The Effect of Encoder and Decoder Stack Depth of Transformer Model to Performance of Machine Translator for Low-resource Languages

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

Automated language translator has wide potential applications especially in plural language countries. Study on low-resource languages is very crucial such as making information accessible to people live in less-connected and technologically underdeveloped areas, making more digital content available, and making Natural Language Processing models more accessible to low-resource languages. Vanilla transformer model has achieved excellent performance to address machine translation task. Despite its high performance, the model contains adjustable hyperparameters such as the number of encoder-decoder stack depth. This paper presents exploration results on the effect of encoder-decoder stack depth to performance of the vanilla transformer model as a neural machine translation of Bahasa Indonesia-Sundanese languages. The empiric results of fine-tuning a pretrained vanilla transformer model showed that average performances of vanilla transformer model with 2, 4, or 6 stack depth are higher than average performance of the model with 8 stack depth. The highest performances achieved by the transformer model with 2 stack depth are: 0.99 average training accuracy, 0.97 average validation accuracy, and 0.99 average testing similarity. Interestingly, according to non-parameteric significance test results with 95% confidence interval, there is no siginificant difference on performance of vanilla transformer model with 2, 4, 6, and 8 stack depths. These results showed that using vanilla transformer with less number of depth stack is favourable for machine translation as it has less number of model parameters but it gives acceptable model performance. From experimentation results, it showed that vanilla transformer model with 2 stack depth is potential to be explored further.

Keywords
Neural Machine Translation, Transformer Model.

1. Introduction
Machine translation task has achieved wide research interest. Many of the prominent successful machine translation models are based on Recurrent Neural Networks (Graves, 2012)·(Siddique et al., 2021) and Long Short-term Memory model families (Hegde et al., 2021)·(Minh et al., 2021). The next generation of neural machine translation models which achieved high performance are sequence-to-sequence model families (Bahdanau et al., 2014)·(Sutskever et al., 2014). One sub-class of the sequence-to-sequence models is transformer models that showed better performance than its predecessor sequence-to-sequence models to address machine translation task (Raganato et al., 2018)·(Liu et al., 2020)·(Kasai et al., 2020). A vast number of Natural Language Processing (NL) researches that have been published, including neural machine translation, mostly involved high-resource languages, such as English due to high availability of its linguistic resources but only a few of published research on machine translation involving languages with limited linguistic resources (low-resource language). The main reason is that neural machine translation model requires sufficient quantity and quality of dataset for training the model.

According to (Joshi et al., 2019), studies on low-resource languages are very crucial for several reasons namely: making information accessible to people live in less-connected and technologically underdeveloped areas, making more digital content available, and making NL models more accessible to low-resource languages. Ranathuga et al., (Ranathunga et al., 2021) have analyzed a number of reports on research advancements in low-resource language using neural machine translation models and some analysis to identify feasible neural machine translation techniques to address the task. One feasible strategy to build a neural machine translation for low-resource language is to use a pre-trained model and use transfer leaning approach to fine-tune the model using availabel sample dataset of low-resource language to address a downstream task, e.g., machine translation.

Indonesia is a plural and language rich country. With many local languages used in many regions beside the national language: Bahasa Indonesia, many Indonesian peoples are bilingual, speak both Bahasa Indonesia and local languages. Sundanese language is the second major local language in Indonesia after Javanese language with 32 million (15 percent of Indonesian Population) active speakers mostly live in West Java provinces. In many rural areas many Indonesian peoples mostly speak local language in their daily communications. Hence, automated machine translation is very instrumental in many areas of Indonesia as a supporting facility in many public services to deliver Government's official announcement or use in signages.

Transformer model is a sequence-to-sequence (seq2seq) model based on encoder-decoder architecture which is firstly proposed by (Vaswani et al., 2017) which becomes a state-of-the-art model to solve Natural Language Processing (NL) problems. The model has been firstly demonstrated by (Vaswani et al., 2017) for translating text from English to German with high accuracy. Following (Vaswani et al., 2017) many publications on the use of transformer model in many fields have been reported such as: drug research (Grechishnikova, 2021), emotional classification (X. Wang & Tong, 2021), and speech recognition (Y. Wang et al., 2021). The advent of transformer model to address several NL tasks resulted in many proposed models available in literature. The common properties of these model are the deeper the transformer model structure, the more accurate the model will be (Simonyan & Zisserman, 2014)·(He et al., 2016)·(Chen et al., 2017). In contrast to pursue higher accuracy models, some other researches focus more on finding optimum architecture to optimize the use of computation facilities for training and testing models (Araabi & Monz, 2020)·(Ma et al., 2020). In many practical applications of the transformer models, a light structure of transformer model is favorable as it reduces the number of model parameters to be trained and less space needed to run the model. The study by (Narang et al., 2021), for example, has concluded various factors contributes to performance of the transformer model such as: the depth of the transformer model structure and activation function.

