(I^{'}) &= \text{Max}(0, I^{'}W_2 + b_2)W_3 + b_3 \\ I^{''} &= \text{LN}(I^{'} + \text{FFN}(I^{'})) \end{aligned} \tag{6}$$

where $W_2 \in \mathbb{R}^{d \times 4d}$, $b_2 \in \mathbb{R}^{4d}$, $W_3 \in \mathbb{R}^{4d \times d}$, $b_3 \in \mathbb{R}^{d}$ are trainable variables, $I^{''} \in \mathbb{R}^{T_i \times d}$ represents the encoded visual features.

3.2 Syntax-Aware Memory Enriched Decoder

The decoder also consists of stacked attention blocks as shown in Fig. 2. Note that the structures of syntactic and textual blocks are the same as those of encoder-decoder attention blocks. Specifically, to predict the word y_{t_e} at t_e-th time step, we utilize $E_{<t_e} \in \mathbb{R}^{t_e \times d}$ that denotes the embeddings of "BOS" token and the words whose time steps are less than t_e. The process of the masked

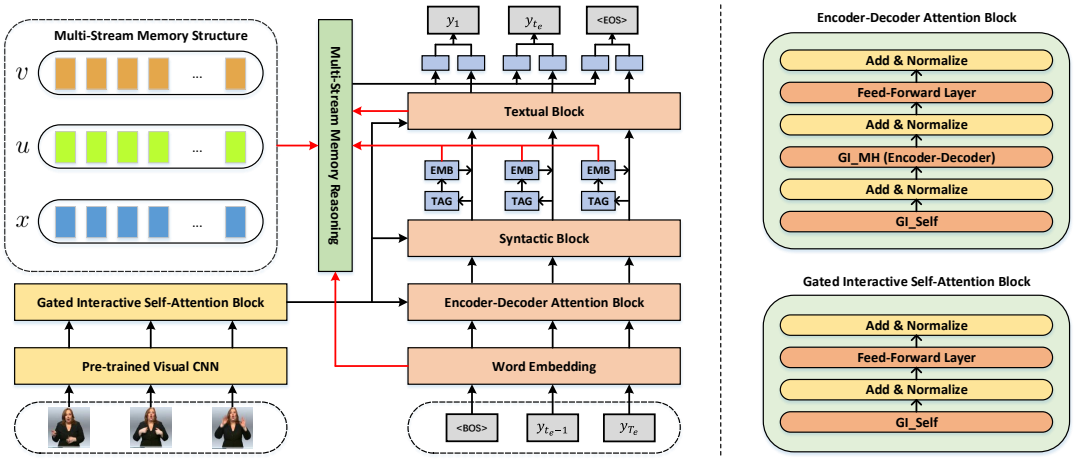

Figure 2: Left is the overall framework of PET, where the encoder processes extracted video features with stacked gated interactive self-attention blocks and the decoder makes full use of the visual features with encoder-decoder attention blocks. Note that the structures of syntactic and textual blocks are the same as those of encoder-decoder attention blocks. "TAG" and "EMB" denote POS tag and embedding, respectively. The multi-stream memory structure is leveraged for auxiliary decoding, where v, u, and x denote visual, textual, and syntactic memory, respectively. Right is the structures of self-attention block and encoder-decoder attention block.

self-attention layer and the following normalization layer is formulated as:

$$E'_{<t_e} = \text{LN}(E_{<t_e} + \text{GI\_Self}(E_{<t_e})) \quad (7)$$

where we also perform adaptive gated interaction for self-attention mechanism. The obtained $E'_{<t_e}$ are utilized to correlate the encoded visual features in the following layer with cross-modal attention:

$$\begin{aligned} E''_{<t_e} &= \text{LN}(E'_{<t_e} + \text{GI\_MH}(E'_{<t_e}, I''|g)) \\ O &= \text{LN}(E''_{<t_e} + \text{FFN}(E''_{<t_e})) \end{aligned} \quad (8)$$

where $E'_{<t_e}$ and I'' are treated as query and key, respectively. $O \in \mathbb{R}^{t_e \times d}$ denotes the output of one encoder-decoder attention block.

