Toward Optimum Transformer Model for Sequence-to-Sequence Data Transformation under Low-resource Computation Constraint

Yaya Heryadi and Cuk Tho
Computer Science Department, BINUS Graduate Program – Doctor of Computer Science
Bina Nusantara University,
Jakarta 11480, Indonesia
yayaheryadi@binus.edu, cuktho@binus.edu

Bambang Dwi Wijanarko
Computer Science Department, BINUS Online Learning
Bina Nusantara University,
Jakarta 11480, Indonesia
bwijanarko@binus.edu

Dina Fitria Murad
Information Systems Department, BINUS Online Learning
Bina Nusantara University
Jakarta 11480, Indonesia
dmurad@binus.edu

Kiyota Hashimoto
Interdisciplinary Graduate School or Earth System Science and Andaman Natural Disaster Management
Prince of Songkla University
Phuket Campus, Phuket 83120, Thailand
kiyota.h@phuket.psu.ac.th

Abstract

Accurate language translator applications running on low-resource computing devices such as smartphone is very instrumental to support tourism industry. The main challenge to achieve such objective is how to optimize performance of machine translation model targeted to limited resource of computing devices. Vanilla transformer model has been well known as one of state-of-the-art neural machine translation model. However, the drawback of this model is its large number of parameter models which might not be suitable for low-resource computing devices. This paper presents study findings in efforts to optimize 2 encoder-decoder stack depth of vanilla transformer by exploring several activation functions using fine-tuning approach. The pre-trained transformer model is fine-tuned using parallel corpus Bahasa Indonesia-Sundanese language to address machine translation task. The experiment results found that Sigmoid gives the highest model performance (0.993 average training accuracy and 0.987 average testing similarity) and GeLU gives the lowest model performance (0.987 average training accuracy and 0.980 average testing similarity) of the tested vanilla transformer models.

Keywords

Neural Machine Translation, transformer model, low-resource computing.

1. Introduction

Indonesia is one of many countries which are famous with natural beauty and rich cultural heritages and traditions that have attracted peoples from many countries to visit. However, the existence of many local languages used by active speakers living in Indonesia might become a challenge for foreigners who visit the country as a tourist to make effective but casual communications with local peoples. As a plural and language rich countries in the world, there are hundreds living languages are spoken in Indonesia. These indigenous (local) languages spoken in Indonesia are mostly used at the district and province or regional levels beside Bahasa Indonesia as the national language. Whilst several major local languages such as Javanese and Sundanese have million active speakers; in contrast, some local languages have active speakers ranging from only a few to several thousand peoples. Sundanese language has approximately 39 million native speakers in the western of Java (Banten and West Java Provinces). Sundanese ethnic group represent about 15% of Indonesia's total population. Therefore, Sundanese language is chosen as the target translation language as the second most active speakers in Indonesia after Javanese language(Lewis et al., 2014).

Despite the existence of local languages that makes many Indonesians as bilingual peoples, it is very common that people in rural areas would rather speak their local languages with people they know than speak Bahasa Indonesia. Therefore, a facility to translate text from international language such English or Bahasa Indonesia, as the national language, to a local language in Indonesia is very instrumental for tourism industry to help tourists communicate with local people living in the visited tourist spots. Unfortunately, as many Indonesian local languages have limited linguistic resources, such as parallel corpus, there is only limited number of studies have been published that explored machine translation involving Indonesian local languages.

A study reported by (Kellerman, 2008) concluded that provision of bilingual information for information localization in public services is crucial. Several reasons are assisting public facilities users for remembering environment, dictating spatial flow of users who have time constraints in using the public facility, reducing language barrier, and controlling safety of public facility environment. In line with these reasons, many studies have been reported to develop a computer-based system to translate Bahasa Indonesia to Indonesian local languages. The target beneficiary of the systems are both foreigners who only understand Bahasa Indonesia and local peoples who want to improve their understanding about Indonesian cultures. In addition, the proliferation of smartphone usage in Indonesia has become a consideration to choose a class of low-resource computing devices as the target of machine translation model.

