International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.010

An Insight into Neural Machine Translation

SanjuktaGoswami
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Chennai,
India sg1990@srmist.edu.in

Dr. G Vadivu
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Chennai, India
vadivug@srmist.edu.in

AkellaSampath Sri Ram
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Chennai, India
as5907@srmist.edu.in

Dr. Evangelos E. Milios
Faculty of Computer Science Dalhousie University Halifax,
Canada
eem@cs.dal.ca

Dr. Jeba Sonia J
Department of Data Science and Business Systems, SRM Institute of Science and Technology
Chennai,
India jebas@srmist.edu.in

***Abstract* — Automatic language translation between two languages has encountered a quantum leap in perspective as of late in the field of machine learning. The term "neural machine translation" was developed in response to statistical machine translation, which relies on various count-based models and long dominated MT research. In contrast to conventional statistical machine translation, neural machine translation aims to build a single neural network that may be mutually changed to maximize translation efficiency. The current NMT models may be traced to previous versions of the encoder-decoder network family as well as to word and sentence embedding in this study. We will conclude with a succinct outline of recent developments in fields like NMT's bidirectional training (BiT).**

Keywords—Statistical Machine learning, Machine translation, word embeddings, Bi directional Training(BiT)

I. Introduction

One of the first objectives of text between languages was the automatic translation. The dawn of NMT positively checks one of the significant achievements throughout the entire existence of MT, and has prompted an extremist and unexpected departure of mainstream research from numerous past research lines. Given the fluidity of human language, machine or programmed translation may be among the most difficult AI undertakings. Earlier, rule-based frameworks were used for this task, but statistical techniques took their place in the 1990s. The field of neural machine translation, appropriately named, has more recently seen cutting-edge results from deep neural network models. The target language's feedforward neural language models were used to rank translation lattices in earlier attempts[1][2][3]. The principal neural models that also took into account the source language were established by using a similar model with bilingual tuples in place of specific linguistic words [4], directly scoring phrase pairs with a feedforward net [5,] or including a source defined range in the neural language model [6]. In this paper, we will discuss the origin of the NMT and try to give a basic overview of the concepts of NMT, Bi directionally training(BiT) it, and other current research in the field.

II. Word Embeddings

One of NLP's models most essential elements is the representation of words or phrases as continuous vectors. A d-dimensional real number vector should be used torepresent the word x. In general, a size d for the embeddinglayer that is noticeably smaller is chosen than the size of the vocabulary (d |Σ|). The following can be used to illustrate how a word is translated into its dispersed representation: a matrix of embedding called E ∈ R d×|Σ| [8]. The word x's d- dimensional representation is contained in the xth column of E,which is designated as E:

x. Embedded matrices are frequently learned along the network as a whole in NMT utilizing back - propagation algorithm [9] and a gradient-based optimizer. Many NLP subfields now make extensive use of pre- trained word embeddings made from unlabeled text[10]. The context in which a word commonly appears is usually taken into consideration by techniques for training word embeddings on raw text. [11][12], or enhance embeddings with cross-linguistic data[13][14]. Contextualized depictions make an effort to use the entire input sentence rather than just one word. In a number of NLP benchmarks, contextualized word embeddings have improved current technology. [15][16][17].

III. EmbeddingsWithin Phrases and Sentences

It is recommended to use phrases or sentences rather than single words when carrying out diverse NLP tasks. By utilising a scattered portrayal of the source sentence, for example, the distribution of the target sentences might be constrained. Reiterated autoencoders were a pioneering method for phrase embedding[18][19]. A phrase is designated as a d-dimensional vector by using [20] A word embedding matrix was initially trained by Socher et al. (2011). They then constructed an auto encoder network that uses the input as the fusion of two child representations to iteratively search for d-dimensional portrayals for 2D inputs. The word embedding chosen by the same auto encoder from two different guardians are the kid representations. A binary tree that can be created greedily controls the order in which representations merge. [20] or created with the aid of an Inversion Transduction Grammar[21][22]. However, in MT, the sentence representation must provide enough information to impose conditions on the objective sentence appropriation, and as a result, it must be higher dimensional than the word embeddings.

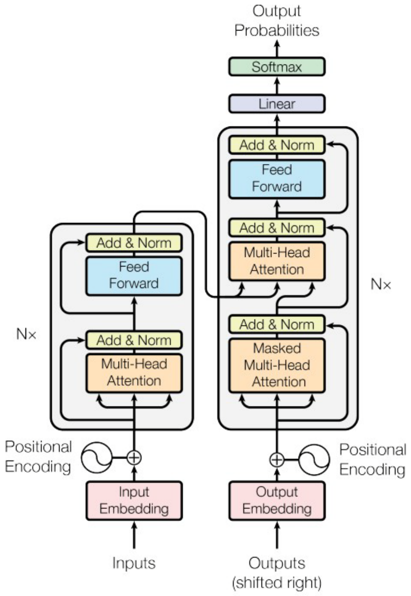
In order to get around the dimensionality issue with recurrent autoencoders, Kalchbrenner and Blunsom (2013)[23] found vector representations of words or sentences using convolution. Recent research finds sentence representations using self-attention much convenient rather than convolution. [24][25][26]. Yu et al. (2018) also looked at the possibility of using (recursive) connection networks. [28][29] which again totals the words in the sentence that are related to one another in pairs. Sentence representation frequently makes use of recurrent structures. It has been discovered that even untrained random RNNs can perform fairly well for a variety of NLP applications[30][31][32][33].

