IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 August 2022, accepted 30 October 2022, date of publication 7 November 2022, date of current version 15 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3220691

RESEARCH ARTICLE

Residual Information Flow for Neural Machine Translation

SHEREEN A. MOHAMED [1], **MOHAMED A. ABDOU** [2], **AND ASHRAF A. ELSAYED** [1,3]

[1] Department of Mathematics and Computer Science, Faculty of Science, Alexandria University, Alexandria 21527, Egypt
[2] Informatics Research Institute, City for Scientific Research and Technology Applications, Alexandria 21527, Egypt
[3] Faculty of Computer Science and Engineering, Al Alamein International University, El-Alamein 21527, Egypt

Corresponding author: Shereen A. Mohamed (shereen.nafie@alexu.edu.eg)

ABSTRACT Automatic machine translation plays an important role in reducing language barriers between people speaking different languages. Deep neural networks (DNN) have attained major success in diverse research fields such as computer vision, information retrieval, language modelling, and recently machine translation. Neural sequence-to-sequence networks have accomplished noteworthy progress for machine translation. Inspired by the success achieved by residual connections in different applications, in this work, we introduce a novel NMT model that adopts residual connections to achieve better performing automatic translation. Evaluation of the proposed model has shown an improvement in translation accuracy by 0.3 BLEU compared to the original model, using an ensemble of 5 LSTMs. Regarding training time complexity, the proposed model saves about 33% of the time needed by the original model to train datasets of short sentences. Deeper neural networks of the proposed model have shown a good performance in dealing with the vanishing/exploding problems. All experiments have been performed over NVIDIA Tesla V100 32G Passive GPU and using the WMT14 English-German translation task.

INDEX TERMS Information flow, neural machine translation, neural sequence-to-sequence networks, residual connections, WMT14 English-German translation task.

I. INTRODUCTION

Alongside increasing globalization and exchange of information come persistent need for means of translation. There are between 6000 and 7000 natural languages around the world [1]. Human translation is slow and expensive, so, machine translation plays an important role in reducing language barriers and facilitating communication between people speaking different languages [2]. Machine translation is the subfield of computational linguistics that aims to study the translation of text and speech from a language to another by means of software [3], [4], [5].

In spite of the achievements made that indicated a near end to all the problems of machine translation, machine translation remains a big challenge. Different languages don't only have different vocabularies but also different styles and structures. So, the development of a machine translation system faces several challenges: [6].

1) Word meaning: various words have distinct meanings. Nevertheless, sometimes the same word may have different meanings. In this case, it's difficult to know which one of these meanings is intended, and consequently, generating the correct translation in the other language is difficult too.
2) Order of the word: there are different word orders. Some languages are subject to the order: subject (S), verb (V), object (O), others follow VSO, SOV, VOS, and other languages may have other word orders. It's important to pay attention to the order of the words when translating from a language to another in order to obtain an accurate translation.
3) Idioms: when gathering words, an expression is formed that has a completely different meaning to the meanings of the individual words. This should be taken into account, so as not to get translation in a different sense.

Due to these and other challenges, research in machine translation has been active for more than five decades.

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeeb Dey.

Since the introduction of Warren Weaver's memorandum in 1949, several machine translation techniques have been developed, including Example-based [7], Rule-based [8], Statistical Machine Translation (SMT) [5], [9], and recently, Neural Machine Translation (NMT) [10].

NMT has arisen as a new technology in the field of machine translation in the past few years. In spite of its short age, NMT has achieved tremendous popularity in this research area due to the encouraging translation results attained, as well as its simple structure [11]. NMT is the technique that uses a single deep neural network to translate a source language text into its target language counterpart [12]. It consists of one large end-to-end neural network containing two sub networks: the encoder and the decoder. First, the encoder gets the words of the source sentence, word by word, and turns them into a representation of semantic vector. Then, the decoder uses this representation to produce the output target sentence [13].

Despite the dominance of SMT technology in the field of automatic machine translation, NMT has shown several advantages in comparison with it [9] and [14]:

1) The SMT system is made up of various components that are adjusted individually. In contrast, The NMT model is one large end-to-end neural network that is responsible for both encoding the input sentences and generating the target sentences.
2) While the SMT system requires numerous precisely defined features to do the translation, relying only on a training corpus, and with little or no feature engineering effort, the NMT model can learn the same translation task.
3) Unlike SMT, NMT can capture significant long-distance dependencies and information of intricated word alignment.
4) The NMT model is different from the SMT. There is no need for a large memory space to store a reordering model, a translation model, and a language model.

Inspired by the idea of using residual connections to improve the performance of models in many tasks such as computer vision, image classification, image segmentation, and NMT, we propose a residual learning framework applied to the NMT model by Sutskever et al. [15]. Thus, the translation system will gain several advantages: 1) Residual connections boost information flow within the neural network and improve the training efficiency. 2) The forward propagation of information can improve the model accuracy, while the backward flow of gradient can speed up the convergence, and enhance the discrimination capability in classification problems. 3) They have been used in training very deep Convolutional Networks and achieved great breakthrough on many datasets, such as ImageNet and MS COCO.

