CS447 Literature Review: Can convolutional networks replace recurrent networks (RNN, LSTM, GRU) in encoder-decoder architecture for Neural Machine Translation?

Pallaw Kumar,
pallawk2@illinois.edu

December 10, 2022

1 Introduction

Natural Language Processing (a.k.a., NLP) is a sub-field of Artificial Intelligence that has close links with Machine Learning. In NLP, the research areas in focus are Machine Translation (MT), Semantic Parsing, Sentiment Analysis, Topic Modeling, Named Entity Recognition (NER), Text Analytics, etc. Deep Learning techniques with Natural Language Processing, such as Neural Machine Translation (NMT) find a lot of applications in domains such as Healthcare, Finance, Manufacturing, Education, Retail, and customer service Kalyanathaya et al. (2019). NMT is a new highly active approach for machine translation, which has shown promising results and due to its success, it has attracted many researchers in the field. Language processing-related impacts research is driving fuel to innovate many tools to build industrial applications.

In this paper, the question of interest is related to neural machine translation. In its vanilla form, a neural machine translation is sequence-to-sequence modeling using an encoder-decoder architecture where an input sentence is first encoded by the encoder, later using this information, the decoder translates to another form, to achieve translation between language pairs such as English-German and vice-versa. In existing research works the recurrent networks such as Recurrent Neural Networks (RNNs) & Long-Short Time Memory (LSTM) are extensively used for building encoder-decoder foundational structures but, recent advancements Meng et al. (2015), Su et al. (2016), Gehring et al. (2017a), Gehring et al. (2017b), Singh et al. (2017), Wang and Xu (2017), Watanabe et al. (2017) are shown using convolutional networks based structures in neural machine translation tasks given better results not only in terms of accuracy but also reducing learning and decoding time drastically. Thereby, such developments raise a pertinent question, ”Can convolutional networks (CNNs) replace recurrent networks (RNN, LSTM, Gated recurrent units (GRU)) in encoder-decoder structures for Neural Machine Translation tasks?”

In this paper, we are going to discuss the above seven paper methodologies, findings, and relevance in detail to understand why it seems promising to query more on the research question of interest. The paper is organized in the following sequences - Section 2 explains the motivation factors behind asking the research question of interest, Section 3 briefs existing neural networks architectures, Section 4 explains the motivation behind the included paper, Section 5 summarizes the above seven research papers, Section 6 provides a work assessment and Section 7 highlights their links with the research question if any, and finally, Section 8 concludes the findings.

2 Motivation

Convolution neural networks are extensively used network types for computer vision problems because of its convolution properties that emphasize more on local features seen in visual data. Using CNNs-based architectures is seen to have better processing speed due to its simplicity to work with hardware such as CUDAS - found to be time efficient to work on big data as this architecture is designed to support parallel processing. Architectures such as RNN, LSTM, and GRU are useful for working with sequence-to-sequence models to solve natural language translation however, it does not support parallel processing. As the result, the existing neural machine translator takes longer training time as they have recurrent networks based encoder-decoder combination. Studying papers discussed in section 4 indicates existing NMT architectures are not fully utilized on large datasets due to their slow processing power. It is hopeful that convolutional networks-based encoder-decoder architectures will achieve faster machine translation. Thus, the motivation is to study if all existing RNNs based architectures can be mechanized using only CNNs would be a breakthrough advancement in neural machine translation.

3 Background

3.1 Convolution Neural Network (CNN)

Convolution neural network (CNN) is the most popular neural network architecture that has gained the most success for computer vision tasks and predictive classification of visual data across various domains ranging from finance, to radiology Yamashita et al. (2018). CNNs are the most utilized deep learning network type and are consumed as foundational components in building the most powerful architectures such as the AlexNet network and High-Resolution network (HR.Net) Alzubaidi et al. (2021). CNNs are fused with the latest transformers to work with challenging vision datasets with considerably less computational cost Guo et al. (2022). A typical CNN architecture is a combination of multi-layers convolution layers followed by max-pooling layers (for down-sampling, optional) and non-linear units such as GLU or ReLU with a flat layer before the output layer. Fig 1. shows the ImageNet classification architecture built using convolutional neural networks.