3.2.1 Syntax-Aware Decoding

Since the decoder has N attention blocks, we distinguish the output of different blocks with superscripts, O^1, O^2,...,$O^N \in \mathbb{R}^{t_e \times d}$. Note that O^{N-1} and O^N are the output of syntactic and textual blocks, respectively. We calculate the probability distributions of different POS tags as:

$$P_{s,t_e} = \text{softmax}(W_s O^{N-1}_{t_e}) \quad (9)$$

where $W_s \in \mathbb{R}^{N_s \times d}$ is trainable, N_s is the vocabulary size of POS tags. We combine the syntactic information and $O^{N-1}_{t_e}$ for the subsequent process. In practice, we project the POS tags into corresponding embeddings: $(O^{N-1}_{t_e})' = O^{N-1}_{t_e} + E^s_{t_e}$, where $E^s_{t_e}$ denotes POS embedding at t_e-th time step. The obtained synthetic representation $(O^{N-1}_{t_e})'$ is considered as the input of textual block. Due to the space limitation, we omit the calculation in textual block which is similar to Eqns. 7 and 8. The output of textual block is used to predict words:

$$P_{b,t_e} = \text{softmax}(W_p O^N_{t_e}) \quad (10)$$

where $W_p \in \mathbb{R}^{N_w \times d}$ is also trainable, N_w is the vocabulary size of words. Overall, we jointly model the multimodal representation and global syntactic structure for sign language translation by developing an end-to-end trainable neural network.

3.2.2 Multi-Stream Memory Structure

We develop a multi-stream memory structure for auxiliary decoding. The rationale behind this design is that a word in the vocabulary may appear in multiple sign language videos. Since a word that comes up with different words may lead to various contextual visual perceptions and one word may correspond to more than one syntactic category, the memory structure is developed to capture the detailed relevant information from different sign language videos where the same word appears, leading to a comprehensive understanding for this word.

(1). Weakly-Aligned Visual Memory: The memory structure is developed to store the descriptive information for each word in the vocabulary. We

construct a dictionary $\langle w, r \rangle$ to record the words w and corresponding representation r. Since the fine-grained alignments between natural words and video frames are not provided, we could not directly obtain the visual memory. However, the end-to-end training of PET provides the weakly-supervised alignments through the cross-modal interaction in the encoder-decoder attention blocks. Therefore, we adopt a separate training scheme. Concretely, we first train a basic sign language translation model with prior knowledge enriched transformer introduced in previous sections to acquire the weakly-supervised alignments between words and video frames. In practice, we only keep the cross-modal attention weights in the textual block. The visual context information $v_{j,i}$ for the j-th word i-th head is modeled as:

$$v_{j,i} = \frac{\sum_{n_v=1}^{N_v} \sum_{n_f=1}^{N_f} (a^i_{n_v,n_f} f^{v,i}_{n_v,n_f})}{\sum_{n_v=1}^{N_v} \sum_{n_f=1}^{N_f} (a^i_{n_v,n_f})} \quad (11)$$
$$v_j = [v_{j,1}, ..., v_{j,i}, ..., v_{j,h}]$$

where N_v denotes the number of related videos in the training set, N_f denotes the number of frames. Note that we only retain the top-N_f relevant video frames to reduce the invalid information. $a^i_{n_v,n_f}$ denotes n_f-th attention weight among the top-N_f weights and $f^{v,i}_{n_v,n_f}$ denotes the corresponding visual features in i-th head. Note that we only focus on the visual features of the last encoding block. The context features are normalized to make the magnitude consistent for words with different frequencies. The final context information v_j is obtained by concatenating the results of all the heads.

(2). Global Syntactic Memory: Considering the fact that a word appearing in multiple sentences may have different syntactic information, we calculate the ratio of different POS categories for each word. The syntactic representation u_j for the j-th word is modeled as:

$$u_j = \sum_{n_s=1}^{N_s} b^s_{n_s} f^s_{n_s}, \ \sum_{n_s=1}^{N_s} b^s_{n_s} = 1 \quad (12)$$

where $b^s_{n_s}$ and $f^s_{n_s}$ denote the weight and embedding of n_s-th POS tag, respectively.