The contributions in this work include:

1) The development of a new NMT model that provides higher translation accuracy and better performance without increasing the number of training parameters.
2) Applying the residual connections in the proposed model and verifying its efficiency in enhancing the translation performance by comparing the loss and accuracy curves with a state-of-the-art model.
3) Examining the extent to which the proposed model can deal with the vanishing gradient problem in deep models.
4) Evaluation of the original and the modified models on the WMT14 English-German dataset has shown that the modified model achieves better BLEU scores than the original model.

The novelty in this paper is how to use skip connections in the encoder layers of the proposed system to achieve: 1) Increased translation accuracy. 2) Reduced training time compared to the original model. 3) Better performance without increasing the number of training parameters, and consequently without increasing the computational resources. 4) Reducing the problem of vanishing gradient, which is clearly visible with increasing model depth.

The paper is organized into six sections starting with the introduction, followed by a discussion of state of the art work in using residual connections in different areas in Section II. In section III, we propose the sequence-to-sequence NMT model. Implementations and experimental setup are shown in Section IV. Results and comparisons are presented in Section V. Finally, Section VI, concludes the paper.

II. RELATED WORK

In 2016, He et al [16] presented a residual learning framework to enhance the training of deep neural networks. They applied residual connections between every few stacked layers. The proposed framework has been evaluated on the ImageNet [17] and CIFAR-10 [18] datasets. On the ImageNet, the deep network consisted of 152 layers, while on CIFAR-10 dataset, it consisted of 100 and 1000 layers. The results showed that the proposed framework offered less error rates with lower complexity. Although residual connections has been presented originally in image classification, they have proven highly efficient in many other tasks.

Guided by the results of the experiments they carried out, which showed that stacking multiple layers of Recurrent Neural Network (RNN) naively in the decoder would lead to slow training and degraded performance, Shu [19] studied the effect of using residual connections between the RNN layers of the decoder. The researchers used the same NMT architecture of Bahdanau et al. [20] and added another RNN to the decoder. The hidden states of the new RNN are calculated using the original decoder states. Now, instead of using the last RNN output, the Softmax layer uses a summation of the states of the two RNNs to compute its output. The authors declared that their proposed system is expected to abbreviate the back-propagation path of the model and consequently help the optimization. Evaluation of the proposed system has been performed on the ASPEC English-Japanese translation dataset [21]. The results showed an increase in BLEU score compared to the original model. Although the model achieves high translation accuracy, it consumes considerable time. Each time the decoder produces an output word, the model must go through the whole input sequence. This consumed time increases with the lengths of the source and target sequences.

Instead of using the last word yt-1, along with other information from the encoder, to produce the target word yt, Werlen et al. [22] introduced a decoder that relies on residual connection and uses target words from y0 to yt-1 to produce the target word yt. The proposed decoder is an attentive residual recurrent network. At each time step, the decoder makes a decision on which of the previously generated target word should be taken into consideration to predict the next one. The proposed system has been evaluated on three language pairs: English-German, English-Chinese, and Spanish-English, and showed an increase in BLEU scores in comparison to other models. Despite the improvement in translation accuracy achieved by the model, left-to-right decoding reduces the ability to control the appearance of a particular word or value in the target sentence.

III. THE PROPOSED SEQUENCE TO SEQUENCE NMT MODEL

Since residual connections have been improved the models' performance in many applications. Residual connection is a type of skip connection that overtakes the non-linear transformations with an identity mapping and obviously reconstructs the layers as learning residual functions with reference to the layer inputs. The residual connection can be explained by the following formula:

$$X_l = H_l\left(X_{l-1}\right) + X_{l-1} \tag{1}$$

where X_{l-1} and X_l are input and output to the l^{th} layer, and H_l is the residual function applied on the l^{th} layer.

In this section, we study the effect of applying residual connections on the sequence-to-sequence NMT model proposed by Sutskever et al [15]. Figures 1 and 2 show Sutskever model and the proposed model after adding residual connections from the embedding layer to the Softmax layer for a single LSTM layer encoder and a single LSTM layer decoder framework, respectively.

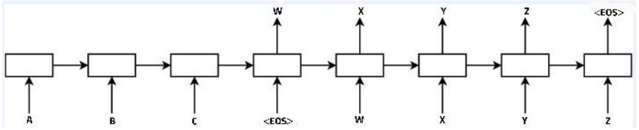


FIGURE 1. Sutskever model [15]. A, B, and C are the words of the input sentence. W, X, Y, and Z are the words of the generated target sentence.

Sutskever et al proposed a translation model that employed Long Short-Term Memory (LSTM) to translate English sentences into their French counterparts. The model consisted of an encoder and a decoder. First, the input sequence is passed through the encoder to convert it to a semantic vector of fixed length. Then, the decoder uses this vector to generate the target sequence. Evaluation of the system, using the WMT 14 English to French test set, showed an achievement of 34.8 BLEU and performance comparable to the phrase-based machine translation system.