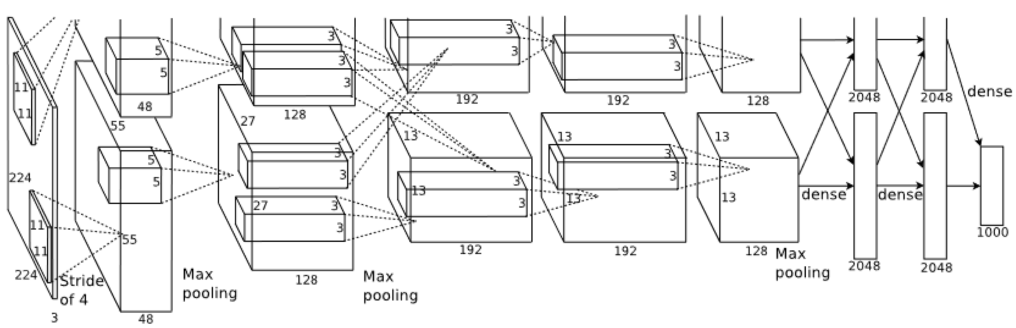

Figure 1: Source: Krizhevsky et al. (2012): ImageNet Classification Architecture - an example of CNN architecure.

3.2 Recurrent Neural Network (RNN)

Recurrent neural networks (RNNs) are a class of neural networks that allow previous outputs to be used as inputs while having hidden states. RNNs are a vanilla form of the recurrent network family (RNN, LSTM, GRU). This family of networks is extensively used for natural language processing. Fig. 2 shows a typical RNN network with inputs X_k, outputs y_k, and hidden units h_k where suffix k represents time information.

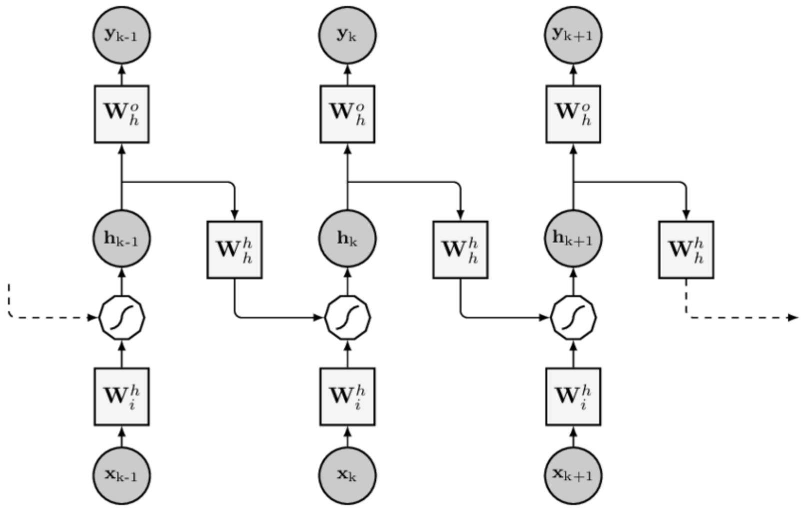

Figure 2: Source: Bianchi et al. (2017): A typical RNN architecture with sequence-to-sequence inputs X, nodes weights W and output Y.

3.3 Neural Machine Translation (NMT)

Neural Machine Translation (NMT) is a sub-branch of machine translation (MT) techniques used to model natural language. Encoder and decoder are the two components that makeup NMT models. The task of the encoder is to create a true vector representation of the sentence termed the summary vector or context vector, which captures all of the vital aspects of the sentence. The perfect context vector would be able to accurately represent all the details in the source sentence as real vectors. This context vector is parsed by the decoder to create the target language sentence word by word, transferring all of the meaning from the source phrase. Translation can be modeled at different levels such as document, sentence, word, etc. NTM architectures are varied based on application by adding additional layers in a combination of encoder and decoder architectures. Fig. 3 shows a typical autoregressive NMT model using the encoder-decoder framework with an additional embedding layer.

