Using Vector Symbolic Architecture in Attention Mechanisms to Improve Machine Translation Models

Exploring Transparency and Organizational Reliance on Human Centric AI Systems within Business

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Kevin Sandoval

April 2, 2024

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Kathryn A. Neeley, Department of Engineering and Society

Introduction

Machine translation models have been a research focus for many years to facilitate communication between people speaking different languages, especially without the need of either person being bilingual. In recent years, there has been a surge in advancements in the machine translation task with the incorporation of the Encoder-Decoder architecture present in the recurrent neural network (RNN) (Lin et al., 2024, p. 3). The key advancement that the RNN implements is long term memory. Long term memory is a necessity for translating text from one language to another, since, to semantically understand a sentence, languages often rely on contextual details from prior or subsequent words and sentences. While RNNs has been implemented in current neutral machine translation models with some success, when the input sentences to be translated becomes very long, past contextual details can become forgotten. This phenomenon is known as the vanishing gradient problem.

To best understand the vanishing gradient problem, researchers Cashman et al. developed a RNN gradient visualization tool (2018, p. 39-40, 46). A large corpus of the C programming language was used to train a RNN model. The model attempted to generate a for-loop programming construct which looks like “for (iter…”. The vanishing gradient problem becomes apparent when the model predicts “oreapner”, where the space character, the left parenthesis, and beginning part of the phrase “iter” is incorrectly predicted, thereby displaying certain contextual details of the for-loop construct vanished from the model.

Relating back to machine translation, this vanishing gradient problem, if not addressed, can provide severe miscommunications between two parties if the translated text provides a semantic meaning different than originally intended. This can have drastic consequences, especially in critical domains such as foreign affairs, business transactions, legal matters, and

many more. Furthermore, a diminishing performance of these neural machine translation models can result in a lack of common acceptance into existing socio-technical systems. While efforts have been made to address the vanishing gradient problem, such as the introduction of the attention mechanism, several issues exist which the technical topic will address. Attention mechanisms decide which words of the text are relevant to keep, especially in consideration of translating one specific word. Yet, various issues affect the computational costs of the attention mechanism (Zhaoyang et al., 2021, p. 59). The technical topic will use vector symbolic architecture (VSA) within the attention mechanism to lessen its computational costs when translating large bodies of text. The STS topic will explore transparency and organizational reliance on human centric AI systems, specifically within the business world.

Technical Topic: Using Vector Symbolic Architecture in Attention Mechanisms to Improve Machine Translation Models

A common type of bias present in human cognition is known as anchoring bias, in which a person relies heavily on a certain piece of information (Charvi et al., 2022, 83:2). Attention mechanisms seek to address this limitation by calculating weights for each feature of the input, and based on a certain threshold, choose to either remember or forget that piece of information (Zhaoyang et al., 2021, p. 48). This specific process is illustrated in Figure 1, in which attention weights are calculated through some distribution function to ultimately produce a context vector that summarizes which parts of the input to focus on.