Although, many previous studies have explored the effect of the depth of model structure to performance of neural network models, for example by (Adil et al., 2020), architecture of the neural network model under study are mostly model with shallow structure. Whilst there is no guideline to choose the number of layers for transformer model architecture, exploration on various layer of the transformer model for particular application is very crucial.

1.1 Objectives
The objective of this study is to explore the effect of several stack depth of the encoder and decoder of the transformer model to its performance as a neural machine translation using a parallel corpus of Indonesian and Sundanese languages.

2. Literature Review
2.1 The Need of Automated Language Translation
Bilingual information in public services in Indonesia has been provided in many forms such as announcement by announcer and posted signage (road signs, announcement boards). However, provision of bilingual information in public services maintained by local authorities is crucial for many reason such as to: assist facility users for remembering environment, dictate spatial flow of users who have time constraints in using the facility, reduce language barrier, and control safety of facility environment (Kellerman, 2008). Localization of standard or official information using local language has been practiced by local government and used in many areas of Indonesia. For example: the use of bi/multilingual announcements in some international airports in Indonesia to deliver information. Among those airports are Sam Ratulangi International Airport in Manado, and Kualanamu International Airport in Medan have used several languages including English, Arabic, Chinese, Bahasa Indonesia, and local languages. Adisutjipto International Airport in Jogyakarta, and Juanda International Airport in Surabaya have used English, Bahasa Indonesia, and Javanese languages (Susanti, 2018). Moreover, some announcement in new Jogyakarta International Airport a written in old Javanese script.

Translating official announcements to local languages, however, is not an easy task. The semantic of the announcement in a source language and the target language should be equivalent. Moreover, as many announcements should be delivered from many Government agencies quickly, the translation tasks will no longer efficient to be handled manually. Therefore, one of the main objective of this study is to develop a robust transformer-based model as machine translation model to translate official announcement from Indonesian to Sundanese language as the second most active speakers in Indonesia after Javanese language (Lewis et al., 2014). This study can be an initial step to replicate the machine translation from text input in Bahasa Indonesia to several local languages in Indonesia.

2.2 Neural Machine Translation
Machine translation which is a subfield of NL aims to find a model that maps a sequence of input words or text in a languages to an equivalent sequence of target words in another languages. Given a sequence of input words with length n, $\boldsymbol{x} = [x_1, x_2, \dots, x_n]$, a machine translation task is to produce a sequence of input words with length m, $\boldsymbol{y} = [y, y_2, \dots, y_m]$. In the past several years, a vast number of machine translation approaches have been proposed. According to (Chatzikoumi, 2020), various machine transaction can be categorized broadly into: rule-based, statistical, hybrid, and neural machine translation. Several studies showed some evidences that Neural Machine Translation is the most feasible machine translation approach (Koehn & Knowles, 2017). In particular the study reported by (Koehn & Knowles, 2017)pointed out the main challenges of neural machine translation approach which potentially reduce its performance namely: different domains, low-resource language, low-frequency words, long sentences, word alignment model, and beam search decoding. Some studies on machine translation for low-resources languages have been reported. For example automated translator from Japanese, Lao, Malay, and Vietnamese aligned to English (Rubino et al., 2020), Hindi to English (Gangar et al., 2021)·(Dave et al., 2001), Assamese to English (Laskar et al., 2021), Marathi to English (Shirsath et al., 2021). In particular, some works on neural machine translation from Bahasa Indonesia to Sundanese language have been reported, for example study report by (Primandhika & Saifullah, 2021).

2.3 Sequence-to-Sequence Models
Sequence-to-sequence (seq2seq) models are neural network-based family models that take sequence of input and produces sequence as output. This model is very instrumental to build a machine translation as both input and output are a sequence of words. For example, a machine translation from French to German (Vaswani et al., 2017) which takes a sentence in French as input and produces a sentence in German with similar meaning. Seq2seq model can be used to model a language (Sutskever et al., 2014). Architecture of the seq2seq model is typically comprises an encoder-decoder such that: the encoder takes input sequence and convert it into an internal sequence (context vector or sentence embeddings).