4 The motivation of included papers

The motivation for including the seven papers mentioned in Section 1 is that these papers are related to natural language processing (NLP) and worked using convolutional neural networks which is related to the research question asked in this paper. The papers have worked on either improving the existing approaches or proposing new mechanisms using convolutional neural networks. Table

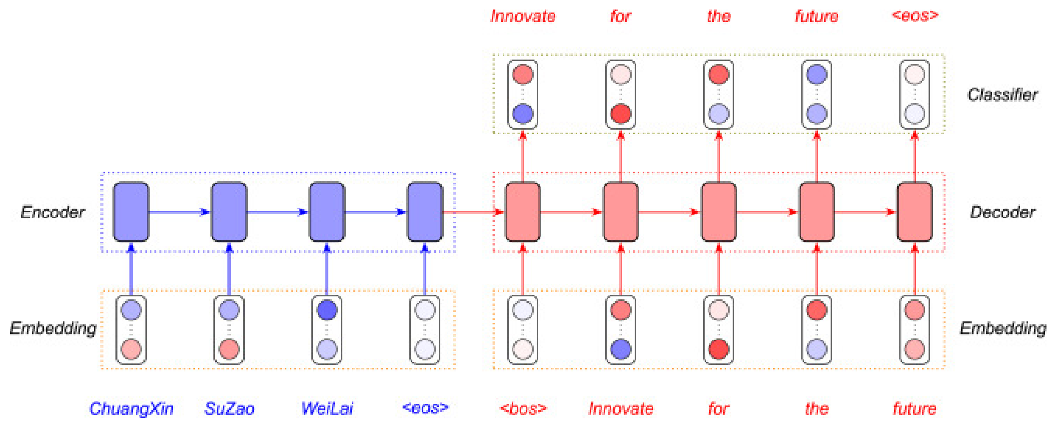

Figure 3: Source: Tan et al. (2020): An overview of the NMT architecture, which consists of embedding layers, a classification layer, an encoder network, and a decoder network. We use different colors to distinguish different languages.

S.N.	Included Research Papers	Natural Language Processing Research Areas
1	Meng et al. (2015)	Neural Machine Translation
2	Su et al. (2016)	Bilingual Semantic Modeling
3	Gehring et al. (2017a)	Neural Machine Translation
4	Gehring et al. (2017b)	Neural Machine Translation
5	Singh et al. (2017)	Neural Machine Translation
6	Wang and Xu (2017)	Word Segmentation
7	Watanabe et al. (2017)	Neural Machine Translation

Table 1: List of papers linked with the research question in this article. Five of them are on Neural Machine Translation (NMT) research areas, one of them on unsupervised learning-based word segmentation and one focuses on semantic parsing area. The commonness across all these papers is that they have some sort of improvements related to Convolutional Neural Networks (CNNs).

1 listed the paper discussed in this article - five of the research papers focus on is Neural Machine Translation (NMT) research area by making convolutional networks-based improvements, and the other two papers are related to areas: word segmentation and bilingual semantic modeling which is a field of Statistics Machine Translation (SMT) also proposing CNN based improvements. Therefore, all of these papers propose convolutional network-based solutions which is the main focus of the research question.

5 Summarization of Papers

Paper-1: Gehring et al. (2017a) presents a faster and relatively simple convolution neural network-based model for neural linguistic translation prior modeled using bi-directional LSTM. The proposed architecture speeds up the decoding process more than twice, achieving similar accuracy on the WMT'16 English-Romanian dataset and better accuracy on the WMT'15 English-Germany dataset compared to the start-of-art.

Paper-2: Meng et al. (2015) proposed a systematic approach to augment n-gram language models for machine translation by summarizing the source information through a convolutional architecture guided by target information unlike choosing heuristic information from source context window presented in Devlin et al. (2014). The architecture works supremely by suggesting the parts of a source sentence that is relevant to predicting the target word and combining the target word with the context entire source sentence to form a unified representation. Specifically designed convolution+gating architectures achieved significantly better results on two NIST Chinese-English translation tasks.

Paper-3: Singh et al. (2017) compares two encoder-decoder neural machine translation architectures where the first architecture has convolutional sequence to sequence model (ConvS2S) and the second recurrent sequence to sequence (RNNS2S). The performance is compared using the BLEU metric on English-Hindi language datasets. It is submitted that ConvS2S performance was found to be improving with hyper-parameters tunning, however, the threshold of the parameters keeping processing fixed is open to study.

Paper-4: Gehring et al. (2017b) introduces a convolutional-based architecture for sequence-to-sequence learning in place of recurrent deep LSTM-based networks. Using CNNs architecture shows several advantages: the learning is paralleled during training to fully use GPU hardware and hyper-parameters optimization becomes easier with limited fixed non-linearity functions with the process being independent of input sequence length. The experiments report better accuracy on WMT'14 English-German and WMT'14 English-French translation under Wu et al. (2016) setup.