(3). Adjacent Textual Memory: The vanilla transformer-based decoder does not model the compatibility between adjacent words explicitly. Thus, the textual memory is designed to capture the information of adjacent words. Concretely, we set the maximal adjacent step to N_a, which means that we retain the word embeddings of adjacent words and the threshold is N_a. The context representation x_j for the j-th word is modeled as:

$$x_j = \frac{\sum_{n_v=1}^{N_v} \sum_{n_a=1}^{2N_a+1} f^t_{n_v,n_a}}{N_v(2N_a+1)} \quad (13)$$

where $f^t_{n_v,n_a}$ denotes the n_a-th word embedding among the $2N_a+1$ adjacent embeddings. We also employ normalization for the final result. In summary, we obtain the multi-stream memory structure which records full-spectrum information r_j for each word w_j with a map structure: $\langle w_j, r_j \rangle = \langle w_j, \{v_j, u_j, x_j\} \rangle$.

3.2.3 Memory Enriched Decoding

We employ the constructed multi-stream memory structure to build an auxiliary decoding system, where the translation results are further combined with the generated sentences by the syntax-aware decoding system. In this way, the translation quality is improved.

In detail, the memory enriched decoding system is built upon the backbone of the syntax-aware decoding system as an auxiliary module. The probability distributions of different words are calculated similarly to Eqn. 10:

$$P_{m,t_e} = \text{softmax}(Q_{t_e}) \quad (14)$$

where $Q_{t_e} \in \mathbb{R}^{N_w}$ denotes the relevance scores of different words and $Q_{t_e,j} \in \mathbb{R}$ denotes the j-th element among them. We employ $Q_{t_e,j}$ to measure the qualification of j-th word for t_e-th time step based on its memory contents:

$$Q_{t_e,j} = w_p^{\mathbf{T}} \tanh(W_v[v_j, O^N_{t_e}] + W_u[u_j, E^s_{t_e}] + W_x[x_j, E^y_{t_e\text{-}1}]) \quad (15)$$

where we concatenate the memory contents (v_j, u_j, x_j) with corresponding representation ($O^N_{t_e}$, $E^s_{t_e}$, $E^y_{t_e\text{-}1}$) at t_e-th time step. For instance, $u_j, E^s_{t_e} \in \mathbb{R}^d$ both denote syntactic information, $x_j, E^y_{t_e\text{-}1} \in \mathbb{R}^d$ both denote textual information. $W_v, W_u, W_x \in \mathbb{R}^{d \times 2d}$, $w_p \in \mathbb{R}^d$ are all trainable variables.

3.3 Training

The optimization goal of sign language translation is to minimize the cross-entropy loss function defined as accumulative loss from all the time steps. Consequently, the syntax-aware decoder is trained by minimizing the combined loss:

$$L_b = -\sum_{t_e=1}^{T_e}\left[\log P_{b,t_e}(y_{t_e}) + \lambda \log P_{s,t_e}(s_{t_e})\right] \quad (16)$$

where y_{t_e} and s_{t_e} denote the ground-truth word and POS tag at t_e-th time step, respectively. λ is a hyper-parameter to balance the two losses. In practice, we set it to 0.5. The memory enriched decoder is trained in a similar way:

$$L_m = -\sum_{t_e=1}^{T_e} \log P_{m,t_e}(y_{t_e}) \quad (17)$$

The syntax-aware decoder and memory enriched decoder are trained in order. We fix the trainable variables except for those in Eqn. 15 when training memory enriched decoder. During inference, we combine the generated results of both decoders.

4 Experiments

In this section, we present the experimental settings of sign language translation and report the results on the benchmark datasets.

Table 1: The statistical results of PHOENIX14T, where the total number of samples is 8257.