Given input representation $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$, the encoder produces $\boldsymbol{z} = [z_1, z_2, \ldots, z_n]$ as internal representation sequence. The internal sequence is expected to summarize the meaning of the input sequence. The decoder takes z sequence as input and produces $\boldsymbol{y} = [y_1, y_2, \ldots, y_m]$ as output sequence. However, the seq2seq model has several the disadvantages namely: fixed-length context vector is incapable to remembering long sentences, and the beginning part of the input sequence are often forgotten once the processing has completed the whole input. To address this problem, Vasmani et al., (Vaswani et al., 2017) proposes attention mechanism to improve the seq2seq model.

2.4 Vanilla Transformer Model

The first transformer model or vanilla transformer model which is proposed by (Vaswani et al., 2017) is a neural machine translation model designed as an extended version of seq2seq by adding an attention mechanism. The transformer model replaces recurrence links in previous seq2seq model with an attention mechanism to estimate global dependencies between input and output. The transformer model has to components. *First*, the encoder part consists of multi-head attention and feed-forward layer which are stacked on top of each other several times 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. *Second*, the decoder part, on the other hand, consists of masked multi-head attention, multi-head attention, and feed-forward layer which are stacked on top of each other several times that process encoder's output iteratively one layer after another. In a transformer model architecture, the depth of stacks in encoder and decoder is similar.

In the vanilla transformer model (see Figure 1), the encoder and the encoder stacks have similar depth. Various stack depths in the transformer models have been reported, for example: a variant of BERT transformer model typically has 12 or 24 stack depth (Devlin et al., 2018). Vasmani et al., (Vaswani et al., 2017) proposed 6 depth for both encoder and decoder part of the vanilla transformer model. However, there is no explanation in their study report on how the stack depth is chosen. Although there are no study results that conclude the most optimum stack depth should encoder and decoder have, some previous studies showed that at some points the number of layers in encoder and decoder of a transformer model affects the transformer model performance.

Another important component of the transformer performance is attention mechanism which serves as a connection between the encoder and decoder parts and 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. Several attention scoring functions have been proposed such as: additive attention (Bahdanau et al., 2014), scaled dot product (Vaswani et al., 2017), content-based attention (Graves et al., 2014), local and hard attention (Luong et al., 2015), general attention (Luong et al., 2015), and dot-product attention (Luong et al., 2015).

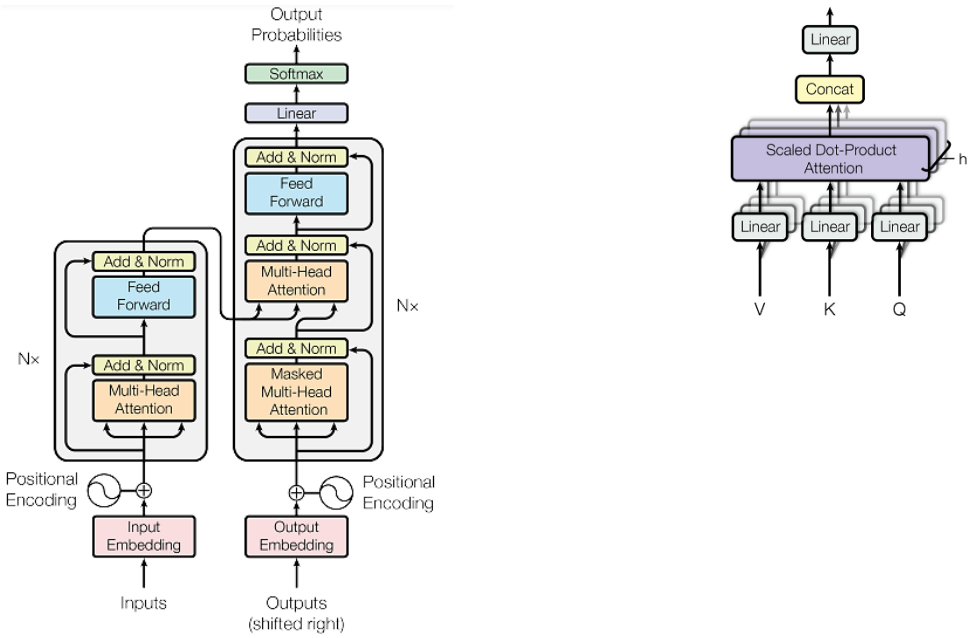

Figure 1. (Left) Architecture of Vanila Transformer Model and (Right) Attention Mechanism (Vaswani et al., 2017)