IV. Networks of Encoders and Decoders with Fixed Length Sentence Encodings

The primary authors who shaped the target sentence distribution using a distributed fixed-length representation of the source sentence were Kalchbrenner and Blunsom (2013)[23]. They modelled their recurrent continuous translation models (RCTM) I and II after the class of encoder-decoder systems [34], which is the most effective NMT design at the moment.

V. Architecture of an Attention Model

Sentences of various lengths pass varying amounts of information. Early NMT models had the drawback of commonly producing bad interpretations for lengthy words [35]. [36] Cho et al. (2014a) alluded that this error is due to the fixed-length source sentence encoding. A vector of constant length “does not have enough capacity to encode a long sentence with complicated structure and meaning”[36]



$\frac{1}{\sqrt{d_k}}$ e Transformer – Model Architecture

A. *Stacks of Encoders and Decoders*

Encoder: The encoder is made up of N = 6 discrete layers placed on top of one another. Each layer is composed oftwo sublayers. The first is a multiple-head self-attention mechanism, whereas the other is a standard feed-forward network thatis entirely related with positions. We employ a residual connection and then layer standardization to encircle each of the two sub-layers. For each sublayer, the outcome is Layer Norm(x + Sublayer(x)), where Sublayer(x) denotes the task performed by sublayer itself. Aspect model = 512 produces results for all model sub-layers and the embedding layers that can be used with the remaining associations[37].

Decoder: Similarly, the decoder is built from a stack of N= 6 similar layers. We employ lingering associations surrounding each sub-layer, similar to the encoder, followed by layer normalization. The self-consideration sub-layer in the decoder stack is also modified in orderto prevent positions from caring for resulting positions. The position's expectations due to this veiling and the fact that the result embedding are balanced by one position[37].

B. *Attention*

A set of vectors like key-value pairs, a planned inquiry, and a result can all be used to define an attention function. The weights assigned to each value are based on how well the question fits with its associated key, and the answer is generatedas a weighted sum of the values.

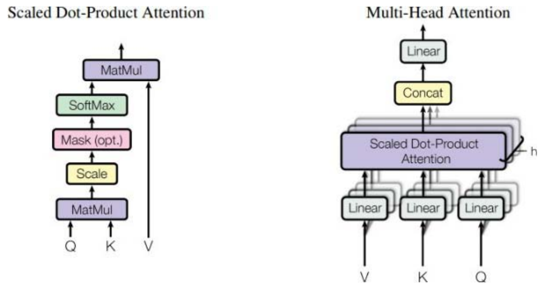


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi- Head Attention consists of several attention layers running in parallel.[37]

C. *Scaled Dot-Product Attention*

As shown in figure 2, the input for "Scaled Dot-Product Attention" consists of queries, keys of dimension dk, and values of dimension dv. The weights for the values are obtained by dividing each key by dk, the softmax function, and the query's product with each key.

We simultaneously register the attention function on many queries that are integrated into a Q matrix. Additionally, the keys as well as values are merged into the matrices K and V. The output matrix is processed as follows:

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Additive attention and dot-product (multiplicative) attention are the two commonly utilized attention functions. For greater values of dk, the dot product grows enormous in magnitude, driving the softmax function into areas where it has small gradients. To balance this impact,we scale the dot products by[37].

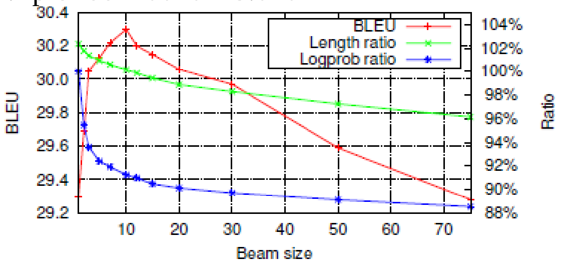


Figure 3: Transform model execution with varied beam widths on the English-German (WMT15) channel. At beam size 10, the BLEU score achieves its maximum, although the length proportion (Length of Hypothesis/Length of Reference) is less than 1. The proportion of the log-probabilities for greedy decoding is shown.

D. *Multi-Head Attention*

Instead of employing a single attention function with model-dimensional keys, values, and queries, we discovered that it is more effective to repeatedly direct project the keys, values, and queries to the dk and dv dimensions. The attention function is then applied simultaneously on every extended queries, keys, and values, producing dv-dimensional output values. Multi- head attention allows the model to process data from many portrayal subspaces simultaneously at various locations[37].