As the first step to build a universal machine translation system from Bahasa Indonesia to several Indonesian local languages that runs on a low-resource computing devices, this study aims to explore the effect of activation function to performance of transformer model. Vanilla transformer model proposed by (Vaswani et al., 2017) is one of the state-of-the-art model in Natural Language (NL) processing. However, the large number of models parameter is prohibited to be implemented in low-resource computing devices. The long run aims of this study is to develop a transformer model for low resource computation device e.g., smart phone or cheap laptop. As a start step, for implication, no further definition of "low resource" is defined. As the first effort to explore a variant of the transformer model whose architecture can be simplified further to fit a low-resource (memory or computing power) computing device for further study, this study started to explore vanilla transformer model with low encoder-decoder stack depth (depth is 2) so that the model has relatively low number of model parameters among the other variants of vanilla transformer models. Although previous studies on exploring the role of activation functions have been reported (Glorot et al., 2011)·(Mishkin et al., 2017), the targeted model in the previous studies is more general.

1.1 Objectives

The objective of this study is to explore several activation functions in the transformer model as a model of interest as a neural machine translation to address a downstream task namely translating text from Bahasa Indonesia to Sundanese languages.

2. Literature Review

2.1 Machine Translation

Machine translation task aims to transform a given text in a source language to an equivalent text in a target language. Given a sequence of words with length n, $\boldsymbol{x} = [x_1, x_2, \dots, x_n]$ in a source language as input to a machine translation model, the model aims to produce a sequence of input words with length m, $\boldsymbol{y} = [y_1, y_2, \dots, y_m]$ in a target language. In the past several years, a plethora of machine translation models have been proposed which can be categorized into several categories including rule-based, statistical, hybrid, and neural machine translation methods (Chatzikoumi, 2020). A study by (Koehn & Knowles, 2017) showed some evidences that neural machine translation is the most practical machine translation approach. Machine translation for low-resource language is still consider a challenging task. Some prominent studies on machine translation for low-resources languages have been reported (Gangar et al., 2021),(Laskar et al., 2021),(Shirsath et al., 2021). Some works on neural machine translation from Bahasa Indonesia to Sundanese language also have been reported (Primandhika & Saifullah, 2021).

2.2 Sequence-to-Sequence Model

Sequence-to-sequence (seq2seq) model is a member of neural network model family which takes sequence of input and produces sequence as output (Sutskever et al., 2014). Some prominent studies on machine translation using seq2eq models have been reported by (Vaswani et al., 2017) to translate French to German. Seq2seq model used encode-decoder architecture. The encoder takes input sequence $\boldsymbol{x} = [x_1, x_2, \dots, x_n]$ and convert it into an internal sequence $\boldsymbol{z} = [z_1, z_2, \dots, z_n]$ as a semantic representation of input sequence. The decoder takes z as input and produces $\boldsymbol{y} = [y_1, y_2, \dots, y_m]$ as output sequence. The disadvantages of seq2seq model are: (1) the fixed-length context vector design is incapability of remembering long sentences, and (2) the beginning part of the input sequence are often forgotten once the processing has completed the whole input. To address this problem, (Bahdanau et al., 2014) added attention mechanism to seq2seq model.

2.4 Vanilla Transformer Model

Vanilla transformer model is proposed by (Vaswani et al., 2017) (see Figure 1) as a seq2seq model and successor of prominent Recurrent Neural Network model family including Long Short-term memory and Gated Recurrent Unit models. In contrast to previous Recurrent Neural Network models, a Transformer replaces recurrent building blocks with attention mechanism. This attention mechanism provides the transformer model capability to estimate global dependencies between input and output. In addition, the mechanism makes it possible for the transformer to process input sequences in parallel which reduces computation cost for transforming a sequence data to another equivalent sequence data.

The vanilla transformer model might be the first transformer model proposed by (Vaswani et al., 2017). The model is characterized as a recurrence-free, convolution-free, having attention mechanism to increase training speed and supporting parallel computation. The main component of the transformer model (see Figure 1) is: embedding layer, positional encoding, encoder stack, decoder stack, self-attention, feed-forward layer, residual connections and normalization, the final linear layer, and SoftMax (Sigmoid) layer.

