Journal of Computing and Communication Vol.1 , No.2 , PP. 1-8 , 2022

A Deep Learning Approach for Gloss Sign Language Translation using Transformer

Ammar Mohammed*[a,b], Mohamed Amin[a], Hesham Hefny[a]

[a]Department of Computer Science, Faculty of Graduate Studies for Statistical Research, Cairo University, Cairo, Egypt

[b]Department of Computer Science, Misr International University, Cairo, Egypt

*Corresponding Author: Ammar Mohammed [***ammar@cu.edu.eg***]

ARTICLE DATA

Article history:
Received 27 April 2022
Revised 08 August 2022
Accepted 09 August 2022
Available online

Keywords:
Neural Machine Translation
Sequence to Sequence Model
Sign Language
Deep learning
Transformer

ABSTRACT

One of the most recent applications of machine learning is to translate sign language into natural language. Many studies have attempted to classify sign language based on whether it is gesture or facial expression. These efforts, however, ignore genuine sentences' linguistic structure and context. The quality of traditional translation methods is poor, and their underlying models are not salable. They also take a long time to complete. The contribution of this paper is that it suggests utilizing a transformer to perform bidirectional translation using a deep learning approach. The proposed models experiment on the ASLG-PC12 corpus. The experimental results reveal that the proposed models outperform other approaches to the same corpus in both directions of translation, with ROUGE and BLEU scores of 98.78% and 96.89%, respectively, when translating from text to gloss. Additionally, the results indicate that the model with two layers achieves the best result with ROUGE and BLEU scores of 96.90% and 84.82% when translating from gloss to text.

1. Introduction

Sign languages are visual-gesture-based languages regarded as the deaf community's mainstream language. Gestures and visual channels communicate in this language [1]. Hand gestures, body movements, and facial expressions are utilized for communicating in sign language. The World Health Organization estimates that 466 million people worldwide suffer from hearing loss, with 34 million children. Over 900 million individuals are expected to experience hearing loss or communication issues by 2050, according to estimates [2]. Almost 121 forms of sign language are used worldwide now [3], with a shortage of sign language interpreters to deal with the diversity. As a result, translation technologies that make the translation process faster and more precise are needed. The first stage in automatic translation is to standardize sign language. Stokoe [4], HamNoSys [5], SignWriting [6], and Gloss Notation are just a few examples of sign language Forms. Facial expressions and body movements are not included in Stokoe notation. As a result, this sign language is limited and unsuitable for deaf translation. In addition, the HamNoSys form uses a 3D animated avatar to formalize any sign language. However, It does not give a simple way to describe body movements and facial expressions. The SignWriting notation employs highly iconic symbols, which are difficult to decipher using a computer. Gloss notation, on the other hand, is a formal sign language similar to Braille, Morse code, and finger-spelling. It annotates, represents, and explains visual-gestural language sequences based on natural language word labels. This is a simple approach to represent an idea in sign language articulated in natural language. Glossing has received much attention in sign language translation because of its simplicity, expressiveness, and formal representation of sign language [7-9, 3]. For years, machine and deep learning have demonstrated remarkable effectiveness in various application domains.

Several researchers have expressed interest in employing a neural network to translate sign languages using machine translation [10-14].Neural Machine Translation (NMT) is a modern neural network-based

translation technique [15]. It is an end-to-end learning approach for automated translation. It has two parts: an encoder and a decoder. An attention mechanism [16] has recently been developed to allow a neural network to pay attention to only a specific area of an input sentence while creating a translation comparable to human translations to improve the learning process. Even though NMT approaches are more successful than traditional machine translation approaches, most neural-based studies disregard the linguistic features of sign language. They believe there is only a one-to-one correspondence between signs and spoken words.

Furthermore, most modern neural networks concentrate on translating from gloss sign language to natural language. However, completely automating translation systems in both directions require the second direction from natural language to gloss sign language. The following are the contributions of this paper. First, it provides sequence-to-sequence deep learning models that translate gloss sign language to natural language text using a transformer. Second, it introduces a deep learning model that translates natural language text to sign language gloss using a sequence-to-sequence approach. Third, this study evaluates the proposed models on the ASLGPC12 corpus [17, 18]. Different metrics, such as BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation), are used to evaluate the performance outcomes. In addition, the best model of the experiments is compared to previous research on the same corpus. The remainder of the paper is laid out: The second section provides some background information on sign languages. Section 3 discusses several related works. Section 4 introduces the proposed approach. The experimental results are discussed in Section 5. Finally, Section 6 concludes the paper.

