TRANSFORMER EXPLAINER: Interactive Learning of Text-Generative Models

Aeree Cho[1*], Grace C. Kim[1*], Alexander Karpekov[1*], Alec Helbling[1], Zijie J. Wang[1], Seongmin Lee[1], Benjamin Hoover[1, 2], Duen Horng (Polo) Chau[1]

[1]Georgia Tech,
[2]IBM Research
{aeree|gracekim3|alex.karpekov|alechelbling|jayw|seongmin|bhoov|polo}@gatech.edu

Abstract

Transformers have revolutionized machine learning, yet their inner workings remain opaque to many. We present TRANSFORMER EXPLAINER, an interactive visualization tool designed for non-experts to learn about Transformers through the GPT-2 model. Our tool helps users understand complex Transformer concepts by integrating a model overview and smooth transitions across abstraction levels of math operations and model structures. It runs a live GPT-2 model locally in the user's browser, empowering users to experiment with their own input and observe in real-time how the internal components and parameters of the Transformer work together to predict the next tokens. 125,000 users have used our open-source tool at https://poloclub.github.io/transformer-explainer/.

Introduction

The Transformer (Vaswani et al. 2017) has gained popularity as a neural network architecture across numerous tasks, from text to vision. Despite its increasing widespread use, particularly in AI chatbots ChatGPT, its inner workings often remain opaque, hindering understanding and engagement (Yeh et al. 2023). Demystifying this architecture is crucial for non-experts interested in learning about it. Existing resources, such as blog posts (Alammar 2018), video tutorials (Karpathy 2024), and 3D visualizations (Bycroft 2023) often emphasize mathematical intricacies and model implementations, which can overwhelm beginners. Meanwhile, visualization tools designed for AI practitioners focus on interpretability at the neuron and layer levels, which can be challenging for non-experts to grasp (Braşoveanu and Andonie 2020). To address this research gap, we contribute:

TRANSFORMER EXPLAINER, an open-source, web-based interactive visualization tool designed for non-experts to learn about both the Transformer's high-level model structure and low-level mathematical operations (Fig. 1). Our tool explains the Transformer through its application in text generation, one of its most recognized uses. It adopts the Sankey diagram visual design, inspired by recent studies viewing the Transformer as a dynamic system (Dutta et al. 2021), to emphasize how input data "flows" through the model's components (Elhage et al. 2021). The Sankey diagram effectively illustrates how information moves through the model, showcasing how inputs undergo processing and transformations via Transformer operations. Our tool helps users gain a comprehensive understanding of the complex concepts within Transformers by tightly integrating a model overview that summarizes a Transformer's structure and enabling users to smoothly transition between multiple abstraction levels to visualize the interplay between low-level mathematical operations and high-level model structures (Fig. 1C) (Wang et al. 2020).

Open-source, web-based implementation with real-time inference. Unlike many existing tools that require custom software installations or lack inference capabilities (Braşoveanu and Andonie 2020), TRANSFORMER EXPLAINER integrates a live GPT-2 model that runs locally in the user's browser using modern front-end frameworks. Users can interactively experiment with their own input text (Fig. 1A), and observe in real time how the internal components and parameters of the Transformer work together to predict the next tokens (Fig. 1B). Our approach broadens educational access to modern generative AI techniques without requiring advanced computational resources, installations, or programming skills. We have chosen GPT-2 for its widespread familiarity, fast inference speed, and architectural similarities to more advanced models like GPT-3 and GPT-4, making it ideal for educational purposes.

System Design and Implementation

TRANSFORMER EXPLAINER visualizes how a trained Transformer-based GPT-2 model processes text input and predicts the next token. The frontend uses Svelte and D3 for interactive visualizations, while the backend uses ONNX runtime and HuggingFace Transformers to run the GPT-2 model in the browser. A major challenge in designing our tool is managing the underlying architecture's complexity, which can be overwhelming if all details are presented. To address this, we draw on two key design principles:

Reducing Complexity via Multi-Level Abstractions. We structured the tool to present information at varying levels of abstraction. This approach allows users to start with a high-level overview and drill down into details as needed, preventing information overload (Kahng et al. 2017). At the

*These authors contributed equally.