Signer	1	2	3	4	5	6	7	8	9
All	2191	95	683	1207	1933	47	866	966	269

4.1 Dataset and Protocols

PHOENIX14T (Signer-Dependent) is the first complete sign language understanding dataset, where a training or testing sample contains a sign language video and the corresponding signer, gloss annotations, natural language translation. Concretely, PHOENIX14**T** is labeled by 9 different signers (the training, validation, and test sets all contain these signers) with a vocabulary of 1066 different sign glosses. In general, one gloss may correspond to multiple natural words, and some words that do not carry visual information are added to guarantee the fluency of sentences, leading to a vocabulary of 2887 words for translation into German language.

PHOENIX14T (Signer-Independent) is obtained by re-dividing the original PHOENIX14**T** dataset. Since the 9 signers are in both the training set and test set, there are no unseen signers for evaluating the generalization. We simply choose the signers 8, 9 (1235 samples) for testing and the other signers (7022 samples) for training and validation, the statistical info is shown in Table 1.

Table 2: Evaluation results on PHOENIX14**T** (Signer-Dependent), where B@{1, 2, 3, 4} denotes BLEU-{1, 2, 3, 4} and R denotes ROUGE-L.

Method	PHOENIX14**T**				
	B@1	B@2	B@3	B@4	R
Multitask	37.22	23.88	17.08	13.25	36.28
DeepHand	38.50	25.64	18.59	14.56	38.05
Mul-Ch.	-	-	-	19.51	45.90
NSLT	32.24	19.03	12.83	9.58	31.80
TSPNet	36.10	23.12	16.88	13.41	34.96
SL-Trans.	47.20	34.46	26.75	21.80	-
ST-Trans.	48.61	35.97	28.37	23.65	-
STMC-T	48.73	36.53	29.03	24.00	46.77
PET	**49.54**	**37.19**	**29.30**	**24.02**	**49.97**

Table 3: Evaluation results on PHOENIX14**T** (Signer-Independent), where * denotes that we implement the methods by ourselves, since none of the previous work conducts experiments on signer-independent setting.

Method	PHOENIX14**T**				
	B@1	B@2	B@3	B@4	R
NSLT*	26.01	13.84	8.95	6.28	25.22
TSPNet*	28.10	16.81	11.82	9.15	31.00
SL-Trans.*	40.15	26.70	19.22	14.78	40.22
PET	**41.72**	**28.97**	**21.36**	**16.94**	**42.45**

We follow the commonly used protocol **Sign2Text (S2T)** in the previous work (Camgoz et al., 2020b), which aims to directly translate the sign language videos into natural sentences without converting the input into intermediate products. Since the visual and textual modalities are not aligned strictly in a weakly-supervised manner, the difficulties of Sign2Text mainly lie in the multimodal alignments.

4.2 Implementation Details

Framework: Following (Camgoz et al., 2020b), a modified version of JoeyNMT (Kreutzer et al., 2019) is employed to implement PET. We utilize PyTorch and Tensorflow frameworks. Except for the CTC beam search decoding module which is

Table 4: Evaluation results of style-specific interaction, where P14**T** (SD) and P14**T** (SI) denote PHOENIX14**T** with signer-dependent and singer-independent settings.

Method	P14**T** (SD)					P14**T** (SI)				
	B@1	B@2	B@3	B@4	R	B@1	B@2	B@3	B@4	R
w/o. GI	48.61	35.24	27.58	22.89	48.34	40.22	27.36	20.05	15.42	40.36
Add	49.04	36.05	28.32	23.40	48.88	40.91	28.19	20.65	16.36	40.76
Enc.	49.45	36.57	28.95	23.45	49.15	41.37	28.54	20.57	16.66	41.54
Dec.	49.30	36.32	28.84	23.42	49.08	41.43	28.52	20.89	16.72	41.28
PET	**49.54**	**37.19**	**29.30**	**24.02**	**49.97**	**41.72**	**28.97**	**21.36**	**16.94**	**42.45**

implemented with Tensorflow, the other modules are developed with PyTorch.

Network Details: The hidden size is set to 512 for all the multi-head attention mechanisms. The numbers of heads and attention blocks are 8 and 3, respectively. The ground-truth POS tags could be obtained by Stanford POS Tagger, which are divided into 13 categories: ADJ, ADV, ADP, VERB, NOUN, DET, PRON, AUX, CONJ, PROPN, NUM, UNK, PUNCT, we project them into 512-dimensional syntactic embeddings. We train all of the networks from scratch.