VI. Neural Machine Translation Decoding

Up to this point, we have seen how NMT defines thetranslation probability P(y|x).They do not explicitly explainhow to create a target sentence (y) from a given source sentence (x), despite this being the goal of machine translation. For two reasons, NMT decoding essentially expands significantly with sequence length, the search space first appears to be very large.

A. *Greedy And Beam Search*

To create the sequence outputs of tokens from a neural network model, greedy search and beam search are used. Both methods concentrate on models that go from sequence to sequence. Both algorithms operate simultaneously. A limited subset of the fractional hypotheses that have lengths (up to) j are chosen for extension in the following time step after being matched with one another in each iteration j. After a very large number of cycles have been completed, the algorithms stop, or all or the best of the selected hypotheses contain the finish-of-sentence symbol.

Beam search appears to be more exact, but there is no assurance that it will always lead to an interpretation with a higher or comparable score than greedy decoding. [38] According to Stahlberg and Byrne (2019), beam search has a huge number of searching errors.

VII. NMT Model Errors

In comparison to multi-level SMT systems like Hiero [39], which look at very broad search spaces, NMT beam search seems unnecessarily basic. This hypothesis claims that when the decoder fails to find the translation with the highest score, translation failures in NMT are more likely to be the consequence of search errors than model defects. It's interesting to note that this isn't always the case. Stahlberg et al. (2018), [38]Stahlberg and Byrne, [40]Niehues et al. (2017), [41][42]Stahlberg et al. (2018), and (2019). In particular, [38]Stahlberg and Byrne (2019) showed that the NMT decoding had a substantial number of search mistakes. Although, despite its theoretical advantage, NMT also experiences a variety of model mistakes in practice, as we will demonstrate in this section.

A. *Sentence Length*

Because the translated texts are becoming overly brief as a result of extensive beams, translation execution consistently declines (green curve). The log-probabilities of the found interpretations, as shown by the blue curve, are, nevertheless, declining as the beam size is increased. Anyhow, a big shaft for a bar search keeps the green path, thus it's seen as the more. This is encouraging right off the bat: outstanding translations may now be found with a rapid beam search and a small beam size. However, it indicates that providing search mistakes will fix the model fault of short translations with a thin beam is equivalent to retaliation. This means that any new NMT training approach will need to make adjustments to the beam size, which is another vital boundary.

VIII. Using Monolingual Training Data

The availability of data for concurrent MT training is often limited and costly, in contrast to the abundance of untranslated monolingual data. For instance, the translation grammar in Hiero [39] covers a large range of possible translations but fails miserably to assign points to them. Most of the time, it is the LM's responsibility to select a cohesive and fluid translation from that space. The NMT decoder should be integrated with an independently created RNNLM, according to Gulcehre et al. (2015, 2017). In a similar manner to traditional SMT, they also began combining the outcomes of RNN-LM and NMT using a log-linear model (a procedure known as "shallow fusion"). They demonstrated considerably better performance using "deep fusion," which makes use of a regulator network that gradually modifies the weights between RNN-LM and NMT. There have been a few increases in WMT assessment frameworks as a result of thorough integration and counting-based language models for n-best re-ranking [46][47]. The translation model is trained using the "simple fusion" method [48] to predictthe leftover probability of the training dataset mixed with the presumption of a fixed, pre-built LM.

Leftover probability of the training dataset mixed with the presumption of a fixed, pre-built LM. In the second research line, monolingual text is used to enrich data. By including monolingual data in the intended language, the natural concurrent training corpus will be enlarged. There are other ways to complete the source side of these sentences, including using a single false token[49] or replicating the intended sentence to the source side[50]. Reverse translation is the best method, and it makes use of a separate translation system to produce source sentences for sentences in a monolingual target language in the reverse direction. The performance of the final translation can, however, be significantly improved by improving the reverse system's quality if there are enough computational resources available. [52].

The amount of interleaving that must be stabilized with the amount of simulated data greatly restricts back-translation. [49][53][54]. So, the back-translation technique can utilize part of the readily available unilingual data. Over- sampling, which involves multiplying real training samples by the size of the synthetic data, can partially correct an imbalance between real and synthetic data. Anyhow, in practice, really high over-sampling rates generally don't perform well. In order The manufactured sentence pairs are used to produce a richer training signal, [55]Edunov et al. (2018a) has recommended adding noise in the sentences that were reverse-translated. Additionally, [56]Wang et al. have confirmed the efficacy of enhancing data in NMT with noise (2018b). These techniques broaden the training data set, which complicates model fitting and ultimately generates additional training signals. By selecting different sentences from the reverse translation model, one can also enhance the number of synthetic sentences in back-translation [57].