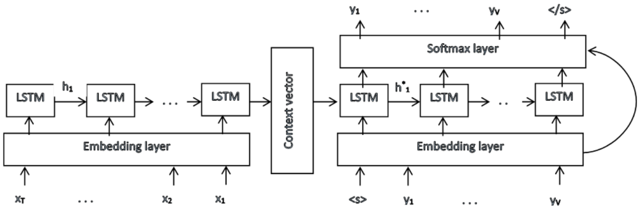


FIGURE 2. Proposed model.

Given a dataset of N sentence pairs, the proposed model is trained to learn a set of parameters θ that maximize the log-likelihood function:

$$\sum\nolimits_{i=1}^{N} \log P(Y^i|X^i, \theta) \tag{2}$$

For every pair of sentence (X, Y) ∈ N. Given an input sentence X=(x1, x2, ..., xT) and an output sentence Y=(y1, y2, ..., yV), the encoder starts to read the input sequence word by word, and generates a fixed dimensional context vector $\underset{c}{\longrightarrow}$ that is the summary of all information in the input sentence so that:

$$h_{[j]} = f\left(x_{[j]}, h_{[j-1]}\right) \tag{3}$$

$$\underset{c}{\longrightarrow} = q\left(\{h_1, \ldots, h_T\}\right) \tag{4}$$

where $h_{[j]}$ is the hidden state at time j, f is a non-linear function, and $q = h[T]$ is the last hidden state generated by the encoder.

Residual connections allow the propagation of features from lower layers to upper ones. This way, the upper layers not only refine the previous presentations, but also create new features, which consequently improves the learning performance.

The embedding layer generates distributed vector representations of words. These representations are important as they capture a large number of accurate syntactic and semantic word relationships.

In the proposed model, residual connections are added between the embedding layer and the Softmax layer to enhance the model performance by using embeddings of inputs to the decoder. Using the fixed dimensional context vector generated be the encode as an initial hidden state, the decoder starts generating the target sentence one word at a time, as shown in figure 2, using the formula:

$$y_i = g\left(y_{i-1}, h^*_{i-1}, v\right) \tag{5}$$

where g is a non-linear function, y_{i-1} is the output word at time i-1, h^*_{i-1} is the decoder hidden state at time i-1, and v is the output of the embedding layer.

IV. EXPERIMENTAL SETUP

This section begins by presenting the dataset preprocessing. The evaluation criteria is based on: the log-likelihood in equation (1), training time, and the BLEU Score Variation (BSV) introduced later in this section. Training LSTM networks require a lot of resources and training time; the LSTM cell is very complex with several gates added to its design. The training time increases with increased sentence length. In the following sections, we study this effect of variable sentence length on the improvement of both translation accuracy and training time.

A. DATASET PREPROCESSING

The WMT14 English-German dataset [23] that is available on the Stanford Natural Language Processing Group website, has been used in the training and evaluation of both the original Sutskever model and the proposed model. The dataset contains about 4.5M (4,468,840) sentence pairs, consisting of 694,766 unique English words and 1,531,652 unique German words. The maximum English sequence length is 100 tokens, and the maximum German sequence length is 100 tokens too.

To work properly, the dataset in [23] has been divided into five different datasets according to the sentence length: the first group consists of 313,966, the second of 1,651,851, the third of 2,900,904, the fourth of 3,686,003, and finally the fifth has 4,094,299 sentences respectively. This data classification is assumed to help study the effect of sentence length and vocabulary size on the performance of the baseline model and the proposed model. Table 1 summarizes the proposed dataset classification method.

TABLE 1. Proposed datasets used in training experiments.

Dataset	Number of sentences	The English vocabulary size	The German vocabulary size
Dataset10	313,966	73,535	129,524
Dataset20	1,651,851	285,484	592,964
Dataset30	2,900,904	449,339	981,641
Dataset40	3,686,003	554,365	1,230,211
Dataset50	4,094,299	616,500	1,373,220

The testing datasets available were used in the validation and the evaluation of the models. Newstest2012 and Newstest2014 sets were used for validation; Newstest2013 and Newstest2015 were used for testing. Table 2 shows the number of sentences.

In the incoming sections, two experiments will be shown: the first considers a single layer encoder/decoder model, while the second considers two layers encoder/decoder model3. The aim of this work is to study the effect of applying residual connections on translation accuracy and training time complexity. Also to examine how residual connections can deal with the vanishing and exploding problems.

TABLE 2. Test datasets used to evaluate the models.