In particular, Vaswani et al., (Vaswani et al., 2017) proposed to use multi-head self-attention mechanism for the vanilla transformer model. In this attention mechanism, the encoded representation of the input is a set of key-value pairs, (K,V), both has dimension n (input sequence length). In the context of machine translation, both the keys and values are the encoder hidden states. 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 in attention mechanism is implemented as scaled dot-product attention which can be formulated as:

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

The multi-head self-attention mechanism (see FIGURE. 1(b)) comprises of h dot product attention layers run in parallel. The multihead self-attention mechanism as represented in the following equation makes it possible for the transformer model to jointly attend to information from different representation subspaces at different position.

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

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 $d_k = d_v = \frac{1}{h} d_{model} = 64$.

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. The parallel corpus of Bahasa Indonesia-Sundanese languages for this study is prepared manually by bilingual linguists who understand both Indonesian 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 are including: 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. (Figure 2)

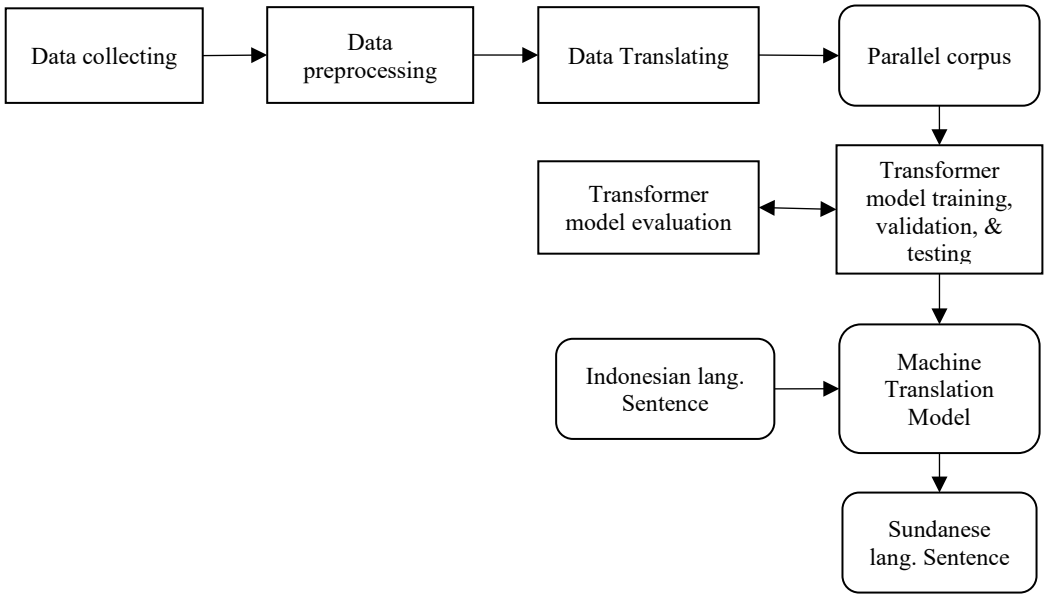

Figure 2. Process Flow of the Experiment

3.2. Experiment Setup

The main object in this study is the pre-trained vanilla transformer model proposed by Vasmani et al., (Vaswani et al., 2017). The transformer model implementation is taken from Hugging Face's transformers library (Wolf et al., 2020). Based on the study by (Kitaev et al., 2020) who reported that training transformer-based architectures takes a quite times when the input data comprises of long sentences. Therefore, maximum length of the tested sentence in this experiment is 100 words.

In this study, it is hypothesized that performance of the transformer model is predominantly by the encoder and decoder stack depth. Unit analysis of this experiment is the pre-trained vanilla transformer models which are fine-tuned with parallel Bahasa Indonesia-Sundanese language samples. Each transformer model is set out with encoder-decoder stack having 2, 4, 6, and 8 stack depth. Following (Vaswani et al., 2017), for each stack depth the feed-forward layer of the transformer model is set with Sigmoid activation function. Each transformer model setting is executed four times ($n = 5$).