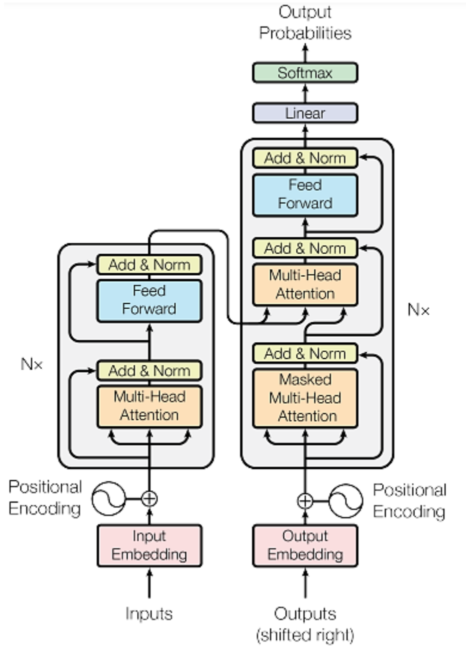
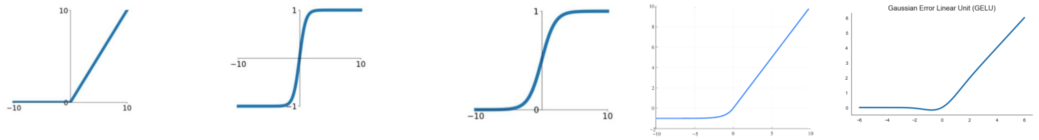


Figure 1. Architecture of Vanila Transformer Model (Vaswani et al., 2017)



(a) ReLU (b) Tanh (c) Sigmoid (d) ELU (e) GeLU

Figure 2. Several Activation Functions

The encoder part of vanilla transformer model consists of a stack of multi-head attention and feed-forward layer that process the input iteratively one layer after another. Each encoder layer generates and passes the encodings to the next encoder layer as inputs. The passed encoding coded information related to part of the inputs are relevant to each other. The decoder part consists of a stack of masked multi-head attention, multi-head attention, and feed-forward layer that process encoder's output iteratively one layer after another. In a transformer model architecture, the number of stacks in encoder and decoder is similar. (Figure 2)

Attention mechanism in a transformer model serves as a connection between the encoder and decoder parts which works as follows: it looks at input sequence to decide at each step which other parts of the input sequence that are important; finally, the data from attention are used as additional information to the decoder. In the vanilla transformer model, (Vaswani et al., 2017) proposed to use multi-head self-attention mechanism in which the encoded representation of the input is a set of key-value pairs, (K,V), both has dimension n (input sequence length). In the decoder part, the previous output is compressed into a query (Q of dimension m) and the next output is produced by mapping this query and the set of keys and values. The attention score function is scaled dot-product attention which can be formulated as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{n}}\right)V, \tag{1}$$

The vanilla transformer model implements multiread self-attention comprises of h dot product attention layers run in parallel. The multiread self-attention mechanism as represented in Equation (2) make it possible for the transformer model to jointly attend to information from different representation subspaces at different position.

$$MultiHead(Q,K,V) = concat(head_1, .., head_h)W^0, \tag{2}$$

where: $head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$, $W^0 \in \mathbb{R}^{d_{model} \times hd_v}$, $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, $h = 8$, and $d_k = d_v = \frac{1}{h} d_{model} = 64$.

The point-wise feed-forward network is an important part of the transformer model as it has two-thirds of the model parameters (Geva et al., 2020). The point-wise feed-forward networks takes 3-dimensional input shape (batch size, sequence length, and feature size). The network comprises of two dense layers that applies to the last dimension. Each layer acts as key-value memories where each key associated with textual patterns in the training dataset and each values induces a distribution over the output vocabulary. Architecture of the feed-forward layer is a fully connected neural network in which each neuron has an activation function. Several activation functions which are widely used in transformer models are: REL, Tanh, Sigmoid, ELU, and GELU. The number of model parameters in feed-forward layers almost two-thirds of a transformer model's parameters; however, to the best of our knowledge, the role of the feed-forward layer in the transformer model is still under-explored. Therefore, the objective of this study is to explore the effect of activation function in the feed-forward layer to performance of the whole transformer model.

3. Methods

3.1. Dataset

The data sources for this study comprises ORCAS dataset; dataset scraped from official airline websites such as: Garuda (https://www.garuda-indonesia.com/), NAM (https://www.flynamair.com/), Asia Air (https://www.airasia.com/), Lion Air (https://www.lionair.co.id/), and Batik Air (https://www.batikair.com); kereta api indonesia (https://www.kai.id); su.wikipedia.org; id.wikipedia.org; several local government websites. The raw data are written in either Indonesian or

Sundanese languages with various length. Next, a Bahasa Indonesia-Sundanese parallel corpus is prepared manually by bilingual linguists who understand both Bahasa Indonesia and Sundanese language. The final input dataset comprises of 38,712 samples of parallel sentence in Indonesian and Sundanese languages.