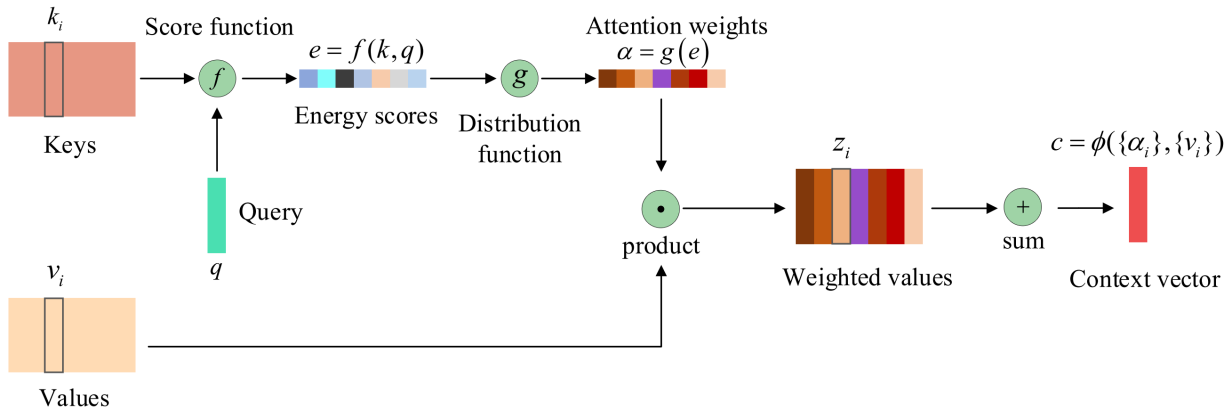

Figure 1. Diagram of an attention mechanism. The architecture and process of generating a context vector is shown as the weighted sum of the values. (Zhaoyang et al., 2021, p. 50).

In the context of machine translation models, attention mechanisms allow the model to focus on relevant details of the given input sentence(s) in the process of generating the translated text. While the attention mechanism has successfully been implemented and even integrated into recurrent neural networks (the architecture behind modern neural machine translation models), several uncertainties and problems exist which impact its performance. One such uncertainty is the representation of keys and queries. In Figure 1, the keys and queries are represented as separate components. However, current research is being done to explore the combination of these components in some way for more efficient computations and better performance. In the context of machine translation models, keys currently represent the input text and the queries represent the target language text in a process known as multi-head attention (Vaswani et al, 2017, p. 5)

Another issue of attention mechanisms is the distribution function, which may have room for improvement, once again, in terms of performance and computational cost reduction (Zhaoyang et al., 2021, p. 59). The distribution function ultimately calculates which parts of the input to focus on by calculating attention weights. These uncertainties ultimately pose high costs

in generating attention bias in the translational model, in addition to decreased performance (Fuchs et al., 2023, 4:38). Furthermore, by not addressing these uncertainties in the attention mechanism, it will fail to reach similar or better performance to current translational models and as a result, would be more likely to produce incorrect translations. In focusing on improving these uncertainties, the reliability of the attention mechanism increases, thereby increasing its trustworthiness in high-stakes environments.

There are multiple directions to investigate and address the current issues with attention mechanisms. However, the primary investigation will rely on VSA, which can handle tasks like learning various tasks simultaneously and efficiently (Cheung et al., 2019, p. 2). In human cognition, binding is known as a process that is used to represent an environment. For example, a task might be binding words to their numeric position within a sentence. In VSA, vectors of dimension N are used to represent the properties of different objects (such as position and the words themselves) (Hiratani et al., 2023, p. 2). To define a specific object in the environment, a pair of vectors is "bundled", which involves a binding operator. Figure 2 depicts a VSA representation of an environment of shapes, which Hiratani et al. (2023) explains "should enable us to answer a question like 'what is the left-most object?' (Answer: pink-cube)…" (p. 2). Research is being conducted to find the optimal binding method that can be generalized to many types of vectors. This VSA architecture might prove to address the current limitations with attention mechanisms since vectors are heavily involved at every step of the model, and generating a context vector describes a binding task.

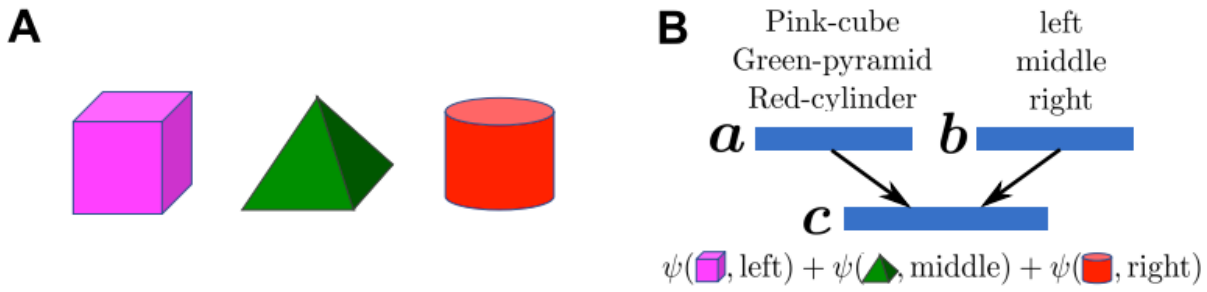

Figure 2: VSA representation of a binding problem. The environment in A) can be represented by forming tuples from vectors *a* and *b*. (Hiratani et al., 2023, p. 2)