Training: In the training stage, we utilize Adam algorithm (Kingma and Ba, 2014) to optimize the loss function. The batch size is set to 64. The learning rate is set to 5×10^{-4} initially. We evaluate our network every 100 iterations. If the metric on validation set does not improve for 9 evaluation steps, we decrease the learning rate by a factor of 0.5. When the learning rate is less than 10^{-6}, we finish the training stage.

Testing: Since the test set may have unseen signers, we calculate the style embedding with mean-pooling operation for the acquired visual features similarly. Beam search is a commonly used method to decode words during evaluation. We adopt the beam size 5. We employ the commonly-used metrics, BLEU-n and ROUGE-L.

4.3 Compared Baseline Methods

NSLT (Camgoz et al., 2018): NSLT first proposes the SLT task and employs LSTM-based structure to translate sign language videos.

Multitask (Orbay and Akarun, 2020): Multitask employs joint learning scheme to enhance the SLT performance.

DeepHand (Orbay and Akarun, 2020): DeepHand transfers the knowledge of hand dataset to the SLT task.

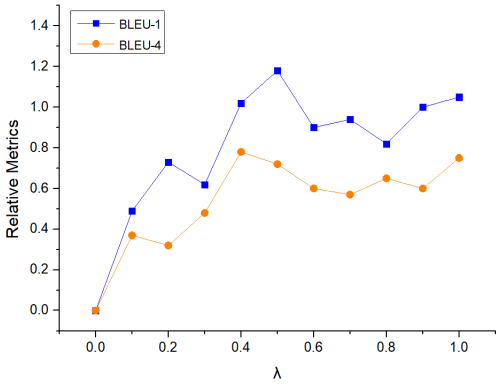

Figure 3: The trade-off between different losses in Eqn. 16, where we set $\lambda = 0$ as the baseline.

SL-Trans. (Camgoz et al., 2020b): SL-Trans. is the recent mainstream method for SLT, the encoder and decoder both consist of Transformer modules.

TSPNet (Li et al., 2020): TSPNet employs video segment representation with multiple temporal granularities to develop a semantic pyramid network.

Mul-Ch. (Camgoz et al., 2020a): Mul-Ch. combines multiple articulatory channels with anchoring losses and proposes a novel multi-channel transformer architecture for sign language translation.

ST-Trans. (Voskou et al., 2021): ST-Trans. equips Transformer with stochastically competing linear units and performs variational Bayesian inference over all connection weights, throughout the network.

STMC-T (Yin and Read, 2020): STMC-T employs spatial-temporal multi-channel Transformer to solve the SLT task.

4.4 Quantitative Results

We compare PET with the recent state-of-the-art methods. Following the previous work (Camgoz et al., 2020b), for PHOENIX14**T** (Signer-Dependent), we develop the gloss-based PET by

Table 5: Evaluation of memory-enriched decoding

Method	P14**T** (SD)					P14**T** (SI)				
	B@1	B@2	B@3	B@4	R	B@1	B@2	B@3	B@4	R
w/o. Vis	49.63	36.28	28.58	23.40	49.32	41.24	28.35	20.89	16.30	41.46
w/o. Tex	49.52	36.54	28.83	23.44	49.12	41.05	28.16	20.74	16.44	40.93
w/o. Syn	**49.69**	36.42	28.75	23.55	49.48	41.58	28.55	21.07	16.64	41.32
w/o. Mem	48.94	35.64	28.07	22.71	49.05	40.54	27.53	20.25	15.56	40.64
PET	49.54	**37.19**	**29.30**	**24.02**	**49.97**	**41.72**	**28.97**	**21.36**	**16.94**	**42.45**

adding the gloss supervision with CTC loss in the encoder. Table 2 shows the experimental results, we could find that PET (model-based) outperforms all the model-based and feature-based methods, NSLT (Camgoz et al., 2018), Multitask (Orbay and Akarun, 2020), DeepHand (Orbay and Akarun, 2020), SL-Trans. (Camgoz et al., 2020b), TSPNet (Li et al., 2020), Mul-Ch. (Camgoz et al., 2020a), ST-Trans.(Voskou et al., 2021) and STMC-T (Yin and Read, 2020) on all the metrics. In particular, PET achieves 49.97% on ROUGE-L, making a large improvement of 3.20% over STMC-T.