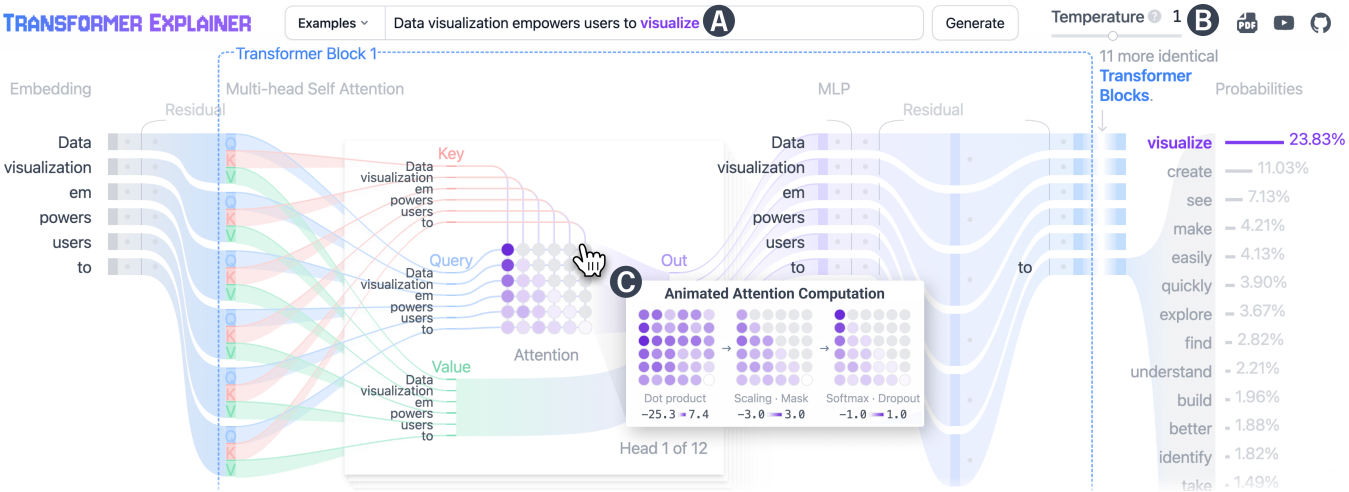

Figure 1: TRANSFORMER EXPLAINER helps users (A) visually examine how Transformer model (GPT-2) transforms input text to predict next tokens, (B) interactively experiment in real time with model parameters like *temperature* to understand prediction determinism, and (C) transition seamlessly between low-level mathematical operations and high-level model structures.

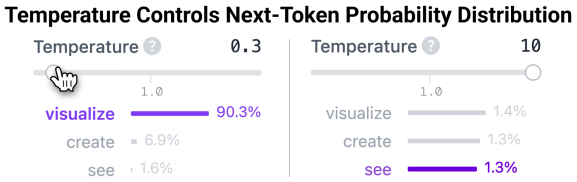

Figure 2: The temperature parameter slider lets users experiment with its impact on the next token's probability distribution. **Left**: lower temperatures sharpen the distribution, making outputs more predictable. **Right**: higher temperatures smooth it, resulting in less predictable outputs.

highest level, our tool illustrates the complete pipeline: from taking user-provided text as input (Fig. 1A), embedding it, processing it through multiple Transformer blocks, to using this data to rank the most likely next token predictions. Intermediate operations, such as calculations for the attention matrix (Fig. 1C), are collapsed by default to visualize the significance of the computational result, with the option to expand to inspect its derivation through animation. We employed a consistent visual language, such as stacking Attention Heads and collapsing repeated Transformer Blocks, to help users recognize repeating patterns in the architecture, while maintaining the end-to-end flow of data.