The anticipated deliverable will be a technical algorithm that utilizes the work of Cheung, Hiratani, et al. to efficiently create the models in attention mechanism to increase its computational performance and lower its high costs in translation tasks. In particular, a methodology will be explored to translate the attention mechanism's vector-based keys and queries to the VSA environment. The translation will be the most challenging to implement since research about this topic is not fully explored and the translated problem needs to be precisely equivalent to the original problem. An optimal binding operator will then be researched using the research of Hiratani to minimize the unbinding performance necessary to efficiently generate the weights and context vector. Parts of an existing implementation of an attention mechanism will be evaluated and compared with the performance of the algorithmic developments found through this technical research proposal.

STS Topic: Exploring the Transparency and Organizational Reliance on Human Centric AI Systems within Business Environments

Human-Centric AI (HCAI) seeks to integrate AI into human activity by interpreting human behavior, predicting their choices, and "[orchestrating] between actions performed directly by humans and those delegated to AI agents in autonomy" (Fuchs, 2023, 4:2). However, modern AI systems are becoming increasingly complex to the point where their internal

mechanisms become unrecognizable and difficult to understand, a concept known as a "black box". This complexity ultimately raises the concern of trusting such systems within everyday use, warranting a warning of relying on systems that aren't fully understood because they are more error prone (Von Braun et al., 2021, p. 165). To achieve the HCAI goal of integrating AI systems within human activity, people must be able to trust that these systems will benefit their lives in terms of productivity.

Prior research by Ammanath (2022) studies the various factors influencing the trust of AI systems, specifically within a business environment (Robust and Reliable). These factors include robustness, which increases an organization's likelihood to incorporate the system within any part of the business ecosystem when present within AI systems. Organizational actors such as CEOs, corporate leaders, and managers are typically hesitant to rely on AI systems since even a minor failure within a small part of an AI system can cause a significant cost to a company. To illustrate this, Ammanath (2022) provides the example where "…if the machine itself is set a foot higher or lower, the tool's accuracy may degrade or suffer a catastrophic failure" (The Challenge of Generalizable Deep Learning). This cost highlights that an AI system can disrupt or halt the entire operations of a business, imposing monetary costs and customer dissatisfaction.

Various approaches have been developed to address the organizational hesitancy in integrating AI systems and mitigating associated costs. Figure 3 shows an overview of a process that outlines how organizations can increase trust in AI systems. The goal of a business in using AI systems is to improve the company's productivity, which requires mitigating the costs associated with those systems. In the previous example, catastrophic failure is identified as a potential consequence, but as depicted in Figure 3, this can be mitigated by increasing the robustness of AI models. There are multiple methodologies to enhance the robustness of AI since

many models exist. However, a common approach is to increase the diversity of the data used when training an AI model so that it is more likely to succeed in various environments.

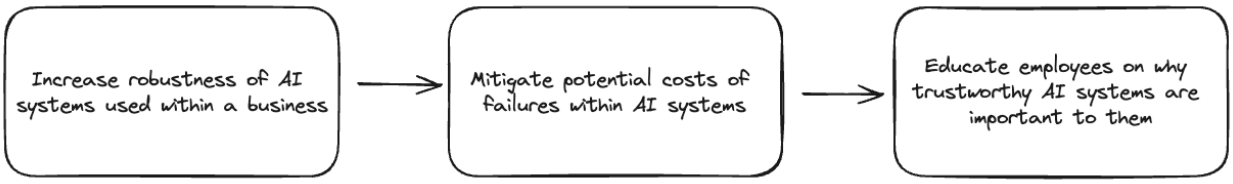

Figure 3: Process to incorporate trust of AI systems within businesses. Business leaders increase the robustness of AI systems, thereby decreasing failures in those systems. However, they still need to influence the employees to trust those systems they might use. (Created by Author).