Table 3 shows the results on PHOENIX14**T** (Signer-Independent), we implement several state-of-the-art methods manually, since none of the previous work conducts experiments on the signer-independent setting (PET is model-based method, so we mainly reproduce the model-based methods, since the methods of other types are compatible with PET). Note that, to keep fairness, we employ the same method of feature extraction in the original paper for NSLT, TSPNet, and SL-Transformer, respectively. The experimental results demonstrate the generalization of PET for unseen signers.

4.5 Ablation Study

In this section, we evaluate the effectiveness of all the contributions with ablation experiments.

4.5.1 Effect of Adaptive Gated Interaction

As shown in Table 4, we design four control experiments to demonstrate the effectiveness of adaptive gated interaction, where **w/o. GI** denotes that we remove the adaptive gated interaction from all attention blocks and keep the other contributions, **Add** denotes that we add the style embedding to the multimodal features, **Enc (only)** denotes that we only keep the adaptive gated interaction in the encoder, while **Dec (only)** denotes that we discard the adaptive gated interaction in the encoder. It is observed that PET outperforms four ablation methods on the benchmark datasets and **w/o. GI** achieves the worst performances on both BLEU and ROUGE-L, which demonstrates that the translation results benefit from the style information. The remaining ablation results illustrate that gated interaction is better than naive addition. In addition, the adaptive gated interaction enhances the multimodal alignments, corresponding results are shown in the appendix.

4.5.2 Effect of Syntax-Aware Auxiliary

We adjust the ratio of different losses in Eqn. 16 and obtain the experimental results that are shown in Fig. 3. To make the comparison more intuitive, we set $\lambda = 0$ as the baseline and provide the relative performances of BLEU-1 and BLEU-4 on PHOENIX14**T** (SD). We find that the performances improve as the λ increases when λ is less than 0.5. Subsequently, the performances are beginning to level off. Such results demonstrate the effectiveness of syntax-aware auxiliary.

4.5.3 Effect of Memory-Enriched Decoding

As shown in Table 5, we also design several control experiments to evaluate the impact of the memory enriched decoding, where **w/o. Mem** denotes the model without memory mechanism, **w/o. Vis** denotes the model only without visual memory, **w/o. Tex**, **w/o. Syn** denote the models without textual memory and syntactic memory, respectively. We find that PET outperforms all the ablation methods on both BLEU-4 and ROUGE-L. Particularly, compared with **w/o. Mem**, PET achieves a significant improvement on BLEU-4 (1.38% for SI, 1.31% for SD).

4.6 Qualitative Results

We would like to investigate the generation process of our model by qualitative results in this section.

Table 6: Qualitative results of PET, where "Ref" denotes reference, "SL-Trans." denotes SL-Transformer. As the annotations in the PHOENIX14**T** dataset are in German, we share both the produced sentences and their translations in English. Note that the words highlighted in red are those that require critical translation, the words highlighted in blue are the failure cases of current mainstream method SL-Transformer.

Ref:	und zum wochenende wird es dann sogar wieder ein bisschen kälter .
	(and at the weekend it even gets a little colder again .)
SL-Trans.:	und der januar .
	(and january .)
PET:	und das wird dann am wochenende ein bisschen kälter .
	(and that gets a bit colder on the weekend .)
Ref:	ganz ähnliche temperaturen wie heute zwischen sechs und elf grad .
	(very similar temperatures as today between six and eleven degrees .)
SL-Trans.:	hier und da ähnliche temperaturen wie heute meist ein grad .
	(here and there temperatures similar to today, mostly one degree .)
PET:	ähnliches wetter heute nacht nur sechs bis elf grad .
	(similar weather tonight only six to eleven degrees .)
Ref:	deutschland liegt morgen unter hochdruckeinfluss der die wolken weitgehend vertreibt .
	(tomorrow germany will be under the influence of high pressure which will largely drive away the clouds .)
SL-Trans.:	in deutschland liegt morgen unter tiefdruckeinfluss und wolken .
	(in germany tomorrow is under the influence of low pressure and clouds .)
PET:	Morgen wird Deutschland von hohem Druck betroffen sein .
	(tomorrow germany will be hit by high pressure .)