To accommodate for monolingual data, the third group of techniques alters the NMT training loss function. As an illustration, According to Escolano et al. (2018), the training goal should include auto encoder words that characterize how well a phrase can be translated into its original form and then reconstructed ([58]Cheng et al. (2016b), [59] Tu et al. (2017), and [60]). Additionally, (unsupervised) parallel learning techniques depend on the utilization of the reconstruction error ([61]He et al., 2016a; [62]Hassan et al., 2018;[63] Wang et al., 2018c). However, it is often expensive to compute and requires approximations to train for the new loss. Alternative methods for combining source-side [64] and target-sidemonolingual data include execute multi-task learning. Starting Seq2Seq training using already-trained encoder and decoder networks is another method for leveraging monolingual data in both the source and the destination languages ([66] Ramachandran et al., 2017; Skorokhodov et al., [67] (2018)). Unsupervised NMT is an extravagant sort of lever-aging monolingual training data since it eliminates the necessity for parallel training data [68][69].

IX NMT Training

Cross-entropy loss and backpropagation [70] are two typesof methods of optimizing like Ad delta 1 [71] are typically used to train NMT models. Recent NMT architectures such as fading gradients are one of the common training difficulties that are addressed by The Transformer, ConvS2S, or recurrent networks combining LSTM or GRU cells [72].

Research on training is still quite active. Now, early shallow models have been replaced with profound encoders and decoders with several layers. Deep architectures, particularly recurrent ones, are vulnerable to disappearing gradients[73], making training them more challenging because additional layers are required to transmit the gradients backwards. In the layer stack, residual connections [74] are quick associations that avoid more complicated sub-networks. Another method to prevent vanishing gradients is batch normalization [75], which uniformly sets each layer's hidden activations in tiny batches to have a mean of 0 and a variance of 1. Recurrent networks benefit most from layer normalization [76], a batch size-independent improvement to batch normalization.

A. *Regularization*

To aid in training, current NMT architectures are severely over-parameterized [77]. The model may be prone to over- fitting due to the huge number of features: The model perfectly matches the training data. Regularizers are techniques designed to stop neural networks that over fit and have too many parameters. The two manageable regularization techniques, according to one argument, are L1 and L2. It is meant to penalize the size of the weights in the network by include words in the loss function. Of course, these fines reduce a lot of variables to zero and make them irrelevant. Accordingly, the potential of the model is essentially constrained by L1 and L2. Label smoothing, early halting and dropout are the three regularization methods used most frequently for NMT. Dropout randomly resets the training exercises for both visible and concealed units to zero. It may be considered an effective regularizer in this way. Label smoothing significantly modifies the training objective, resulting in smoother distributions from the model.

B. *NMT by Bidirectional Training*

Bidirectional training is a quick and efficient pre-training method. The model will be bi-directionally updated at the earlier stage, and then tweaked as usual. The training samples can be reconstructed from "srctgt" to "src+tgttgt+src" to update bi-directionally without requiring any complex model adjustments. The suggested approach can be used in conjunction with current data manipulation techniques including back translation, data distillation, and data diversification. Large-scale investigations reveal that the methodology works as an innovative bilingual code-switcher, obtaining a better bilingual arrangement. Fortunately, with the help of BiT, our system [80] took first place in the low-resource track of IWSLT20218 for BLEU scores. Integrating BiT into our current systems [81][82] and confirming its efficacy in industrial level competitions will be intriguing.

C. *Results*

Outcomes on Multiple Data Scales: Various data sizes were(BiT) collected for 10 language directions, including IWSLT14 EnDe, WMT16 EnRo, IWSLT21 EnSw, WMT14 EnDe,and WMT19 EnDe, in order to test the method's utility. The largest direction has 38M sentence pairs, whereas the lowest direction only has 160K sentences. Table 1 displays the outcomes. The efficacy and thoroughness of the BiT are demonstrated by the fact that it significantly beats the solid standard Transformer in 7 of the 10 dimensions (importance test, p 0.01) and in remaining 3 directions (importance test, p 0.05) of the suggested bidirectional pre-training methodology. One advantage of BiT is that it reduces training time for the reverse direction by one- third. This benefit demonstrates that BiT may be a successful training method for multilingualism, such as multilingual pre training [83].

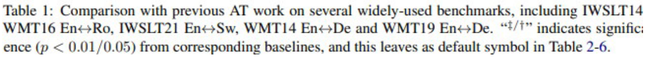
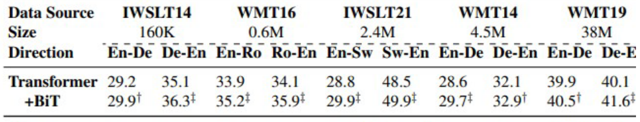


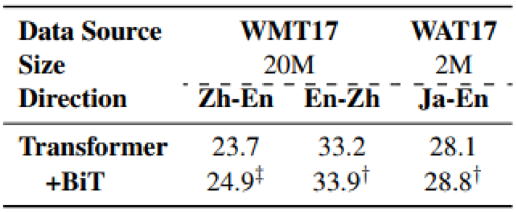
Table 1: Comparison with previous AT work on several widely-used benchmarks, including IWSLT14 WMT16 En↔Ro, IWSLT21 En↔Sw, WMT14 En↔De and WMT19 En↔De. "‡/†" indicates signific ence ($p < 0.01/0.05$) from corresponding baselines, and this leaves as default symbol in Table 2-6.