Enhancing Understanding and Engagement via Interactivity. The *temperature* parameter is crucial in controlling a Transformer's output probability distribution, affecting whether the next-token prediction will be more deterministic (at *low* temperature) or random (*high* temperature). Existing educational resources on Transformers often overlook this aspect (Alammar 2018). Our tool enables users to adjust the temperature in real time (Fig. 1B) and visualize its critical role in controlling the prediction determinism (Fig. 2). Additionally, users can select from provided examples or enter their own input text (Fig. 1A). Supporting custom input text engages users, by allowing them to analyze the model's behavior under various conditions and interactively test their own hypotheses on diverse text inputs.

Usage scenario. Professor Rousseau is modernizing the Natural Language Processing course to highlight recent advances in generative AI. Noticing that some students view Transformer as "magic", she introduces TRANSFORMER EXPLAINER, an interactive tool that provides an overview of Transformer (Fig. 1) for hands-on learning. With over 300 students, its browser-based access without installation eliminates setup concerns. The tool uses animations and interactive abstractions (Fig. 1C) to explain complex operations like attention computation, helping students grasp both high-level concepts and lower-level details that produce the results. She is also aware that the technical capabilities of Transformers are sometimes obscured by anthropomorphism (e.g., the *temperature* parameter being viewed as a "creativity" control). By having students experiment with the temperature slider (Fig. 1B), she demonstrates that it actually modifies the probability distribution of the next token (Fig. 2), balancing prediction randomness and determinism. Additionally, as the system visualizes the token processing flow, students see there is no "magic" involved — regardless of the input texts (Fig. 1A), the model follows a clear sequence of operations using the Transformer architecture, sampling one token at a time before repeating the process.

Impact and Future Work

Upon release, our tool sparked overwhelming excitement, amassing over 125,000 users. We plan to boost inference speed with WebGPU and reduce model size through compression techniques (e.g., quantization, palettization) (Hohman et al. 2024). We also plan to conduct user studies to assess our tool's efficacy and usability, observe how newcomers to AI, students, educators, and practitioners use it, and gather feedback for new features.

References

Alammar, J. 2018. The Illustrated Transformer. https://jalammar.github.io/illustrated-transformer/.

Braşoveanu, A. M. P.; and Andonie, R. 2020. Visualizing Transformers for NLP: A Brief Survey. In *24th International Conference Information Visualisation (IV)*, 270–279.

Bycroft, B. 2023. LLM Visualization. https://bbycroft.net/llm.

Dutta, S.; Gautam, T.; Chakrabarti, S.; and Chakraborty, T. 2021. Redesigning the Transformer Architecture with Insights from Multi-particle Dynamical Systems. In *NeurIPS*.

Elhage, N.; et al. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.

Hohman, F.; Wang, C.; Lee, J.; Görtler, J.; Moritz, D.; Bigham, J. P.; Ren, Z.; Foret, C.; Shan, Q.; and Zhang, X. 2024. Talaria: Interactively optimizing machine learning models for efficient inference. In *CHI*.

Kahng, M.; Andrews, P. Y.; Kalro, A.; and Chau, D. H. 2017. ActiVis: Visual exploration of industry-scale deep neural network models. *IEEE TVCG*, 24(1): 88–97.

Karpathy, A. 2024. Let's build GPT: from scratch, in code, spelled out. https://youtu.be/kCc8FmEb1nY.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. NIPS'17, 6000–6010. Curran Associates Inc. ISBN 9781510860964.

Wang, Z. J.; Turko, R.; Shaikh, O.; Park, H.; Das, N.; Hohman, F.; Kahng, M.; and Chau, D. H. 2020. CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization. *TVCG*.

Yeh, C.; Chen, Y.; Wu, A.; Chen, C.; Viégas, F.; and Wattenberg, M. 2023. Attentionviz: A global view of transformer attention. *TVCG*.