Data preprocessing in this study (see Figure 3) are including standard processes such as: converting the input text to lowercase, removing double whitespaces, removing new line characters, removing non ASCII characters, and removing single space remaining at the beginning and end of the text, and tokenization. Following (Vaswani et al., 2017), input text is represented using embedding layer which is learned on the fly by the transformer model. The purpose of this embedding technique is to ensure that each word can be mapped to a vector properly without missing out any word in the input text.

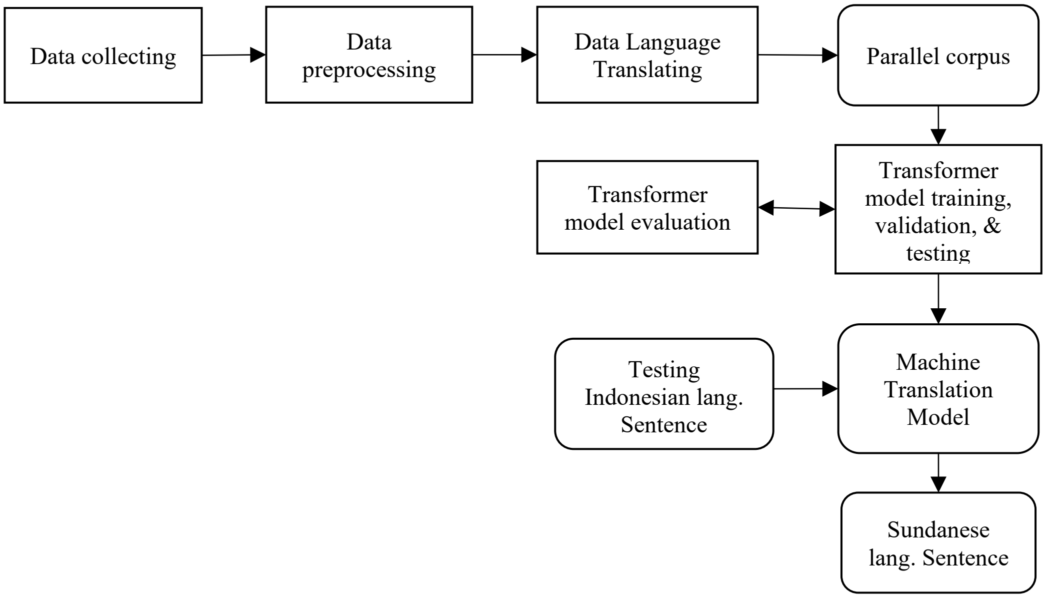

Data collecting
Data preprocessing
Data Language Translating
Parallel corpus
Transformer model training, validation, & testing
Transformer model evaluation
Testing Indonesian lang. Sentence
Machine Translation Model
Sundanese lang. Sentence

Figure 3. Process Flow of the Experiment

3.2. Fine-tuning Process of the Transformer Model

Basic neural machine translation model used in this study is the pre-trained vanilla transformer model proposed by Vasmani et al., (Vaswani et al., 2017). Transfer learning in this study aims to adjust the pre-trained model with language translation from Bahasa Indonesia to Sundanese as the new task. The transfer learning is implemented as follows. First, complexity of the pre-trained model architecture is reduced by only using stack depth = 2 for each encoder and decoder. Second, activation function in feed-forward layer is tested by one of the following activation functions namely: Rectified Linear Unit (ReLU), Sigmoid, Tanh, Exponential Linear Unit (ELU), and Gaussian Error Linear Unit (GELU). Finally, the model is fine-tuned using Adam optimization algorithm and parallel corpus data to adjust the model parameters for language translation task.

Each of the transformer model in this study is fine-tuned using 38,324 (99 %) parallel samples which are chosen randomly from the input dataset and used as training and validation dataset and 388 (1 %) parallel samples as testing dataset of parallel corpus Indonesian and Sundanese languages. In this study, maximum length of the tested sentence is 100 words. The model is retrained in 250 epochs. During fine-tuning process, the model performances are measured using several metrices namely: training accuracy, validation accuracy, training loss, and validation loss. Finally, the fine-tuned model is tested using the testing dataset. For simplicity, similarity between predicted and actual texts is measured using cosine distance function.