2. Background

The concepts of sign language and machine translation are briefly introduced in this section.

2.1. Sign Language

The visual-manual form is used to express meaning in sign languages. When compared to spoken languages, sign languages have unique articulators. In spoken languages, the throat, nose, and mouth serve primary articulators, whereas the fingers, hands, and arms serve as primary articulators in sign languages. One of the most common linguistic elements of sign language is the so-called non-manual feature. It is a parameter of a meaningful sign. It is not created by hand. But with eyebrow movement, facial expression, mouth patterns, movement of the eyes/cheeks, movement of the upper body, tilting of the head, and shoulder movements. A sign language statement will be worthless without a non-manual feature, regardless of whether the syntax is in the appropriate order.

Non-manual signals are used in sign language to distinguish between declarative, imperative, and interrogative statements. Stokoe [4], HamNoSys [5], SignWriting [6], and Gloss Notation [19] are examples of several ways to express sign language. Stokoe, HamNoSys, and SignWriting are iconic representations of a sign language that is difficult for deaf people to read and interpret and are used by translation systems to generate 3D animations. Gloss notation, on the other hand, is a visual-gestural language based on labels-words used to annotate, express, and describe sequences of signs. Linguists use it for transcription, and it's called an interlinear translation. Gloss notation is a good technique to concentrate on grammar and word order while keeping vocabulary distinct. Gloss notation is also written in CAPITAL letters above the natural words. Table 1 shows pairs of sentences (in English and American sign language).

TABLE 1: English and American Sign Language pairs

English Sentences	ASL Gloss
What is your name?	NAME YOU WHAT^{WH}
He doesn't like pizza.	PIZZA IX-boy DOESN'T-LIKE
Help me.	HELP-ME (one sign)
See you later.	SEE-YOU-LATER (one sign)
Don't know.	DON'T-KNOW (one sign)
Today is Friday, October 28th.	NOW+DAY FRIDAY fs-OCT 28

2.2. Machine Translation

In machine translation, the NMT technique has made significant progress. It's an end-to-end approach to automated translation learning. [15]. NMT outperforms other traditional techniques due to a variety of factors. First, NMT optimizes all translation learning parameters simultaneously to reduce network output loss automatically. Second, it has distributed representations with many improvements by sharing statistical strengths across similar words or phrases. Third, it can exploit the context of translations better. NMT can learn a larger context if there is more source and destination text. As a result, NMT outperforms other methods in terms of efficiency and quality.

A sequence-to-sequence model implemented as a coupled network of encoder and decoder with an attention mechanism is one of the NMT techniques. In this model, a source sentence $x = \{x_1, x_2, .., x_I\}$ of length I words is given, The model converts this sentence into a target sentence $y = \{y_1, y_2, .., y_J\}$. The encoder network converts source sequences into a list of vectors for each input. On the other hand, the decoder network produces one symbol at a time until the particular end-of-sentence symbol is generated.

3. Related Work

Many attempts to automate sign language translations have recently been made. Various algorithms and machine translation approaches are used in these efforts. Several authors used neural machine translation of sign languages, similar to the method proposed in this paper. In [10], for example, the authors presented a neural sign Language translation that converts gloss sign language to natural language. They used a sequence-to-sequence neural model in their research and evaluated the results on the Phoenix-2014T2 corpus. The BLEU scores for their proposed GRU model with the Luong attention mechanism were 44.13%, 31.47%, 23.89%, and 19.26%, respectively, and the ROUGE score was 45.45%. [12] reported on a similar study that employed the sequence-to-sequence approach. The authors propose that gloss sign language be translated into text. For their studies with three different attention functions: dot, general, and concat, they used ASLG-PC12 corpus on several network architectures. Using GRU with dot attention function hidden size 800 units, the evaluation of BLEU score in the range of 1 to 4 grams attained are 86.70%, 79.50%, 73.20%, and 65.90%. In [13], the authors proposed a sequence-to-sequence translation model based on a human keypoint estimate. They create their work's KETI sign language corpus [13], including 14,672 high-resolution and high-quality movies with gloss translations. 64% of the corpus was used for training, 7% for development, and 29% for testing. On the top level, their model, built on a sequence-to-sequence model based on GRU cells, scored 55.28% accuracy, 52.63% BLEU score, and 63.53% ROUGE score. In addition, the authors of [20] proposed sign language transformers: joint end-to-end sign language recognition and translation. They evaluated their proposed work on the Phoenix-2014T dataset, and the BLEU scores for their proposed model are 48.9%, 36.88%, 29.45%, and 24.54%. In [11], the authors also proposed a translation system based on transformer models. They evaluated their proposed work on the Phoenix-2014T [10] and ASLG-PC12 [17, 18] corpora. Using Transformer on the Phoenix-2014T dataset, their proposed model achieved BLEU in the range of 1 to 4 grams with scores of 48.40%, 36.90%, 29.70%, and 24.90%.