Dataset	Number of sentences
TestDataset10	797
TestDataset20	2561
TestDataset30	3957
TestDataset40	4686
TestDataset50	4980

B. SINGLE LAYER ENCODER/DECODER MODEL SINGLE LAYER ENCODER/DECODER MODEL: TRAINING SETTINGS

An ensemble of 5 LSTMs has been proposed to evaluate both the original model and the proposed model, using almost the same training settings of Sutskever et al. The complete training criteria could be summarized as:

1) LSTM has been used in both the encoder and the decoder with 1000 cells at each layer, and word embeddings of 1000 dimensions.
2) All of LSTM parameters are initialized with the uniform distribution interval [-0.08, 0.08].
3) The stochastic gradient descent (SGD) optimizer has been used without momentum, and the initial learning rate is set to 0.7. After 5 epochs, the learning rate decreases by half every half epoch. Each model has been trained for 7 epochs.
4) The dataset is divided into batches of 128 sentences.
5) To avoid exploding gradients, a hard constrained has been set on the norm of the gradient. The gradient g is scaled to 5g/s when $s > 5$, where $s = ||g||_2$.
6) The input vocabulary consists of 160,000 words and the output vocabulary consists of 80,000 words. The other words have been replaced with "UNK" token.
7) The input sentences are entered in reversed word order.

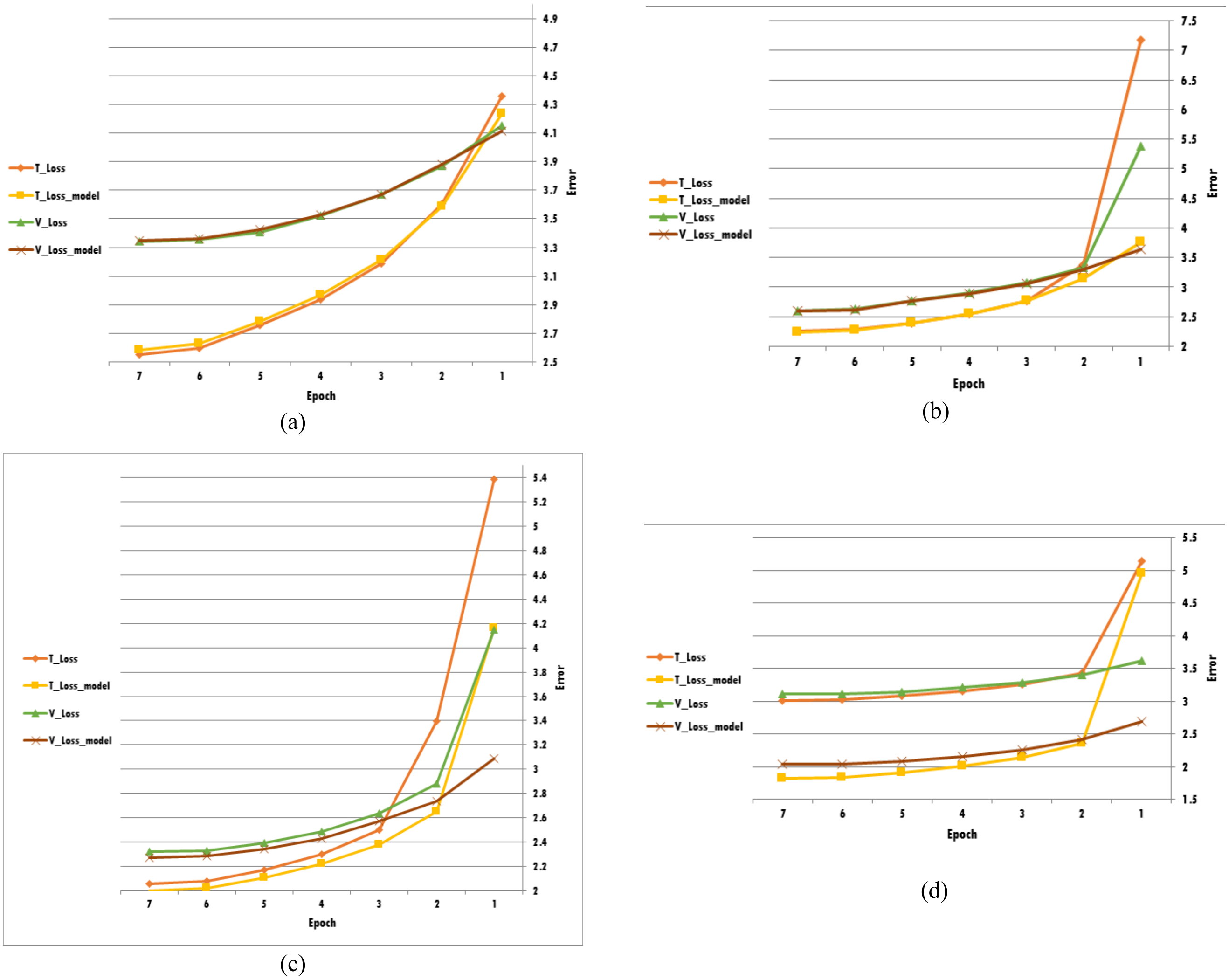
All experiments have been carried out on a single NVIDIA Tesla V100 32G Passive GPU.

SINGLE LAYER ENCODER/DECODER MODEL: RESULTS AND COMPARISONS

In this section, the proposed sequence-to-sequence NMT model is tested using the testing datasets, as mentioned in a previous section. To verify the efficiency of the proposed model, and prove how the residual connections has enhanced the performance, we present both the training loss and the training accuracy curves. Figure 3 shows the training loss curves for both training and development sets for all the datasets that have been used in the training. Figure (3-a) presents the training loss curves for both training and development sets of Dataset10, moreover figure (3-b) gives the training loss curves for both training and development sets of Dataset20, Figure (3-c) and (3-d) presents the training loss curves for both training and development sets of Dataset30 and Dataset40 respectively.