4. Results and Discussion

Model performance of the transformer model from fine-tuning process can be summarized in Table 1.

Table 1. Summary of Model Performance Metrics

Activation Function	n	Average Training Accuracy	Average Validation Accuracy	Average Training Loss	Average Validation Loss	Average Testing Similarity
Sigmoid	19	0.993	0.980	0.084	0.134	0.987
Tanh	19	0.992	0.981	0.048	0.109	0.984
ReLU	19	0.990	0.971	0.061	0.232	0.985
ELU	19	0.989	0.970	0.067	0.248	0.983
Gelu	19	0.987	0.957	0.087	0.297	0.980

In this study, non-parametric statistics technique is used for testing hypothesis because of the following reasons: (1) the sample size is relatively small ($n = 5$) due to high computation workload during training of the Vanilla Transofmer model which takes almost 10 hours to complete 300 epoch in order to obtain significant convergence of training error, and (2) non-parametric tests typically have fewer assumptions about the data distribution such as normality. The result of significance test to performance metrics of the model fine-tuning process can summarized as follows.

1) The results of Mann-Whitney U-test for hypothesis testing method with sample size $n = 19$ and 95% confidence interval for each vanilla transformer model with stack depth = 2 can be summarized in Table 2. The null hypothesis (H_0) for this test is that there is no different of average training accuracy between the two models with the tested activation function. The alternative hypothesis (H_1) for this test is that one of the activation functions of the tested model gives higher model average training accuracy than the other.

   As can be seen from the Table 2, it can be concluded that the highest average training accuracy given by the model that uses Sigmoid in the feed-forward layers; whilst the lowest average training accuracy given by the model that uses GeLU activation function. The significance test showed that average training accuracy given of the model with GeLU activation is significantly different from that given by other activation functions.

Table 2. Significance Test on Average Training Accuracy ($\boldsymbol{n = 19}$, $\boldsymbol{\alpha = 0.05}$)

Activation Function	Sigmoid	Tanh	ReLU	ELU	GeLU	Average Training Accuracy
Sigmoid		nS	S	S	S	0.993
Tanh			nS	S	S	0.992
ReLU				nS	S	0.990
ELU					nS	0.989
GeLU						0.987

Note: nS: not significant, S: significant

Table 3. Significance Test on Average Testing Similarity ($\boldsymbol{n = 19}$, $\boldsymbol{\alpha = 0.05}$)

Activation Function	Sigmoid	Tanh	ReLU	ELU	GeLU	Average Training Similarity
Sigmoid		nS	nS	S	S	0.987
Tanh			nS	S	S	0.984
ReLU				nS	S	0.985
ELU					S	0.983
GeLU						0.980

Note: nS: not significant, S: significant

2) The results of Mann-Whitney U-test for hypothesis testing method with sample size $n = 19$ and 95% confidence interval for each vanilla transformer model with stack depth = 2 can be summarized in the Table 3. The Null hypothesis (H_0) for this test is that there is no different of average testing similarity of the two model with the tested activation function. The alternative hypothesis (H_1) for this test is that one of the activation functions of the tested model give higher model average testing similarity than the other. As can be seen from Table 3, the significance test results showed there are four pair of activation functions that give almost similar average testing similarity to the tested transformer model namely: Sigmoid-Tanh, Tanh-ReLU, ReLU-ELU, and ELU-GeLU. In addition, as can be seen from the Table 3, it can be concluded that the highest average testing similarity given by the model in which the feed-forward layers use Sigmoid; whilst, the lowest average testing similarity given by GeLU activation function. The average testing similarity given by GeLU activation is significantly different from that given by other activation functions. The results is consistent with the previous significance test summarized in Table 2.

This experiment findings validate the choice of Vashmani et al., (Vaswani et al., 2017) to use Sigmoid for feed-forward layers as the highest performed activation functions among the tested activation functions in this study. However, this experiment findings disagree with previous study reported by (Glorot et al., 2011) which concludes that ReLU has better performance than Sigmoid and Tanh activation function. Some results of the model training are shown in Table 4 and Table 5. Table 4 shows several samples of testing results which are correctly predicted by the fine-tuned models.