Furthermore, they used Transformer on ASLG-PC12 to achieve BLEU scores of 92.88%, 89.22%, 85.95%, and 82.87%. In [14], the author also uses Generative Adversarial Networks (GANs) to create a sign language video of a human signer rather than a skeletal pose. Even though the final video is visually appealing, the approach still relies on concatenating isolated signs, ignoring the signs' grammatical syntax. The authors in [21] proposed data augmentation for sign language gloss translation. They experimented with their proposed work on Phoenix-2014T. The evaluation of their proposed model on Phoenix-2014T achieved BLEU-4 23.35% and COMET scores 13.65% using Transformer. The authors in [22] proposed approaching sign language gloss translation as a low-resource machine translation task. They experimented with their proposed work on Phoenix-2014T. Their proposed model on Phoenix-2014T evaluated BLEU-4 at 24.38% using Transformer. The authors in [23] proposed conditional sentence generation and cross-modal reranking for Sign Language translation. They experimented with their proposed work on Phoenix-2014T. The evaluation of their proposed model on Phoenix-2014T achieved BLEU4 15.18% and ROUGE score of 38.85% using Transformer. Similarly, the authors in [24] proposed A Simple Multi-Modality Transfer Learning Baseline for Sign Language Translation. They experimented with their proposed work on Phoenix-2014T and CSL-Daily. The evaluation of their proposed model on Phoenix-2014T achieved a BLEU-4 of 26.70% and ROUGE score of 52.54% using Transformer. Despite the success of previous neural network translation approaches, most of these approaches, except this paper, focus on only one way of translation from gloss sign language to natural language.

4. Proposed Approach

This section demonstrates the proposed method for translating natural language text to gloss sign language and vice versa. The proposed strategy is split into two parts. Text is translated into gloss notation in the first direction, and gloss notation is translated into text in the second. The following are the specifics for each direction.

4.1. Text to Gloss Notation Approach

Figure 1 shows the text to gloss notation method. The NMT receives the input text and converts it to gloss notation. The NMT is divided into preprocessing and encoding/decoding phases. Convert natural language text to lowercase and convert gloss notation to uppercase during the preprocessing step. Remove digits and punctuation, as well as any white spaces. Following that, the text is embedded in a continuous vector space. The second phase consists of a self-attention-based encoder-decoder neural network model that transforms the embedded text into gloss notation language. Each encoder layer is divided into two sub-layers: a multi-head attention mechanism and a position-wise, fully-connected feed-forward layer. And each sub-layer is also a residual connection followed by layer normalization. The output of each sub-layer is *LayerNorm(x+Sublayer(x)),* where *x* is the encoder input and Sublayer denotes the function applied by the sub-layer itself. The decoder layer contains an additional" Encoder Decoder attention" that works similarly to multi-head attention. Still, it utilizes queries from the layer below it and keys and values from the encoder stack's output. At each time step, the decoder stack outputs a symbol
from the output sentence, which is then fed to the first decoder layer in the next step until the end of the sentence is reached. The self-attention layers in the decoder also mask future positions by setting them to *-inf,* for example, so that the predictions for the *i*-th symbol can only depend on known outputs at positions less than *i*.