From figure (3), it could be concluded that the proposed residual connections criteria improved the conventional model to achieve efficient information flow. The loss curves of the proposed model all surpass those of the original model, for all datasets. Although the difference in the loss values between Sutskever and the proposed models begins with small fractions, as shown in figures 3.a and 3.b; the proposed model maintains consistent performance with increasing sentence length and increasing number of words that are not in the vocabularies as shown in Figure 3.d.

The same could be observed in the accuracy curves, shown in figure (4). Table 3 shows the BLEU scores obtained when



(a)

(b)

(c)

(d)

FIGURE 3. Training (T_Loss) and validation(V_Loss) error curves on: (a) Dataset10. (b) Dataset20. (c) Dataset30. (d) Dataset40.

TABLE 3. BLEU scores gained from an ensemble of 5 models with beam size = 1.

	testDataset10	testDataset20	testDataset30	testDataset40
Sutskever model	25.73688	21.50606	18.3438	16.669225
Proposed model	25.73896	21.65912	18.61096	16.68155
% Improvement	0.002	0.15	0.3	0.012

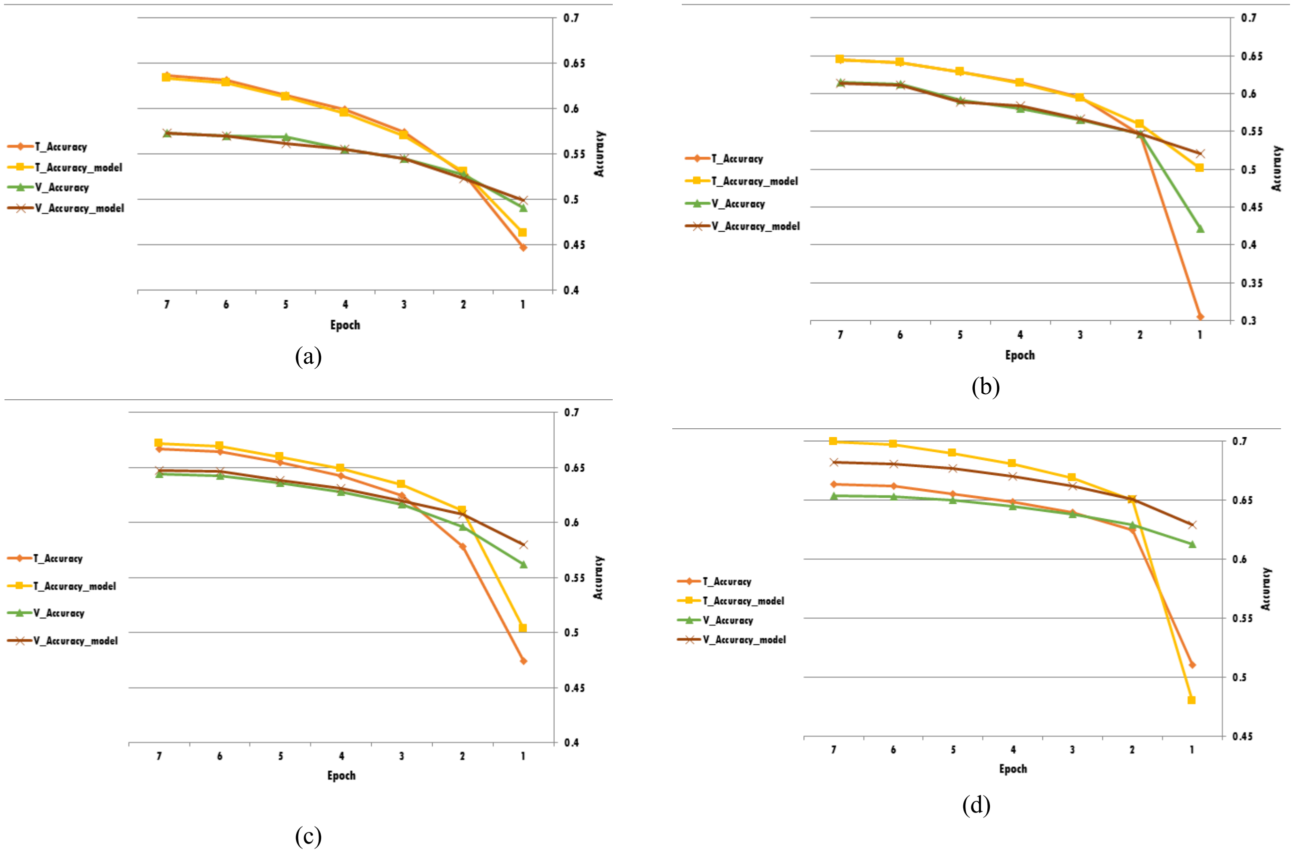
evaluating the proposed model and the original model, using 4 test datasets.