Each of the transformer model in this study is fine-tuned using 38,324 (99 percent) parallel samples as training and validation dataset and 388 (1 percent) parallel samples as testing dataset of parallel corpus Indonesian and Sundanese languages. The model is retrained in 300 epochs using Adam optimization algorithm. The model performance metrics measured in this experiment are: average training accuracy, average validation accuracy, average training loss, and average validation loss. For simplicity, testing similarity between predicted and actual sentences is measured using cosine distance function. Model performances are analyzed using non-parameteric statistical hypothesis testing methods with 5% confidence level (α) or 95% confidence interval.

4. Results and Discussion

As can be seen from Table 1, as expected, the number of transformer model parameters increases with the increasing stack depth. Consequently, computing time for estimating the model parameters increased monotonically with the number of model parameters. However, it is interesting to find that performance of the transformer model does not increase linearly with the increasing number of model parameters.

Table 1. Summary of Trainable Model Parameters

Number of Stacks	Number of Parameters
2	4,217,077
4	5,142,773
6	6,068,469
8	6,994,165

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 results of hypothesis testing are as follows.

1) Mann-Whitney U-test is used to measure the significance of the encoder and decoder stack depth to performance of the transformer model with the sample size $n = 5$ and 95% confidence interval for each tested model. The null hypothesis (H_0) for this test is that the stack depth does not affect average training accuracy of the tested models. The alternative hypothesis (H_1) for this test is that there is a the stack depth gives higher average training accuracy to the tested model than the other. The result of the significance test (see Table 3) showed that each stack depth gives significant effect to average training accuracy of the model. In particular, the depth = 2 of encoder-decoder stack gives the highest average training accuracy; whilst, the depth = 8 of encoder-decoder stack gives the lowest average training accuracy to the tested models. This results is different from the study result reported by Vasmani et al., (Vaswani et al., 2017) who proposed vanilla transformer architecture with 6 stacks depth.

Table 2. Summary of Model Performance Metrices ($\boldsymbol{n = 5}$)

Stack Depth	Average Training Accuracy	Average Validation Accuracy	Average Training Loss	Average Validation Loss	Averagae Testing Similarity.
2 stacks	0.993	0.980	0.084	0.134	0.987
4 stacks	0.970	0.913	0.171	0.527	0.985
6 stacks	0.980	0.928	0.120	0.444	0.974
8 stacks	0.803	0.741	0.699	1.001	0.717

Table 3. Significance Test on Average Training Accuracy ($\boldsymbol{n = 5}$, $\boldsymbol{\alpha = 0.05}$)

Stack Depth	2	4	6	8	Average Training Accuracy
2		S	S	S	0.993
4			S	S	0.970
6				S	0.980
8					0.803

Note: S: significant

2) Mann-Whitney U-test is used to test hypothesis with the sample size $n = 5$ and 95% confidence interval for each tested model. The null hypothesis (H_0) for this test is that the stack depth does not affect average testing similarity of the tested models. The alternative hypothesis (H_1) for this test is that there is a stack depth gives higher average testing similarity to the tested model than the other. The result of the significance test (see Table 4) showed that there is no signifficant different between average testing similarity of the model with stack depth 2 and 4. However, average testing similarity between the model with stack depth 2 or 4 with the model with the other stack depths is significantly different. From this experiment results, the vanila transformer model with the stack depth = 8 of encoder-decoder achieves the lowest average testing similarity.

Table 4. Significance Test on Average Testing Similarity ($\boldsymbol{n = 5}$, $\boldsymbol{\alpha = 0.05}$)

Stack Depth	2	4	6	8	Average Testing Similarity
2		nS	S	S	0.987
4			S	S	0.985
6				S	0.974
8					0.717

Note: nS: not significant, S: significant

Several samples of testing results are given in the Table 5. some samples are unable to be translated correctly (see Table 6) although semantically the translated sentences are still closed to the input sentences.

Table 5. Some Predicted Results which are Similar with the Target Sentences

No	Sentences
1	(Id) Kita sedang berada dalam antrian ke tiga untuk take-off, dan diharapkan untuk mengudara dalam waktu kira-kira sepuluh menit. (Su) Pesawat waktos atos dina antrian katilu pikeun take-off, pesawat bakal hiber dina waktos sapuluh menit deui. (En) We are in line three for take-off, and are expected to be on the air in about ten minutes.
2	(Id) Kartu kredit adalah "uang plastik" yang dikeluarkan oleh bank untuk alat pembayaran di tempat-tempat tertentu seperti hotel, restoran, tempat rekreasi, dan lain-lain.