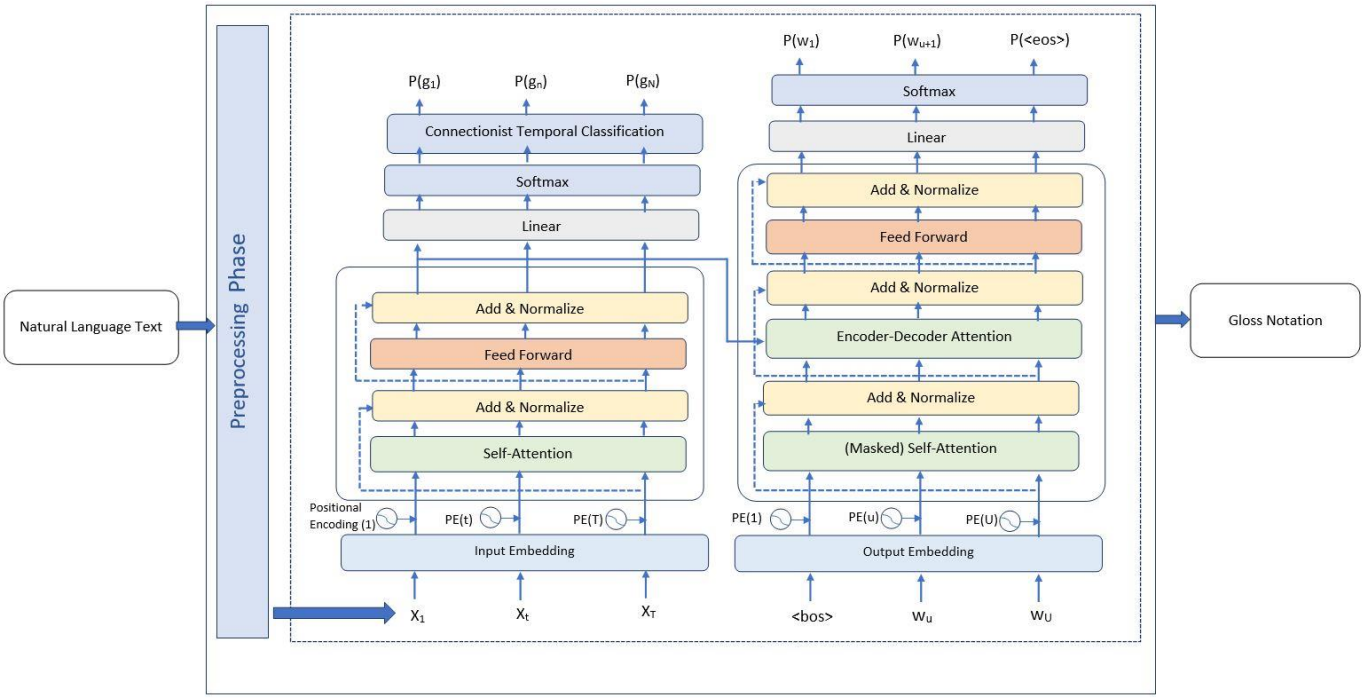

FIGURE 1. Natural Language Text to Sign Language Gloss Model

4.2. Gloss to text Approach

Figure 2 illustrates the second direction of the proposed approach. The primary objective is to convert gloss notations into text. To do that, the machine translation component receives a gloss notation and executes natural language preprocessing operations, where the gloss is embedded in a continuous vector space. Second, the embedded gloss is translated into text using an encoder-decoder with a self-attention mechanism. The encoder and decoder architecture is similar to that shown in Figure 1.

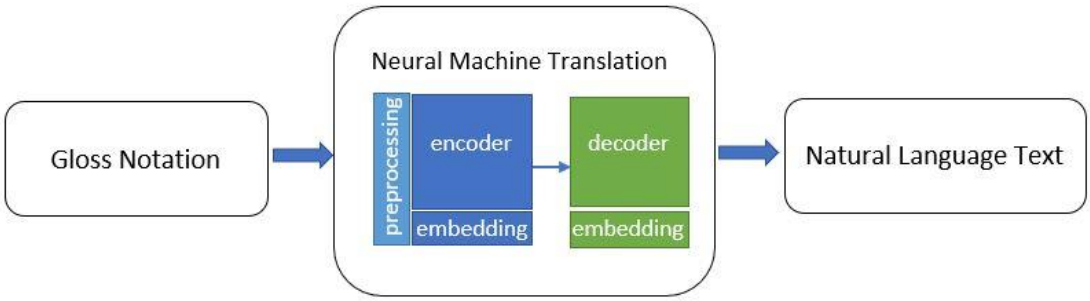

FIGURE 2. Sign Language Gloss to Natural Language Text Model

5. Experimental Results

The experimental results of the proposed approach for the ASL-PC12 corpus are shown in this section. Before showing the results, we first describe the corpus in depth. The corpus's criteria were used to characterize the corpus's criteria: the sentence, running words, vocabulary size, singletons, and out-of-vocabulary (OOV). Sentences represent the number of examples in the corpus. The number of words in the corpus is represented by the running words. Vocabulary size is several tokens that indicate how many words a model knows.