The proposed model outperforms the baseline state of the art work in all test datasets. With sentences no longer than 10 words, the proposed model gains .002% BLEU. With sentences no longer than 20 words, the proposed model outperforms the baseline by .15% BLEU. With sentences no longer than 30 words, the proposed model outperforms the baseline by .3% BLEU. Finally, with sentences no longer than 40 words, the difference in BLEU is 0.012%.

Regarding the training time, the proposed model shows better performance, than the original model, with short sequences. Table 4 shows the time (in hours) needed to train the group of 5 LSTMs in both the original and the proposed models, over the training datasets. As shown, the proposed model consumes less training time than the original model to train datasets of short sequences. For sequences of no more than 10 words, the proposed model consumes 69.2% of the training time the original model needs to train the same dataset. For sentences of no more than 20 words, the proposed model needs only 67.5% of the training time the original model consumes to train the same dataset. The outperformance of the proposed model in saving training time decreases as the sequence length increases.

C. TWO LAYERS ENCODER/DECODER MODEL

From the state of the art work, it could be concluded that while increasing the model's depth, the vanishing and exploding problems increase [16]. In this section, it is important to focus on the effect of increasing the model's depth on the translation



(a)

(b)

(c)

(d)

FIGURE 4. Training (T_Accuracy) and validation (V_ Accuracy) accuracy curves on: (a) Dataset10. (b) Dataset20. (c) Dataset30. (d) Dataset40.

TABLE 4. Training time (in hours) needed to train an ensemble of 5 models with beam size = 1.

Dataset	The original model (base)	The proposed model (prop)	(prop/base)%
Dataset10	5.763	3.989	69.2
Dataset20	45.518	30.716	67.5
Dataset30	106.594	106.448	99.9
Dataset40	169.176	170.291	100.7

accuracy. Several experiments are carried out where each model is trained twice.

1) TWO LAYERS ENCODER/DECODER MODEL: TRAINING SETTINGS

Firstly, the training procedure and hyperparameter values are chosen similar to those used in the original model. However, the results obtained were not satisfactory, as will be explained later. Thus, the following adjustments are to be considered:

1) The models have been trained using different learning rate values (beside 0.7): 0.3, 0.4, and 0.5.
2) The parameters of only the first LSTM have been initialized with the uniform distribution interval [−0.08, 0.08], in both the encoder and the decoder stages. Then, the parameters of all LSTMs have been initialized with the same interval.
3) Other uniform distribution intervals: [−0.06, 0.06] and [−0.07, 0.07] have been used.

All experiments have been carried out on a single NVIDIA Tesla V100 32G Passive GPU.

2) TWO LAYERS ENCODER/DECODER MODEL: RESULTS AND COMPARISONS

a: EXPERIMENT (1): THE EFFECT OF THE LEARNING RATE VALUE

Both the proposed NMT model and the original model have been tested over the testing datasets. Table 5 shows the BLEU scores obtained on testDataset10. The training has been carried out at different learning rate values, and with the parameters of only the first LSTM initialized in the interval [-0.08, 0.08], in both the encoder and the decoder stages. Here, we will define a new parameter: BLEU Score Variation (BSV), which represents the difference in BLEU score between the first and second run.

$$\text{BSV} = |\text{BLEU (First Run) - BLEU (Second Run)}| \quad (6)$$

It is preferred that the variation between obtained BLEU scores be minimum to maintain training system stability.

From table (5), the following observations could be pointed:-

The obtained BSV values, at different learning rates, show that the proposed model outperforms the original model in terms of stability. It is shown that the difference between first

TABLE 5. BLEU scores gained from the two layers encoder/decoder original and proposed models.

Learning rate	testDataset10					
	Sutskever			Proposed model		
	First Run	Second Run	BSV	First Run	Second Run	BSV
LR = 0.3	20.244	25.092	4.848	20.534	22.529	1.995
LR = 0.4	19.237	18.988	0.248	25.524	24.071	1.453
LR = 0.5	15.976	11.131	4.845	23.612	22.946	0.666
LR = 0.7	28.667	15.524	13.143	19.799	24.219	4.420

and second runs in case of the proposed model varies from 0.6659 to 4.4201; however in case of the original model the BSV varies from 0.2484 to 13.1431 which means that the system is too much dependent on the learning rate value and lacks stability. The proposed model shows more consistency than the original model.

At learning rate 0.7, the value of the BSV is maximum in case of the original model and reached 13.1431. Similarly, the proposed model expressed the highest BSV. The learning rate has been too large, resulting in an unstable training process and sub-optimal weights learning. Therefore, both the original model and the proposed models have been trained using other values of learning rates: 0.3, 0.4, and 0.5.

Table 6 shows the minimum BLUE scores obtained from the original and the proposed models, at different learning rate values. From the table, it can be concluded that the proposed model outperforms the original model at all learning rate values.

TABLE 6. Minimum BLEU scores obtained from the two layers encoder/decoder original and proposed models.