Paper-5: According to Su et al. (2016), "Estimating similarities at different levels of linguistic units, such as words, sub-phrases, and phrases, is helpful for measuring semantic similarity of an entire bilingual phrase". The article proposes recursive neural networks (RNNs) and convolutional neural networks (CNNs) combined bilingual phrase structures generator (called ConvBRNN model) that demonstrate the high quality of NIST Chinese-English translation. The proposed RNNs architecture limits bilingual phrase structure learning to be aligned with source words, CNNs architecture integrates vector representations of linguistic units on the structures into bilingual phrase embeddings. It also introduces two max-margin losses to train the ConvBRNN model: one for the phrase structure inference and the other for the semantic similarity model.

Paper-6: In the Wang and Xu (2017) convolution-based learning is proposed to handle two major weaknesses of character-based sequence labeling framework on Chinese word segmentation (CWS). The discussed weakness are: (i) biagram feature require manual feature engineering and (ii) full word information is not fully utilized. The paper proposed a convolutional neural model to generate high-quality n-gram features automatically and tables an effective mechanism to integrate the model with word embedding. Results evaluating two benchmark datasets: PKU and MSR, the model obtains competitive performance — 95.7% on PKU and 97.3% on MSR without any feature engineering. Armed with word embedding, the model achieves 96.5% on PKU and 98.0% on MSR.

Paper-7: Watanabe et al. (2017) proposes a new attention algorithm to enhance the performance of convolutional neural networks (CNNs) based sequence to sequence encoder-decoder neural machine translation. The attention represents every combination of source words generated by neural networks. Attentions incorporated NMT gains 0.66 BLUE points better than state-of-art on the

Asian Scientific Paper Excerpt Corpus (ASPEC) English-Japanese translation task through decoding a target sentence on basis of the attention scores of the hidden states of CNNs.

6 Assessment, Interpretation, Analysis, or Discussion of the Papers

Gehring et al. (2017a) compares recurrent neural machine translation (RNMT) approach with soft attention first introduced in Bahdanau et al. (2015) and non-recurrent (NRNMT) architectures. In RNMT the input sentences (each of fixed length, 'M') are encoded and implemented using bi-directional LSTMs, that in turn are decoded to achieve linguistic translation. In the NRNMT model, the encoder consists of two stacked convolutional networks without adding pooling layers where the number 'two' is decided empirically, however, the author has not commented on the performance of more than two-stacked networks. It is clarified that the proposed model performance is independent of 'M' which shows the model can be used for a sentence of any fixed length with no new experimental setup. The results show that the translation models with convolutional encoders can translate twice as fast as strong baselines with bi-directional recurrent encoders. In the future, as the author says, CNN-based NMT parameters can be optimized to achieve higher speed and the model can be used to translate other sequence-to-sequence tasks such as summarizing constituency parsing, and dialog modeling. Also, the proposed architecture can be explored on a character-level encoder where input is way longer for words.

Meng et al. (2015) proposes a better architecture for neural machine translation using a joint model with CNNs encoder to model source language for extraction of relevant source information instead heuristically choosing a word to gain improved neural machine translation. The model use convolutional network encoded representation to predict the next words. The generic CNN architecture has fixed six layers. The author has not mentioned if the layers count is achieved empirically. The architecture has two forms of CNN named tagCNN and inCNN worth studying. It is mentioned in section 5 that, "We only keep sentence pairs that the length of source part no longer than 40 words, which covers over 90% of the sentence." but not pointed to any comment on model performance if "40" is changed.

Gehring et al. (2017b), in section 5, the author has achieved better metrics with proposed architectures using lower processing power hardware than reference work Wu et al. (2016) which seems to be promising as the result can be interesting and motivating to see with high processing power infrastructure. Section 5.5 shows that multi-step attention outperformed single layers. Further study is required to understand the threshold count for the layers and again can be outbreak finding. The paper uses 1D convolution for designing the proposed architecture. 2D convolution is considered to be capturing both temporal and context information. Given that input data used in this approach has a fixed length, experimentation with 2D convolutions architecture should be studied.