Data Source	IWSLT14		WMT16		IWSLT21		WMT14		WMT19	
Size	160K		0.6M		2.4M		4.5M		38M	
Direction	En-De	De-En	En-Ro	Ro-En	En-Sw	Sw-En	En-De	De-En	En-De	De-E
Transformer	29.2	35.1	33.9	34.1	28.8	48.5	28.6	32.1	39.9	40.1
+BiT	29.9†	36.3‡	35.2‡	35.9‡	29.9‡	49.9‡	29.7‡	32.9†	40.5†	41.6‡

Statistics for Distant Language Pairs: Inspired by [84], we present the BiT findings for the distant language pairs Zh'en and Ja'en, which are members of various language families. This clears up any confusion regarding the use of BiT and languages that belong to the same linguistic family, such as English and German.

Table 2: Performance on distant language pairs, including WMT17 Zh↔En and WAT17 Ja→En. To perform BiT on languages in different alphabets, we share the sub-words dictionaries between languages.



Data Source	WMT17		WAT17
Size	20M		2M
Direction	Zh-En	En-Zh	Ja-En
Transformer	23.7	33.2	28.1
+BiT	24.9‡	33.9†	28.8†

Table 2 shows the outcomes as they were observed, compared to baselines, and developed through time as a result oftechnique in all cases. BiT improves on average by+0.9 BLEU over the baselines.

X.Conclusion

The most popular and effective type of machine translation has been neural machine translation (NMT) over the years. In this study, word, phrase, and neural language models were used to reconstruct the history of NMT. We examined the repeat, convolution, and attention building blocks of NMT architectures. We then briefly discussed cutting-edge NMT research areas such NMT By Bidirectional Training

References

[1] Y. Bengio, R.Ducharme, P.Vincent, and C. Jauvin, “A neural probabilistic language model”, Journal of Machine Learning Research, pp. 1137–1155, 2003.

[2] Y.Bengio, H.Schwenk, J.S.Senécal, F.Morin, and J. L.Gauvain, “Neural Probabilistic Language Models”, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 137–186, 2006.

[3] H.Schwenk, D.Dechelotte, and J.L.Gauvain, “Continuous space language models for statistical machine translation,” In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, Australia. Association for Computational Linguistics, pp. 723–730, 2006.

[4] F.Zamora-Martinez, M. J.Castro-Bleda, and H. Schwenk, “N- gram-based machine translation enhanced with neural networks for the French-English BTEC-IWSLT’10 task”, In International Workshop on Spoken Language Translation (IWSLT) 2010.

[5] H. Schwenk, “Continuous space translation models for phrase- based statistical machine translation,” In Proceedings of COLING 2012: Posters, Mumbai, India. The COLING 2012 Organizing Committee, pp. 1071–1080, 2012.

[6] H.S.Le, A.Allauzen, and F. Yvon, “Continuous space translation models with neural networks,” In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montréal, Canada. Association for Computational Linguistics, pp. 39– 48, 2012.

[7] J.Devlin, R.Zbib, Z.Huang, T.Lamar, R.Schwartz, and J.Makhoul, “Fast and robust neural network joint models for statistical machine translation”, In proceedings of the 52nd annual meeting of the Association for Computational Linguistics,vol. 1, Long Papers, pp. 1370-1380, June, 2014.

[8] R.Collobert, and J. Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” In Proceedings of the 25th International Conference on Machine Learning, pp. 160-167, July, 2008.

[9] D. E.Rumelhart, J. L.McClelland, and PDP Research Group,“Parallel distributed processing”, New York: IEEE, vol. 1, pp. 354-362, 1988.

[10] R.Collobert, J.Weston, L.Bottou, M.Karlen, K.Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” J. Mach. Learn. Res., vol. 12, pp. 2493–2537, 2011.

[11] J.Pennington, R.Socher, and C. D. Manning, “Glove: Global vectors for word representation”, In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) pp. 1532-1543, October, 2014.

[12] T.Mikolov, Q. V.Le, and I. Sutskever, “Exploiting similarities among languages for machine translation,”arXiv preprint arXiv:1309.4168, 2013.

[13] T.Mikolov, K.Chen, G.Corrado, and J. Dean, “Efficient estimation of word representations in vector space”,arXiv preprint arXiv:1301.3781, 2013.

[14] S.Upadhyay, M.Faruqui, C.Dyer, and D. Roth, “Cross-lingual models of word embeddings: An empirical comparison,”arXiv preprint arXiv:1604.00425, 2016.