No	Sentences
	(Su) Kartu kredit nyaeta "duit plastik" anu dikaluarkeun ku bank pikeun alat pambayaran di tempat-tempat nu tangtu samisal jiga di hotel, restoran, tempat rekreasi jeung sajabana. (En) Credit cards are "plastic money" issued by banks for payment instruments at certain places such as hotels, restaurants, recreation areas, and others.
3	(Id) Bagi anda yang akan mengakhiri perjalanan di stasiun Purwokerto kami persilahkan untuk mempersiapkan diri. (Su) Pikeun anu perjalananana mung dugi ka stasiun Purwokerto supados siap-siap. (En) For those of you who will end the trip at Purwokerto station, we suggest you to prepare yourself.
4	(Id) Di Belanda pusat kotanya adalah Amsterdam, tetapi kota pemerintahannya adalah Den Haag. Di kota terakhir ini juga ada perwakilan dari luar negeri. (Su) Di nagri Walanda puseur dayeuhna nyaeta Amsterdam, tapi dayeuh pamarentahanana nyaeta Den Haag. Di dayeuh pamungkas ieu oge aya wawakil-wawakil ti nagara deungeun. (En) In the Netherlands the capital city is Amsterdam, but the city of government is the Hague. In this later city there are also representatives from overseas countries.
5	(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 train passengers, good evening and welcome to Bandung.

Table 6. Some Predicted Results which are not Closely Similar with the Target Sentences

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 gentlemen.

5. Conclusion

Based on the experiment results, it can be concluded as follows. *First*, transfer learning to build a neural machine translation for low-resource languages is a feasible approach as it does not require a large size of dataset to estimate many model parameters from scratch. In this aspect, fine-tuning a pre-trained vanilla transformer model (Vaswani et al., 2017) using a parallel corpus of low-resource language becomes a feasible way for developing a neural machine translation model to address a downstream task such as machine transation. *Second*, despite the encoder-decoder stack depth does not affect average training accuracy but it affects testing similarity of the vanila transformer model. The highest average testing similarity of the machine translation is achieved by the vanila transformer model with 2 and 4 stack depth; whilst the lowest testing similarity is achieved by the model with 8 stack depth. With a smaller number of model parameters than the transformer with more than 2 stack depth has made vanilla transformer model with 2 stack depth is potential to be explored further. *Third*, the vanilla transformer model still has a room for further optimization without compromising its performance. *Finally*, the experiment results showed that it is feasible to build a machine translation based on transformer model from Bahasa Indonesia to any Indonesian local language for socializing many

Governments' official announcements targeting to people in rural areas that speaks mostly local languages. These findings become a foundation to further study machine translation involving other Indonesian local languages.

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

Adil, M., Ullah, R., Noor, S., and Gohar, N., Effect of number of neurons and layers in an artificial neural network for generalized concrete mix design, *Neural Computing and Applications*, vol. 34, pp. 1–9, 2020.

Araabi, A., and Monz, C., Optimizing transformer for low-resource neural machine translation, *ArXiv Preprint ArXiv:2011.02266*, 2020.

Bahdanau, D., Cho, K., & 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*, no. 2, pp. 137–161, 2020.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. *40*. no. 4, pp. 834–848, 2017.

Dave, S., Parikh, J., and Bhattacharyya, P., Interlingua-based English--Hindi machine translation and language divergence, *Machine Translation*, vol. *16*, no. 4, pp. 251–304, 2001.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K., Bert: Pre-training of deep bidirectional transformers for language understanding, *ArXiv Preprint ArXiv:1810.04805,* 2018.

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, 2021.

Graves, A., Sequence transduction with recurrent neural networks, *ArXiv Preprint ArXiv:1211.3711*, 2021.

Graves, A., Wayne, G., and Danihelka, I., Neural turing machines. *ArXiv Preprint ArXiv:1410.5401*, 2014.

Grechishnikova, D., Transformer neural network for protein-specific de novo drug generation as a machine translation problem, *Scientific Reports*, vol. *11*, no. 1, pp. 1–13, 2021.