Table 4. Some Predicted Results which are Similar with the Target Sentence

No	Sentences
1	(Id) *Kami meminta anda untuk memasang sabuk pengaman anda saat ini, dan simpan semua koper di bawah kursi atau di kompartemen atas.* (Su) *Sim kuring nyuhunkeun panumpang sadaya supados mageuhkeun sabuk pengaman sareng nyimpen sadaya koper dina handapeun korsi atanapi dina kompartemen luhureun korsi pangcalikan.* (En) We ask that you fasten your seatbelt at this time, and store all luggage under the seat or in the upper compartment.
2	(Id) *Penumpang kereta Parahyangan yang kami hormati, selamat malam dan selamat datang di Bandung.* (Su) *para panumpang kareta parahyangan anu dipihormat , wilujeng wengi sareng wilujeng sumping ka bandung .* (En) Dear Parahyangan passengers, good evening and welcome to Bandung.
3	(Id) *Kami berharap bisa berjumpa dengan anda lagi dalam penerbangan dalam kesempatan yang akan datang.* (Su) *Pamugi urang pendak deui dina kasempetan salajengna.* (En) We hope to see you again on flights in the near future.
4	(Id) *Tolong keluarkan perangkat elektronik dari dalam tas Anda.* (Su) *Kaluarkeun alat elektronik tina lebet tas Bapa/Ibu.* (En) Please take the electronic device out of your bag.
5	(Id) *AirAsia menawarkan tiket bus murah ke beberapa tempat paling eksotis di Asia.* (Su) *airasia nawiskeun tiket beus anu hargana kahontal pikeun ngajugjug ka sababaraha tempat anu paling endah di asia .* (En) AirAsia offers cheap bus tickets to some of the most exotic places in Asia.
6	(Id) *Untuk keselamatan dan kenyamanan Anda, mohon tetap duduk dengan memakai sabuk pengaman Anda.* (Su) *Supados Bapa/Ibu ngaraos aman sareng sugema, supados tetep calik sareng nganggo sabuk pengaman.* (En) For your safety and comfort, please remain seated with your seat belt on.

However, the fine-tuned model with the dataset used in this study is unable to translate several sentences precisely although semantically the translated sentences are still closed to the input sentences.

Table 5. Some Predicted Results which are not Closely Similar with the Target Sentence.

No	Sentences
1	(Id) *selamat datang* (Su-predicted) *wilujeng sumping ibu / bapa .* (Su-target) *wilujeng sumping* (En) Welcome
2	(Id) *Enak nih, sore-sore ngopi santuy bareng temen-temen!* (Su-predicted) *raos , ngopi sore sareng babaturan !* (Su-target) *enak , ngopi sore sareng babaturan !* (En) It's good, afternoon coffee with friends!
3	(Id) *Selamat pagi.* (Su-predicted) *wilujeng enjing para panumpang.* (Su-target) *wilujeng enjing.* (En) Good morning
4	(Id) *Selamat malam Bapak/Ibu* (Su-predicted) *wilujeng wengi bapa sareng ibu anu dipihormat.* (Su-target) *Wilujeng wengi bapa/ibu.* (En) Good evening ladies and gentlement.

5. Conclusion

The experiment results found that Sigmoid activation function used in the feed-forward layers gives the highest vanilla transformer model performance (0.993 average training accuracy and 0.987 average testing similarity) and GeLU gives the lowest model performance (0.987 average training accuracy and 0.980 average testing similarity) of the tested vanilla transformer models. This experiment findings validate the choice of Vashmani et al., (Vaswani et al., 2017) to use Sigmoid for feed-forward layers as the best activation functions among the tested activation functions. However, this experiment findings disagree with previous study reported by (Glorot et al., 2011) which concludes that ReLU achieves better performance than Sigmoid and Tanh activation function. This analysis results can be used further as a basis for future works in developing a variant of transformer model which is suitable for low-resource computing environment.

Acknowledgements

The authors would like to thank Professor Kiyota Hashimoto from Interdisciplinary Graduate School or Earth System Science and Andaman Natural Disaster Management (ESSAND), Prince of Songkla University, Phuket Campus, Thailand for providing High Performance Computing facilities. This study is funded by Binus University under International Research Grant No: 017/VR.RTT/III/2021

References

Bahdanau, D., Cho, K., and Bengio, Y., Neural machine translation by jointly learning to align and translate, *ArXiv Preprint ArXiv:1409.0473*, 2014.