Here we provide some sign language translation examples in Table 6. As the annotations in the PHOENIX14**T** dataset are in German, we share both the produced sentences and their translations in English. Note that the words highlighted in red are those that require critical translation, the words highlighted in blue are the failure cases of current mainstream method SL-Transformer. Benefiting from the style-specific interaction, syntax-aware auxiliary, and memory enriched decoding, PET could accurately translate some detailed information compared with SL-Transformer and retain the whole contents of the ground truth better than SL-Transformer, which demonstrates the effectiveness again.

5 Conclusion

In this paper, we have proposed a new method called prior knowledge and memory enriched transformer for sign language translation. Specifically, we develop the adaptive gated interaction which associates the multimodal representation and global signing style in all the attention blocks. One POS sequence generator relies on the associated information to predict the global syntactic structure, which is thereafter leveraged to guide the sentence generation. Besides, considering that the visual and textual context information, and additional auxiliary knowledge of a word appear in more than one video, we design a memory structure to store the full-spectrum correspondence between a word and its various relevant information in the training data. The experimental results reveal the effectiveness and generalization of PET.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant No.2020YFC0832505, National Natural Science Foundation of China under Grant No.61836002, No.62072397 and Zhejiang Natural Science Foundation under Grant LR19F020006.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. Subunets: End-to-end hand shape and continuous sign language recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3075–3084. IEEE.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Necati Cihan Camgoz, Ahmet Alp Kindirouglu, and Lale Akarun. 2016. Sign language recognition for assisting the deaf in hospitals. In *International Workshop on Human Behavior Understanding*, pages 89–101. Springer.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020a. Multi-channel transformers for multi-articulatory sign language translation. In *European Conference on Computer Vision*, pages 301–319. Springer.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020b. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033.

Runpeng Cui, Hu Liu, and Changshui Zhang. 2017. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7361–7369.

Runpeng Cui, Hu Liu, and Changshui Zhang. 2019. A deep neural framework for continuous sign language

recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7):1880–1891.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Tao Jin, Siyu Huang, Ming Chen, Yingming Li, and Zhongfei Zhang. 2020. Sbat: Video captioning with sparse boundary-aware transformer. *arXiv preprint arXiv:2007.11888*.

Tao Jin, Siyu Huang, Yingming Li, and Zhongfei Zhang. 2019a. Low-rank hoca: Efficient high-order cross-modal attention for video captioning. *arXiv preprint arXiv:1911.00212*.

Tao Jin, Yingming Li, and Zhongfei Zhang. 2019b. Recurrent convolutional video captioning with global and local attention. *Neurocomputing*, 370:118–127.

Tao Jin and Zhou Zhao. 2021. Contrastive disentangled meta-learning for signer-independent sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5065–5073.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. 2019. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2306–2320.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey nmt: A minimalist nmt toolkit for novices. *arXiv preprint arXiv:1907.12484*.

Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Ben Swift, Hanna Suominen, and Hongdong Li. 2020. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *arXiv preprint arXiv:2010.05468*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Alptekin Orbay and Lale Akarun. 2020. Neural sign language translation by learning tokenization. *arXiv preprint arXiv:2002.00479*.

Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. 2019. Memory-attended recurrent network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8347–8356.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Andreas Voskou, Konstantinos P Panousis, Dimitrios Kosmopoulos, Dimitris N Metaxas, and Sotirios Chatzis. 2021. Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11946–11955.

Shuo Wang, Dan Guo, Wen-gang Zhou, Zheng-Jun Zha, and Meng Wang. 2018. Connectionist temporal fusion for sign language translation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1483–1491.

Kayo Yin and Jesse Read. 2020. Better sign language translation with stmc-transformer. *arXiv preprint arXiv:2004.00588*.