LEARNING RATE	SUTSKEVER	PROPOSED MODEL	IMPROVEMENT
LR = 0.3	20.2443	20.5341	0.2898
LR = 0.4	18.9883	24.0709	5.0826
LR = 0.5	11.1314	22.9461	11.8147
LR = 0.7	15.524	19.7992	4.2752

b: EXPERIMENT (2): THE EFFECT OF INITIALIZING ALL LSTMS PARAMETERS

The proposed and the original models have been re-trained with the parameters of all LSTMs initialized in the interval [−0.08, 0.08], in both the encoder and the decoder stages. Table 7 shows the BLEU scores gained on testDataset10. Initialization of the parameters of the second LSTM in the same interval as the first LSTM has not been useful in creating new features and therefore has not improved performance. The proposed model still shows more consistency than the original model. BSV values of the proposed model are less than their correspondings of the original model at all learning rates.

To further study the effect of initializing the parameters of the second LSTM on the model performance, we have repeated the training with initializing the second LSTM parameters in the interval [−0.07, 0.07], then in the interval [−0.06, 0.06]. Table 8 shows the BLEU scores received from the proposed model at all learning rates.

TABLE 7. BLEU scores obtained when initializing all LSTMs parameters in the interval [−0.08, 0.08].

Learning rate	Dataset10					
	Sutskever			Proposed model		
	First Run	Second Run	BSV	First Run	Second Run	BSV
LR = 0.3	19.0539	21.5693	2.5154	19.7336	19.3541	0.3795
LR = 0.4	8.7713	10.883	2.1117	17.6012	19.5956	1.9944
LR = 0.5	20.6032	6.3559	14.2473	20.1442	14.513	5.6312
LR = 0.7	19.3058	12.0424	7.2634	24.2504	21.5382	2.7122

TABLE 8. BLEU scores obtained when initializing the second LSTM parameters in other intervals.

Learning rate	Dataset10					
	[-0.06, 0.06]			[-0.07, 0.07]		
	First Run	Second Run	BSV	First Run	Second Run	BSV
LR = 0.3	20.8067	20.1097	0.697	21.1828	19.1539	2.0289
LR = 0.4	20.8309	20.6549	0.176	20.4996	20.5341	0.0345
LR = 0.5	25.0164	17.7841	7.2323	13.3363	14.8442	1.5079
LR = 0.7	19.9579	19.7405	0.2174	14.6096	15.769	1.1594

From table (8) it could be shown that with the interval [−0.06, 0.06], the proposed model is more consistent. The BSV values are less compared to those of table 7 except at learning rate 0.5. Also, the BLEU scores have been improved except at learning rate 0.7. With the interval [−0.07, 0.07], the proposed model shows more consistency too. The BSV values are less than those of table 7 except at learning rate 0.3.The BLEU scores have been improved at learning rates 0.3 and 0.4. The minimum BSV have been at learning rate of 0.4 for both intervals.

Although choosing different initialization intervals for each LSTM layer has improved the translation accuracy and model consistency as shown in Table 8, the translation accuracy and consistency are still lower than those shown in Table 5.

So, we conclude that the random initialization of the parameters of the second LSTM, in both the encoder and the decoder, helps the model to create new features, and consequently improves the translation accuracy. Also, the best learning rate to train the model on the WMT14 English-German translation task is 0.4.

V. CONCLUSION

In this work, a residual-connected framework of the sequence-to-sequence NMT model has been presented. The proposed model used residual connections between the embedding layer and the Softmax layer of the decoder. This concept of residual connections has been validated over a single NVIDIA Tesla V100 32G Passive GPU. For single layer encoder/decoder models, the experiments have shown that the proposed model performs better and gives higher

BLEU scores compared to the baseline model. Regarding training time complexity, training datasets of short sentences has been reduced from 45.5 hours to 30.7 hours compared to the state of the art work, which means a reduction of 67.5% in computational time complexity. In the two layers encoder/decoder models, the proposed model has shown better performance than the original model through the vanishing/exploding problems. The stability of the proposed model has been pointed through the values of BSV, where the values were reduced from 13.14 to 4.42 compared to the original model. Future work will focus on the efficiency of using residual connections framework on similar datasets. Furthermore, dense connected networks might be tackled for better performance.

ACKNOWLEDGMENT

The authors express our deepest gratitude to the IRI/Data Center Team at the City for Scientific Research and Technological Applications (SRTA-City) for providing GPU resources to train and evaluate the proposed models presented in this work.

REFERENCES

[1] D. Meulder, "Sign language communities," in *The Palgrave Handbook of Minority Languages and Communities*, G. Hogan-Brun and B. O'Rourke, Eds. Cham, Switzerland: Springer, 2018, pp. 207–232.

[2] W. Yang, "Research on the application of computer aided translation CAT in the field of translation," in *Cyber Security Intelligence and Analytics* (Advances in Intelligent Systems and Computing). Springer, 2019, pp. 148–154, doi: 10.1007/978-3-030-15235-2_23.