Chatzikoumi, E., How to evaluate machine translation: A review of automated and human metrics, *Natural Language Engineering*, vol. 26(2), pp. 137–161, 2020.

Gangar, K., Ruparel, H., and Lele, S., Hindi to english: Transformer-based neural machine translation, *International Conference on Communication, Computing and Electronics Systems*, pp. 337–347, Singapore, 2021.

Geva, M., Schuster, R., Berant, J., and Levy, O., Transformer feed-forward layers are key-value memories, *ArXiv Preprint ArXiv:2012.14913,* 2020.

Glorot, X., Bordes, A., and Bengio, Y., Deep sparse rectifier neural networks, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323, Fort Lauderdale, FL, USA 2011.

Kellerman, A., International airports: Passengers in an environment of 'authorities.', *Mobilities*, vol. *3*(1), pp. 161–178, 2008.

Koehn, P., and Knowles, R., Six challenges for neural machine translation, *ArXiv Preprint ArXiv:1706.03872,* 2017.

Laskar, S. R., Pakray, P., and Bandyopadhyay, S., Neural Machine Translation for Low Resource Assamese--English, *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU,* pp. 35, *Shillong, India*, *170*, 2021.

Lewis, M. P., Simons, G. D., and Fennig, C. D., *Ethnologue: Languages of Asia (pp. 1--558), SIL International*. Global

Publishing, 2014.

Mishkin, D., Sergievskiy, N., and Matas, J., Systematic evaluation of convolution neural network advances on the imagenet. *Computer Vision and Image Understanding*, vol. *161*, pp. 11–19, 2017.

Primandhika, R. B., and Saifullah, M. N. M. A. R., Experiment on a Transformer Model Indonesian-to-Sundanese Neural Machine Translation with Sundanese Speech Level Evaluation, *Thirteenth Conference on Applied Linguistics (CONAPLIN 2020)*, pp. 452–459, Atlantis Press, 2021.

Shirsath, N., Velankar, A., Patil, R., and Shinde, S., Various Approaches of Machine Translation for Marathi to English Language, *ITM Web of Conferences*, vol. *40*, pp. 3026, 2021.

Sutskever, I., Vinyals, O., and Le, Q. V., Sequence to sequence learning with neural networks, *ArXiv Preprint ArXiv:1409.3215,* 2014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., Attention is all you need, *ArXiv Preprint ArXiv:1706.03762*, 2017.

Biographies

Yaya Heryadi is a lecturer, researcher, and Certified Data Scientist at the Doctor of Computer Science (DCS) Department, Binus Graduate Program, Bina Nusantara University with research interests in Artificial Intelligence, Data Science, Machine Learning/Deep Learning, Natural Language Processing, and Computer Vision. He holds a bachelor's degree in Statistics and Computing from the Bogor Agricultural Institute, a Master of Science from Indiana University at Bloomington, USA, a Doctorate in Computer Science from the University of Indonesia, and. During his career as a researcher, he has attended lectures at the University of Kentucky at Lexington, USA, and the sandwich-like program at Michigan State University at East Lansing, USA.

Bambang Dwi Wijanarko is a lecturer and researcher at Binus Online Learning, Bina Nusantara University with research interests in Natural Language Processing and Machine Learning/Deep Learning. He holds a Doctorate degree in Computer Science from Bina Nusantara University.

Dina Fitria Murad is a lecturer and researcher at Binus Online Learning, Bina Nusantara University with research interests in Natural Language Processing and Machine Learning/Deep Learning. She holds a Doctorate degree in Computer Science from Bina Nusantara University.

Cuk Tho is a lecturer and researcher at Bina Nusantara University with research interests in Natural Language Processing. Currently, she is now taking a Doctorate program in Computer Science at Bina Nusantara University.

Kiyota Hashimoto is a lecturer and researcher at Earth Science, Graduate School, Faculty of Technology and Environment, Prince of Songkla University, Phuket Campus, Thailand. He is also a Collaborative Professor at Kanazawa University, Japan. He holds Bachelor of Arts, and Master of Arts degree ib Linguistics from Kyoto University, Japan; and Doctor of Engineering in Information Science from Nara Institute of Science and Technology. His expertise covers Natural Language Processing, Artificial Intelligence, Machine Learning (including Deep Learning), Data Science (main target domains are environment, tourism, education), Research Methodology; Higher Education Reforms; and Theoretical Linguistics.