The number of words appearing only once in the training set is called singletons. The number of words appearing in test data but not training data is expressed as OOV. The ASLG-PC12 corpus [17, 18] was proposed as a large parallel corpus between English written texts and American Sign Language gloss. The ASLG-PC12 is an 87,709-sentence bilingual corpus. In addition to 8,542 singletons for English words and 6,133 for gloss words, the total number of" running words" is 1,034,532 for English words and 913,579 for gloss words. Both sign gloss annotation and spoken language have a vocabulary of 21,600 and 15,782, respectively. We divided the corpus into 82,709 sentences for training, 4,000 for validation, and 1,000 for testing in the experiments. The statistics of the corpus are described in table 2.

TABLE 2: Key statistics of ASLG-PC12

	English			**Gloss**		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
Sentences	82,709	4000	1000	82,709	4000	1000
Running Words	975,942	46,637	11,953	862,046	41,030	10,503
Vocab Size	21,600	5,634	2,609	15,782	4,323	2,150
Singletons	8,542	-	-	6,133	-	-
OOV	-	369	99	-	255	83

With batch size 2048, word embedding size 1024, sinusoidal positional encoding, recurrent layers having 512 hidden units, and Transformer feed-forward layers of 2048 hidden units, all of the suggested Transformer models are generated using PyTorch [25] and the OpenNMT library [26]. For optimization, we used Adam with 0.9 beta1 and 0.998 beta2, 0.1 dropouts, Noam learning rate schedule and gradient clipping with threshold 0, and 0.1 labels were smoothing.

5.1. Results

Tables 3 and 4 show the complete results of the proposed approach on ASLG-PC12 in both text-to-gloss and gloss-to-text translation directions.

TABLE 3: ASLG-PC12 Text to Gloss Model Results

	Test				
Layers	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**ROUGE**
1	98.40	97.75	97.12	96.47	98.46
2	**98.67**	**98.07**	**97.48**	**96.89**	**98.78**
4	98.54	97.88	97.26	96.62	98.68
6	97.89	97.15	96.46	95.76	98.44

TABLE 4: ASLG-PC12 Gloss to Text Model Results

	Test				
Layers	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**ROUGE**
1	91.83	87.51	84.70	80.14	94.99
2	**93.63**	**90.54**	**87.43**	**84.82**	**96.90**
4	92.76	88.43	85.64	82.15	95.52
6	91.60	87.28	83.44	79.86	94.44

When comparing the results of the trained text to gloss models, the encoder-decoder model with two layers achieves the best result, with a ROUGE score of 98.78% and a BLEU-4 score of 96.89%.

Furthermore, the results of the trained gloss-to-text models show that the encoder-decoder model with two layers achieves the best result, with a ROUGE score of 96.90% and a BLEU-4 score of 84.82%.

Table 5 shows our best results when translating ASLG-PC12 gloss to text and compares them to the best models in [11], [12]. To our knowledge, there are no similar models for the second direction of translation from text to gloss sign language.

TABLE 5: Comparison of Test score ASLG-PC12 for Gloss to Text with other works

	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**ROUGE**
Gloss2Text using RNN [12]	86.70	79.50	73.20	65.90	-
Gloss2Text using Transformer [11]	92.88	89.22	85.95	82.87	96.22
Proposed Gloss2Text model	**93.63**	**90.54**	**87.43**	**84.82**	**96.90**

6. Conclusion

This paper proposes a method for translating sign language to natural language and vice versa. We proposed a deep learning approach based on sequence to sequence with a transformer for bidirectional translation from gloss notation to text and text to gloss in both directions. We used self-attention and encoder-decoder. On the ASLG-PC12 corpus, we tested the proposed approach. For each translation direction, we created four encoder-decoder models with different layers. In each translation direction, we compared the outcomes of the four models. The model with two layers performed best using the ROUGE metric with a score of 96.90% and a BLUE-4 score of 84.82% when translating from gloss to text, according to the total experimental findings on eight different models applied to the ASLG-PC12 corpus. Also, when translating text to gloss, the model with two layers performed best, with a ROUGE score of 98.78% and a BLEU-4 score of 96.89%. Furthermore, compared to previous work on the same corpus in one direction of translation, some results show the superiority of the proposed models. The use of so-called position estimation, we think, would be a significant improvement in sign language translations. [27, 28, 29]. As a future study topic, the translation from text to pose estimation and vice versa is very interesting.