[15] M. E.Peters, M.Neumann, L.Zettlemoyer, and W. T. Yih, “Dissecting contextual word embeddings: Architecture and representation,”arXiv preprint arXiv:1808.08949, 2018.

[16] J.Phang, T.Févry, and S. R. Bowman, “Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks,”arXiv preprint arXiv:1811.01088, 2018.

[17] J.Devlin, M. W.Chang, K.Lee, and K. Toutanova, “Bert: Pre- training of deep bidirectional transformers for language understanding,”arXiv preprint arXiv:1810.04805, 2018.

[18] J. R. Bellegarda, “A latent semantic analysis framework for large-span language modeling,” In Fifth European Conference on Speech Communication and Technology, 1997.

[19] Y.Bengio, R.Ducharme, P.Vincent, and C. Jauvin, “A neural probabilistic language model,” Journal of Machine Learning Research 3, Feb (2003), pp. 1137—1155, 2003, Google Scholar Google Scholar Digital Library Digital Library.

[20] R.Socher, J.Pennington, E. H.Huang, A. Y.Ng, and C. A. Manning, “Semi-supervised recursive autoencoders for predicting sentiment distributions,” In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 151-161, July2011.

[21] D. Wu, “Stochastic inversion transduction grammars and bilingual parsing of parallel corpora,” Computational Linguistics, vol. 23, no. 3, pp. 377-403, 1997.

[22] P.Li, Y.Liu,and M. Sun, “Recursive autoencoders for ITG-based translation,” In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing pp. 567-577, October2013.

[23] N.Kalchbrenner, and P. Blunsom, “Recurrent continuous translation models,” In Proceedings of the 2013 conference on empirical methods in natural language processing pp. 1700-1709, October, 2013.

[24] G.Wang, C.Li, W.Wang, Y.Zhang, D.Shen, X.Zhang, ... and L. Carin, “Joint embedding of words and labels for text classification,”arXiv preprint arXiv:1805.04174, 2018.

[25] L.Wu, F.Tian, T.Qin, J.Lai, and T. Y. Liu, “A study of reinforcement learning for neural machine translation,”arXiv preprint arXiv:1808.08866, 2018.

[26] Q.Zhang, S.Liang,and E. Yilmaz, “Variational self-attention model for sentence representation,”arXiv preprint arXiv:1812.11559, 2018.

[27] L.Yu, C. D. M.d'Autume, C.Dyer, P.Blunsom, L.Kong, andW. Ling, “Sentence encoding with tree-constrained relation networks”,arXiv preprint arXiv:1811.10475, 2018.

[28] A.Santoro, D.Raposo, D. G.Barrett, M.Malinowski, R.Pascanu, P.Battaglia,and T.Lillicrap,“A simple neural network module for relational reasoning,” Advances in neural information processing systems, p. 30, 2017.

[29] R.Palm, U.Paquet,and O. Winther, “Recurrent relational networks,” Advances in Neural Information Processing Systems, p. 31, 2018.

[30] A.Conneau, D.Kiela, H.Schwenk, L.Barrault, and A. Bordes, “Supervised learning of universal sentence representations from natural language inference data,”arXiv preprint arXiv:1705.02364, 2017.

[31] A.Conneau, G.Kruszewski, G.Lample, L.Barrault,and M.Baroni,“What you can cram into a single vector: Probing sentence embeddings for linguistic properties”,arXiv preprint arXiv:1805.01070, 2018.

[32] A.Conneau, G.Lample, M. A.Ranzato, L.Denoyer,and H. Jégou, “Word translation without parallel data,”arXiv preprint arXiv:1710.04087, 2017.

[33] J.Wieting, and D. Kiela, “No training required: Exploring random encoders for sentence classification,”arXiv preprint arXiv:1901.10444, 2019.

[34] R. P.Neco,and M. L. Forcada, “Asynchronous translations with recurrent neural nets. In Proceedings of International Conference on Neural Networks (ICNN'97), IEEE, vol. 4, pp. 2535-2540, June1997.

[35] P.Sountsov,and S. Sarawagi, “Length bias in encoder decoder models and a case for global conditioning”,arXiv preprint arXiv:1606.03402, 2016.

[36] K.Cho, B.Van Merriënboer, D.Bahdanau,and Y.Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,”arXiv preprint arXiv:1409.1259, 2014.

[37] A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A. N.Gomez, ... and I. Polosukhin, “Attention is all you need”, Advances in Neural Information Processing Systems, p. 30, 2017.

[38] F.Stahlberg, D.Saunders, A.de Gispert,and B.Byrne, CUED@WMT19:EWCandLMs. In Proceedings of the Fourth Conference on Machine Translation: Shared Task Papers. Association for Computation, 2019.

[39] D. Chiang, “Hierarchical phrase-based translation,”Computational Linguistics, vol. 33, no. 2, pp. 201-228, 2007.