Singh et al. (2017) compares recurrent Gehring et al. (2017b) and convolution Bahdanau et al. (2015) architectures for English-Hindi Neural Machine Translation. The reported result shows that CNNs based model outperformed in terms of accuracy on Hindi to English translation while the RNNS model had better metrics for English-Hindi translation. The reasons, not studies. Also, it is not mentioned how the parameters of both the models varied and used parameters had any correlation which required deep analysis.

Su et al. (2016) presents a convolution-enhanced bilingual recursive neural network to learn bilingual semantic similarity, a research line of statistical machine translation. The paper first introduced a better form of (i) a recursive neural network to exploit words based on the work from paper Zhang et al. (2014). (ii) secondly, it added a variant of a tree-based convolutional neural network to produce bilingual phrase embeddings. The results reported are based on the final architecture, however, it would be of interest to understand the result alone from part (i) (the author has not given any comment). Another interesting study would be trying to achieve part (i) using convolutional networks as CNNs based solution work well with utilizing CUDAS processing power.

Wang and Xu (2017) work is focused on automating Chinese word segmentation. Segmentation is a form of unsupervised learning. The results support that the work done is relevant findings. The author has offered two points in this paper: (i) a proposed new convolutional neural model to capture rich n-gram features without any feature engineering and (ii) an effective approach to integrate the proposed model with word embeddings (word embedding helps to see crosswords context). Few of the values in the experimental setups were chosen randomly. For example, the training batch size is 100 and 'three' convolutional layers are stacked. Now, the number mentioned can be crucial factors influencing model training time. It would be interesting to decide empirically the best values for batch size and the number of CNNs layers stacked. These findings could guide in deciding a good experimental setup for other related works.

Watanabe et al. (2017) proposes a new attention mechanism for neural machine translation (NMT) based on convolutional neural networks (CNNs). The attention unit is a key component in NMT architecture. Better attention logic means higher accuracy in the translation task. The paper incorporates a Convolution neural networks model that imitates the CKY table into the attention mechanism of Attention NMT. The table obtained stores "sentence parsing" using CKY algorithm Kasami (1965) Younger (1967). In proposed approach showed better by 0.66 BLUE metrics, however, ended up consuming more memory and hence, inefficient to work on large datasets. Section 5 explains the experimental setup. 100,000 sentences are used for the training model while, 1812 sentences are used for testing which is less than 2% of total datasets. As a practice, it is recommended to maintain the test datasets size greater than 5% of the total data to hold the Statistical significance of the results. Thus, the results reported can vary drastically on new unseen sentences.

7 Contribution of each paper to the topic

In this paper, the research question asked is if neural machine translation can be achieved successfully to the state of the art using convolutional networks-based encoder-decoder structures. As mentioned in section 4 the seven research papers pointed out different advantages of using non-recurrent NMT and they have partially tried to achieve using convolutional networks either through achieving improvised versions or through proposing new architectures or mechanisms. The below segment explains how the mentioned papers are related to the open questions.

Gehring et al. (2017a) proposed a better encoder architecture based on convolutional networks that not only performed as par or better than existing works but also, helped to achieve faster translation. However, the decoder module of the proposed NMT is still implemented using bi-directional LSTM. What mechanisms can be studied to implement a decoder using CNNs, and how would the performance of a non-recurrent model be interesting to study and add more emphasis to the research question asked in the current article?

Meng et al. (2015) devises CNN-based mechanisms to augment the n-gram target language model, a part of neural machine translation which was originally achieved through random selection. This work is not too relevant to the pointed research question directly, however, it shows that convolutional networks are powerful to solve various linguistic problems and must be studied through all angles.

Gehring et al. (2017b) indicates how using CNNs-based encoder-decoder NTM makes it easier to exploit the GPU hardware. The work in this paper emphasizes why and how crucial it is to mechanize CNN-based architecture in NMT to achieve high-speed translation as CNNs are capable of utilizing existing hardware like GPU and CPU connection due to its architecture nature. Thus, the results shown in this paper are promising and motivating to further study convolution sequence to sequence translation and thus, soliciting to ask the research question.

Singh et al. (2017) works are relevant to the question reached as it compares recurrent network and convolutional networks performance on English-Hindi datasets and concludes that although the former-based encoder-decoder performs well in terms of accuracy, fails to achieve high processing NMT. Fast training decoding during translation is an attractive key selling factor to work on real-time big data.