He, K., Zhang, X., Ren, S., and Sun, J., Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, 2016.

Hegde, A., Gashaw, I., and Hl, S., MUCS@-Machine Translation for Dravidian Languages using Stacked Long Short Term Memory, *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pp. 340–345, Kyiv, 2021.

Joshi, P., Barnes, C., Santy, S., Khanuja, S., Shah, S., Srinivasan, A., Bhattamishra, S., Sitaram, S., Choudhury, M., & Bali, K. (2019). Unsung challenges of building and deploying language technologies for low resource language communities. *ArXiv Preprint ArXiv:1912.03457*, 2019.

Kasai, J., Cross, J., Ghazvininejad, M., and Gu, J., Parallel machine translation with disentangled context transformer, *ArXiv Preprint ArXiv:2001.05136*, 2020.

Kellerman, A., International airports: Passengers in an environment of 'authorities.', *Mobilities*, vol. *3*, no. 1, pp. 161–178, 2008.

Kitaev, N., Kaiser, Ł., and Levskaya, A., Reformer: The efficient transformer, *ArXiv Preprint ArXiv:2001.04451*, 2020.

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, Shillong, India*, *170*, pp. 35, 2021.

Lewis, M. P., Simons, G. D., & Fennig, C. D., *Ethnologue: Languages of Asia (pp. 1--558), SIL International*. Global Publishing, 2014.

Liu, X., Duh, K., Liu, L., and Gao, J., Very deep transformers for neural machine translation, *ArXiv Preprint ArXiv:2008.07772*, 2020.

Luong, M.-T., Pham, H., and Manning, C. D., Effective approaches to attention-based neural machine translation,

ArXiv Preprint ArXiv:1508.04025, 2015.
Ma, S., Zhang, D., and Zhou, M., A simple and effective unified encoder for document-level machine translation, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3505–3511, 2020.
Minh, T. N., Meesad, P., and Nguyen Ha, H. C., English-Vietnamese Machine Translation Using Deep Learning, *International Conference on Computing and Information Technology*, pp. 99–107, 2021.
Narang, S., Chung, H. W., Tay, Y., Fedus, W., Fevry, T., Matena, M., Malkan, K., Fiedel, N., Shazeer, N., Lan, Z., and others, Do Transformer Modifications Transfer Across Implementations and Applications?, *ArXiv Preprint ArXiv:2102.11972*, 2021.
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, 2021.
Raganato, A., Tiedemann, J., and others., An analysis of encoder representations in transformer-based machine translation, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 287–297, Brussels, Belgium, 2018.
Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., and Kaur, R., Neural machine translation for low-resource languages: A survey, *ArXiv Preprint ArXiv:2106.15115*, 2021.
Rubino, R., Marie, B., Dabre, R., Fujita, A., Utiyama, M., and Sumita, E., Extremely low-resource neural machine translation for Asian languages, *Machine Translation*, vol. *34*, no. 4, pp. 347–382, 2021.
Shirsath, N., Velankar, A., Patil, R., & Shinde, S., Various Approaches of Machine Translation for Marathi to English Language. *ITM Web of Conferences*, vol. *40*, no. 4, pp. 3026, (2021).
Siddique, S., Ahmed, T., Talukder, M., Azam, R., Uddin, M., and others, English to bangla machine translation using recurrent neural network, *ArXiv Preprint ArXiv:2106.07225*, 2021.
Simonyan, K., & Zisserman, A., Very deep convolutional networks for large-scale image recognition, *ArXiv Preprint ArXiv:1409.1556*, 2014.
Susanti, D. I., Bahasa daearah Eksistensi Bahasa Daerah (Bahasa Jawa) di Bandara Adisutjipto, Yogyakarta, *Wacana: Jurnal Bahasa, Seni, Dan Pengajaran*, vol. *2*, no. 2, pp. 16–20*,* 2018.
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.
Wang, X., and Tong, Y., Application of an emotional classification model in e-commerce text based on an improved transformer model, *Plos One*, vol. *16*, no. 3, pp. e0247984, 2021.
Wang, Y., Shi, Y., Zhang, F., Wu, C., Chan, J., Yeh, C.-F., and Xiao, A., Transformer in action: a comparative study of transformer-based acoustic models for large scale speech recognition applications, *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6778–6782, Toronto, Canada, 2021.
Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., and others, Transformers: State-of-the-art natural language processing, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

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 Institure 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.