[40] J.Niehues, E.Cho, T. L.Ha,and A.Waibel,“Analyzing neural MT search and model performance,’. arXiv preprint arXiv:1708.00563, 2017.

[41] F.Stahlberg, J.Cross,and V. Stoyanov, “Simple fusion: Return of the language model,” In Proceedings of the Third Conference on Machine Translation: Research Papers, Belgium, Brussels. Association for Computational Linguistics, pp. 204–211, 2018a.

[42] F.Stahlberg, A.de Gispert,and B. Byrne, The University of Cambridge’s machine translation systems for WMT18. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Belgium, Brussels. Association for Computational Linguistics, pp. 504–512, 2018b.

[43] C.Gulcehre, O.Firat, K.Xu, K.Cho, L.Barrault, H.C.Lin, F.Bougares, H.Schwenk,and Y. Bengio, “On using monolingual corpora in neural machine translation,”arXiv preprint arXiv:1503.03535, 2015.

[44] C.Gulcehre, O.Firat, K.Xu, K.Cho,and Y. Bengio, “On integrating a language model into neural machine translation”, Computer Speech and Language, vol. 45, pp. 137 – 148, 2017.

[45] P. Koehn, “Pharaoh: A beam search decoder for phrase-based statistical machine translation models,” In Frederking, R. E., and Taylor,K. B. (Eds.), Machine Translation: From Real Users to Research, Berlin, Heidelberg. Springer Berlin Heidelberg, pp. 115–124, 2004.

[46] S.Jean, O.Firat, K.Cho, R.Memisevic,and Y. Bengio, “Montreal neural machine translation systems for WMT’15,” In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal. Association for Computational Linguistics, pp. 134–140, 2015b.

[47] X.Wang, Z.Lu, Z.Tu, H.Li, D.Xiong,and M. Zhang, “Neural machine translation advised by statistical machine translation”, In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI’17, AAAI Press, pp. 3330–3336, 2017.

[48] F.Stahlberg, J.Cross,and V. Stoyanov, “Simple fusion: Return of the language model”, In Proceedings of the Third Conference on Machine Translation: Research Papers, Belgium, Brussels. Association for Computational Linguistics, pp. 204–211, 2018a.

[49] R.Sennrich, B.Haddow,and A. Birch, “Improving neural machine translation models with monolingual data,” In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany. Association for Computational Linguistics, pp. 86–96, 2016b.

[50] A.Currey, A. V.MiceliBarone,and K. Heafield, “Copied monolingual data improves low-resource neural machine translation,” In Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark. Association for Computational Linguistics, pp. 148–156, 2017.

[51] H. Schwenk, “Investigations on large-scale lightly-supervised training for statistical machine translation,” In International Workshop on Spoken Language Translation (IWSLT), pp. 182–189, 2008.

[52] F.Burlot,and F. Yvon,“Using monolingual data in neural machine translation: A systematic study,” In Proceedings of the Third Conference on Machine Translation: Research Papers, Belgium, Brussels. Association for Computational Linguistics, pp. 144–155, 2018.

[53] R.Sennrich, B.Haddow,and A. Birch, “Edinburgh neural machine translation systems for WMT 16”, In Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics, pp. 371–376, 2016a.

[54] A.Poncelas, D.Shterionov, A.Way, G. M. d. B.Wenniger,and P. Passban, “Investigating back translation in neural machine translation,”arXiv preprint arXiv:1804.06189, 2018.

[55] S.Edunov, M.Ott, M.Auli,and D. Grangier, “Understanding back-translation at scale,” In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Association for Computational Linguistics,pp. 489–500, 2018a.

[56] X.Wang, H.Pham, Z.Dai,and G. Neubig, “SwitchOut: An efficient data augmentation algorithm for neural machine translation,” In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Association for Computational Linguistics, pp. 856–861, 2018b.

[57] K.Imamura, A.Fujita,and E.Sumita,“Enhancement of encoder and attention using target monolingual corpora in neural machine translation,” In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, Melbourne, Australia. Association for Computational Linguistics, pp. 55–63, 2018.

[58] Y.Cheng, W.Xu, Z.He, W.He, H.Wu, M.Sun,and Y. Liu, “Semi-supervised learning for neural machine translation,” In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany. Association for Computational Linguistics, pp. 1965–1974, 2016b.

[59] Z.Tu, Y.Liu, L.Shang, X.Liu,and H. Li, “Neural machine translation with reconstruction, In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI’17, AAAI Press, pp. 3097– 3103, (2017).

[60] C.Escolano, M. R.Costa-jussà,and J. A. Fonollosa, (self- attentive) autoencoder based universal language representation for machine translation. arXiv preprint arXiv:1810.06351, 2018.

[61] D.He, Y.Xia, T.Qin, L.Wang, N.Yu, T.Y.Liu, and W.Y.Ma,“Dual learning for machine translation,” In Lee, D. D., Sugiyama, M.,

Luxburg, U. V., Guyon, I., and Garnett, R. (Eds.), Advances in Neural Information Processing Systems 29, Curran Associates, Inc., pp. 820–828, 2016a.