Su et al. (2016) proposes a new mechanism to enhance bilingual recursive neural networks using the convolutional neural-based solution. So, the paper is important to focus on the semantic similarity of bilingual phrases for translation selection in statistical machine translation. This work is encouraging to see convolutional neural-based mechanics working across all applications of natural language processing giving another motive point to explore the research question.

Wang and Xu (2017) is not relevant to neural machine translation task rather it is related to word segmentation. However, it is motivating to see convolutional networks used for solving unsupervised linguistic problems indicating that CNNs have hidden treasures worth studying.

Watanabe et al. (2017) works validate the research question as the author proposed a new convolution networks-based attention mechanism. The attention is an extension of the encoder-decoder modeling mechanism to select target words that have more contextual fit based on attention score. As the result, attention NMT translations are more accurate than traditional NMT-based translations.

8 Conclusion

Many types of research have been already studied and implemented by the researchers' communities across the world to innovate & develop better Neural Machine Translators which are implemented using recurrent networks such as recurrent neural networks (RNNs), Long Short Term Memory (LSTM), bi-direction LSTM, and Gated Recurrent Unit (GRU) as in the current state of arts. However, the proposed architectures seem to suffer from either and/or both of the issues: low performance and lower processing speed during data training. Based on the study it is the high time to raise an interesting question, ”Can recurrent networks Neural Machine Translation be replaced by convolutional networks powered Neural Machine Translation while maintaining the performance of the state-of-art?”. The question is pertinent and the drawbacks of using RNNs-based NTM encoder-decoder seem to be already noticed in the papers discussed in Section 5. There are substantial works done on comparing two neural networks (RNNs & CNNs) based NTM architectures, reporting that RNNs-based implementations have been suffering from ”lower processing power” due to the longer training time. Relying on the observations and findings of the included research papers, it is demanding and rewarding to study techniques & methodologies to incorporate high-performance convolutional networks powered encoder-decoder architectures for neural machine translation applications that provide both the advantages of the ”faster processing time” and ”parallel processing” for linguistic translation.

References

Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, and Ayad Al-Dujaili. 2021. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. In *Journal of Big Data, 8*, pages 2196–1115. SpringerOne.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Filippo Maria Bianchi, Enrico Maiorino, Michael Kampffmeyer, Antonello Rizzi, and Robert Jenssen. 2017. *Recurrent Neural Network Architectures*, pages 23–29.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017a. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017b. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. 2022. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12175–12185.

Krishna Prakash Kalyanathaya, D Akila, and P Rajesh. 2019. Advances in natural language processing–a survey of current research trends, development tools and industry applications. *International Journal of Recent Technology and Engineering (IJRTE)*, 7:199–202.

Tadao Kasami. 1965. An efficient recognition and syntax algorithm for context-free languages. Technical Report AFCRL-65-758.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Fandong Meng, Zhengdong Lu, Mingxuan Wang, Hang Li, Wenbin Jiang, and Qun Liu. 2015. Encoding source language with convolutional neural network for machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 20–30, Beijing, China. Association for Computational Linguistics.

Sandhya Singh, Ritesh Panjwani, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2017. Comparing recurrent and convolutional architectures for English-Hindi neural machine translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 167–170, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jinsong Su, Biao Zhang, Deyi Xiong, Ruochen Li, and Jianmin Yin. 2016. Convolution-enhanced bilingual recursive neural network for bilingual semantic modeling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3071–3081, Osaka, Japan. The COLING 2016 Organizing Committee.

Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. 2020. Neural machine translation: A review of methods, resources, and tools. volume 1, pages 5–21.

Chunqi Wang and Bo Xu. 2017. Convolutional neural network with word embeddings for Chinese word segmentation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 163–172, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Taiki Watanabe, Akihiro Tamura, and Takashi Ninomiya. 2017. CKY-based convolutional attention for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1–6, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. 2018. Convolutional neural networks: an overview and application in radiology. In *Insights Imaging 9*, page 611–629. SpringerOne.

Daniel H. Younger. 1967. Recognition and parsing of context-free languages in time n3. *Information and Control*, 10(2):189–208.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually-constrained phrase embeddings for machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 111–121.