[3] J. Driscoll. (2018). *Localizing Webpages for Francophone Audiences With Machine Translation*. [Online]. Available: https://cs.union.edu/Archives/SeniorProjects/2018/CS.2018/files/driscolj/driscolj-499-report.pdf

[4] L. Bowker and J. B. Ciro, "Machine translation," in *Machine Translation and Global Research*. Bingley, U.K. Emerald Group Publishing, 2019, pp. 37–54.

[5] A. Garg and M. Agarwal, "Machine translation: A literature review," 2018, *arXiv:1901.01122*.

[6] M. Singh, R. Kumar, and I. Chana, "GA-based machine translation system for Sanskrit to Hindi language," in *Recent Trends in Communication, Computing, and Electronics*. Singapore: Springer, 2018, pp. 419–427, doi: 10.1007/978-981-13-2685-1_40.

[7] C. C. Chua, T. Y. Lim, L.-K. Soon, E. K. Tang, and B. Ranaivo-Malançon, "Meaning preservation in example-based machine translation with structural semantics," *Expert Syst. Appl.*, vol. 78, pp. 242–258, Jul. 2017, doi: 10.1016/j.eswa.2017.02.021.

[8] M. S. H. Ameur, F. Meziane, and A. Guessoum, "Arabic machine translation: A survey of the latest trends and challenges," *Comput. Sci. Rev.*, vol. 38, Nov. 2020, Art. no. 100305, doi: 10.1016/j.cosrev.2020.100305.

[9] D. Moussallem, M. Wauer, and A.-C.-N. Ngomo, "Machine translation using semantic web technologies: A survey," *J. Web Semantics*, vol. 51, pp. 1–19, Aug. 2018.

[10] S. A. Mohamed, A. A. Elsayed, Y. F. Hassan, and M. A. Abdou, "Neural machine translation: Past, present, and future," *Neural Comput. Appl.*, vol. 33, no. 23, pp. 15919–15931, Dec. 2021, doi: 10.1007/s00521-021-06268-0.

[11] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico, "Neural versus phrase-based machine translation quality: A case study," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 257–267. [Online]. Available: http://aclweb.org/anthology/D/D16/D16-1025.pdf

[12] Z. Yang, W. Chen, F. Wang, and B. Xu, "Generative adversarial training for neural machine translation," *Neurocomputing*, vol. 231, pp. 146–155, Dec. 2018.

[13] B. Zhang, D. Xiong, J. Su, and H. Duan, "A context-aware recurrent encoder for neural machine translation," *ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2424–2432, Dec. 2017.

[14] X. Wang, Z. Tu, and M. Zhang, "Incorporating statistical machine translation word knowledge into neural machine translation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 12, pp. 2255–2266, Dec. 2018.

[15] I. Sutskever, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[17] O. Russakovsky, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[18] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009. Accessed: Nov. 4, 2022. [Online]. Available: http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

[19] R. Shu, "Residual stacking of RNNs for neural machine translation," in *Proc. 3rd Workshop Asian Transl.*, 2016, pp. 223–229.

[20] D. Bahdanau, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[21] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara, "ASPEC: Asian scientific paper excerpt corpus," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, May 2016, pp. 2204–2208.

[22] L. M. Werlen, N. Pappas, D. Ram, and A. Popescu-Belis, "Self-attentive residual decoder for neural machine translation," 2017, *arXiv:1709.04849*.

[23] Nlp.Stanford.Edu. (2022). *The Stanford Natural Language Processing Group*. Accessed: Jul. 10, 2022. [Online]. Available: https://nlp.stanford.edu/projects/nmt/

SHEREEN A. MOHAMED received the B.Sc. degree in scientific computations from Ain Shams University, Cairo, Egypt, and the M.Sc. degree in computer science from Alexandria University, Alexandria, Egypt. She is currently a Researcher with the Department of Mathematics and Computer Science, Faculty of Science, Alexandria University. Her research interests include natural language processing, machine translation, deep learning, neural networks, and optimization.

MOHAMED A. ABDOU is currently a Professor of communications and computer engineering and the Vice Dean of the Informatics Research Institute, SRTA-City (www.srtacity.sci.eg). He is also the Academic Chancellor of Electrical Engineering Department, Pharos University in Alexandria (PUA), and a Visiting Professor with Alexandria University, as well. He spent more than 23 years in academia, in research and teaching at undergraduate and postgraduate levels. He has more than 50 publications in indexed journals and conferences. He is conducting his research in machine learning (ML) and deep learning (DL) applications in biomedical, healthcare, machine translation, and other industrial and engineering applications. He possesses a large experience with youth development and mentorship through several national and international projects for more than ten years as a consultant for entrepreneurship education and knowledge technology transfer in national and international levels.

ASHRAF A. ELSAYED received the B.Sc. and M.Sc. degrees in computer science from Alexandria University, Alexandria, Egypt, in 1995 and 2004, respectively, and the Ph.D. degree in computer science from the University of Liverpool, U.K., in 2012. He is currently an Associate Professor with the Faculty of Computer Science and Engineering, Al Alamein International University, Egypt (On leave as an Associate Professor at the Faculty of Science, Alexandria University). His research interests include data science, big data analytics, deep learning, quantum machine learning, and medical image mining.