From an organizational perspective, the robustness of AI systems can be increased so that organizations are more likely to rely on them throughout a business. Yet, one challenge is convincing other actors to trust those systems that they often utilize. In developing a potential algorithm to increase the robustness of the attention mechanism, other users within the business environment may be hesitant to rely on it if they do not fully understand the algorithm. Using the STS methodology of the actor network theory (ANT), the users of the AI systems can be identified in the business system. Through this, connections can be formed to determine how robust AI influences the users' ability to trust AI systems. By better understanding why those users can trust the AI system, the algorithm has a better chance of being utilized in a business ecosystem, starting from the organizational leaders, and ending with the product delivered to customers. Hence, by highlighting the importance of transparency, organizational leaders can convince workers, other organizational members, and customers on why this the proposed algorithm can be accepted within the business. The anticipated STS deliverable will be increasing the awareness of the importance of transparency in HCAI systems.

Conclusion

The deliverable for the technical research will be an algorithm to efficiently create representations of translation tasks that consider biases within the attention mechanism, which

will draw on VSA research conducted by Hiratani et al. The anticipated STS deliverable will be using increased robustness and ANT to improve the transparency of HCAI systems. With the technical deliverable, the hyper-dimensional vector algorithm could improve the outcomes of neural machine translation models to produce better translated text that effectively maintains the semantic meaning of the input text. Combined with the STS deliverable, this algorithm will be shown to perform the translation task at a similar or better reliability and performance than current neural machine translation models that don't employ the attention mechanism. This therefore strengthens the trust of these systems and increasing its likelihood of being used in all parts of a business or other sectors.

References

Ammanath, B., & O'Reilly Online Learning: Academic/Public Library Edition (2022). *Trustworthy AI: A business guide for navigating trust and ethics in AI*. Hoboken, NJ: Wiley.

Cashman, D., Patterson, G., Mosca, A., Watts, N., Robinson, S., & Chang, R. (2018, November 1). RNNbow: Visualizing learning via backpropagation gradients in RNNs. *IEEE Computer Graphics and Applications*, 38(6), 39 - 50.

Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. (2022). Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *Proceedings of the Association for Computing Machinery on Human-Computer Interaction*. 6(CSCW1), Article 83, 1-22.

Cheung, B., Terekhov, A., Chen, Y., Agrawal, P., & Olshausen, B. (2019). Superposition of many models into one. *Advances in Neural Information Processing Systems*, 32, 1-10.

Fuchs, A., Passarella, A., & Conti, M. (2023, June 1). Modeling, replicating, and predicting human behavior: A survey. *Association for Computing Machinery Transactions on Autonomous & Adaptive Systems*, 18(2), 1 - 47.

Hiratani, N., & Sompolinsky, H. (2023). Optimal quadratic binding for relational reasoning in vector symbolic neural architectures. *Neural Computation*, 35(2), 105-155.

Lin, Y., & Wang, Z. (2024, January 2). A novel method for linguistic steganography by English translation using attention mechanism and probability distribution theory. *PLoS ONE*, 19(1), 1 - 23.

Matsumori, K., Koike, Y., & Matsumoto, K. (2018). A biased Bayesian inference for decision-making and cognitive control. *Frontiers in Neuroscience*, 12(1), 1-16.

Ragni, M., Eichhorn, C., Bock, T., Kern-Isberner, G., & Tse, A. (2017, March 1). Formal nonmonotonic theories and properties of human defeasible reasoning. *Minds & Machines*, 27(1), 79 - 117.

Riesterer, N., Brand, D., Dames, H., & Ragni, M. (2020, January 1). Modeling human syllogistic reasoning: The role of "no valid conclusion". *Topics In Cognitive Science*, 12(1), 446 - 459.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems,* 5998 – 6008.

Von Braun, J., Archer, Margaret S. (Margaret Scotford), Reichberg, G. M., & Sánchez Sorondo, M. (Eds.) (2021). *Robotics, AI, and humanity: Science, ethics, and policy*. Cham: Springer. 165.

Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. 2021. A review on the attention mechanism of deep learning. *Neurocomputing* 452 (2021), 48–62.