[62] H.Hassan, A.Aue,C. Chen, V.Chowdhary, J. H. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W.D. Lewis, M. Li, et al., “Achieving human parity on automatic Chinese to English news translation,”arXiv preprint arXiv:1803.05567, 2018.

[63] Y. Wang, Y. Xia, L. Zhao, J.Bian, T. Qin, G. Liu, andT. Y. Liu, “Dual transfer learning for neural machine translation with marginal distribution regularization,” In Thirty-Second AAAI Conference on Artificial Intelligence, 2018c.

[64] J. Zhang, andC. Zong, “Exploiting source-side monolingual data in neural machine translation,” In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas. Association for Computational Linguistics, pp. 1535–1545,2016b.

[65] T. Domhan, andF. Hieber, “Using target-side monolingual data for neural machine translation through multi-task learning,” In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark. Association for Computational Linguistics, pp. 1500–1505,2017.

[66] P. Ramachandran, P. J. Liu, and Q. V. Le, “Unsupervised pretraining for sequence to sequence learning,” In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark. Association for Computational Linguistics, pp. 383–391, 2017.

[67] I. Skorokhodov, A. Rykachevskiy, D. Emelyanenko, S.Slotin, andA. Ponkratov, “Semi-supervised neural machine translation with language models,” In Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018), Boston, MA. Association for Machine Translation in the Americas, pp. 37–44,2018.

[68] Rajesh, M., &Sitharthan, R. (2022). Image fusion and enhancement based on energy of the pixel using Deep Convolutional Neural Network. Multimedia Tools and Applications, 81(1), 873-885.

[69] M. Artetxe, G. Labaka, E. Agirre, andK. Cho, Unsupervised neural machine translation. arXiv preprint arXiv:1710.11041, 2017b.

[70] D.E. Rumelhart, G. E. Hinton, andR.J. Williams, R. J.,“Neurocomputing: Foundations of Research, chap,” Learning Representations by Back-propagating Errors, MIT Press, Cambridge, MA, USA, pp. 696–699.1988.

[71] M.D. Zeiler, ADADELTA: An adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012.

[72] S. Hochreiter, Y.Bengio, P. Frasconi, andJ. Schmidhuber, “Gradient flow in recurrent nets: The difficulty of learning long-term dependencies,” 2001.

[73] R.Pascanu, T.Mikolov,and Y.Bengio,“On the difficulty of training recurrent neural networks,” In Dasgupta, S., andMcAllester, D. (Eds.), Proceedings of the 30th International Conference on MachineLearning, Vol. 28 of Proceedings of Machine Learning Research, Atlanta, Georgia, USA. PMLR, pp. 1310–1318, 2013.

[74] K.He, X.Zhang, S.Ren,and J. Sun, “Deep residual learning for image recognition,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016b.

[75] Moshika, A., Thirumaran, M., Natarajan, B., Andal, K., Sambasivam, G., &Manoharan, R. (2021). Vulnerability assessment in heterogeneous web environment using probabilistic arithmetic automata. IEEE Access, 9, 74659-74673.

[76] J. L.Ba, J. R.Kiros,and G. E. Hinton, Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

[77] F. Stahlberg, andB. Byrne, “Unfolding and shrinking neural machine translation ensembles,” In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark. Association for Computational Linguistics, pp. 1946–1956,2017.

[78] R. Livni, S. Shalev-Shwartz, andO. Shamir, “On the computational efficiency of training neural networks,” In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (Eds.), Advances in Neural Information Processing Systems 27, Curran Associates, Inc., pp. 855–863, 2014.

[79] L.Ding, D.Wu,and D. Tao, “Improving Neural Machine Translation by Bidirectional Training,”arXiv preprint arXiv:2109.07780, 2021.

[80] Liang Ding, Di Wu, and Dacheng Tao, The usyd-jd speech translation system for iwslt2021. In IWSLT, 2021d.

[81] Liang Ding and Dacheng Tao, The University of Sydney's machine translation system for wmt19. In WMT, 2019.

[82] Longyue Wang, ZhaopengTu, Xing Wang, Li Ding,Liang Ding, and Shuming Shi,Tencentai lab machine translation systems for wmt20 chat translation task. In WMT, 2020.

[83] Yinhan Liu, JiataoGu, NamanGoyal, Xian Li, Sergey Edunov, MarjanGhazvininejad, Mike Lewis, and Luke Zettlemoyer,“Multilingual denoising pre-training for neural machine translation,” Transactions of the Association for Computational Linguistics, 2020b.

[84] Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and ZhaopengTu,“Rejuvenating low-frequency words: Making the most of parallel data in non-autoregressive translation,” In ACL, 2021b.

[85] F. Stahlberg, “Neural machine translation: A review,” Journal of Artificial Intelligence Research, vol. 69, pp. 343-418, 2020.