References

[1] Othman, A., & Jemni, M. (2017, December). An XML-gloss annotation system for sign language processing. In 2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA) (pp. 1-7). IEEE.

[2] Olusanya, B. O., & Newton, V. E. (2007). Global burden of childhood hearing impairment and disease control priorities for developing countries. The Lancet, 369(9569), 1314-1317.

[3] Othman, A., & Jemni, M. (2011). Statistical sign language machine translation: from English written text to American sign language gloss. arXiv preprint arXiv:1112.0168.

[4] Stokoe Jr, W. C. (2005). Sign language structure: An outline of the visual communication systems of the American deaf. Journal of deaf studies and deaf education, 10(1), 3-37.

[5] Hanke, T. (2004, May). HamNoSys-representing sign language data in language resources and language processing contexts. In LREC (Vol. 4, pp. 1-6).

[6] Thiessen, S. M. (2011). A grammar of SignWriting (Doctoral dissertation, University of North Dakota).

[7] Bungeroth, J., & Ney, H. (2004, May). Statistical sign language translation. In sign-lang@ LREC 2004 (pp. 105-108). European Language Resources Association (ELRA).

[8] López-Ludeña, V., San-Segundo, R., Montero, J. M., Córdoba, R., Ferreiros, J., & Pardo, J. M. (2012). Automatic categorization for improving Spanish into Spanish Sign Language machine translation. Computer Speech & Language, 26(3), 149-167.

[9] San-Segundo, R., Barra, R., Córdoba, R., D'Haro, L. F., Fernández, F., Ferreiros, J., ... & Pardo, J. M. (2008). Speech to sign language translation system for Spanish. Speech Communication, 50(11-12), 1009-1020.

[10] Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). Neural sign language translation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7784-7793).

[11] Yin, K., & Read, J. (2020, August). Attention is all you sign: sign language translation with transformers. In Sign Language Recognition, Translation and Production (SLRTP) Workshop-Extended Abstracts (Vol. 4).

[12] Arvanitis, N., Constantinopoulos, C., & Kosmopoulos, D. (2019, November). Translation of sign language glosses to text using sequence-to-sequence attention models. In 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (pp. 296-302). IEEE.

[13] Ko, S. K., Kim, C. J., Jung, H., & Cho, C. (2019). Neural sign language translation based on human keypoint estimation. Applied sciences, 9(13), 2683.
[14] Stoll, S., Camgoz, N. C., Hadfield, S., & Bowden, R. (2020). Text2Sign: towards sign language production using neural machine translation and generative adversarial networks. International Journal of Computer Vision, 128(4), 891-908.
[15] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.
[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[17] Othman, A., & Jemni, M. (2012, May). English-asl gloss parallel corpus 2012: Aslg-pc12. In 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon LREC.
[18] Othman, A., Tmar, Z., & Jemni, M. (2012, July). Toward developing a very big sign language parallel corpus. In International Conference on Computers for Handicapped Persons (pp. 192-199). Springer, Berlin, Heidelberg.
[19] Klima, E. S., & Bellugi, U. (1979). The signs of language. Harvard University Press.
[20] Cihan Camgoz, N., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. arXiv e-prints, arXiv-2003.
[21] Moryossef, A., Yin, K., Neubig, G., & Goldberg, Y. (2021). Data augmentation for sign language gloss translation. arXiv preprint arXiv:2105.07476.
[22] Zhang, X., & Duh, K. (2021, August). Approaching Sign Language Gloss Translation as a Low-Resource Machine Translation Task. In Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL) (pp. 60-70).
[23] Zhao, J., Qi, W., Zhou, W., Duan, N., Zhou, M., & Li, H. (2021). Conditional Sentence Generation and Cross-modal Reranking for Sign Language Translation. IEEE Transactions on Multimedia, 24, 2662-2672.
[24] Chen, Y., Wei, F., Sun, X., Wu, Z., & Lin, S. (2022). A simple multi-modality transfer learning baseline for sign language translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5120-5130).
[25] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.
[26] Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810.
[27] Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7291-7299).