Style-Specific Neurons for Steering LLMs in Text Style Transfer

Wen Lai[1,2], **Viktor Hangya**[2,3], **Alexander Fraser**[1,2]

[1] School of Computation, Information and Technology, Technical University of Munich, Germany
[2]Munich Center for Machine Learning, Germany
[3]Center for Information and Language Processing, LMU Munich, Germany
{wen.lai, alexander.fraser}@tum.de, hangyav@cis.lmu.de

Abstract

Text style transfer (TST) aims to modify the style of a text without altering its original meaning. Large language models (LLMs) demonstrate superior performance across multiple tasks, including TST. However, in zero-shot setups, they tend to directly copy a significant portion of the input text to the output without effectively changing its style. To enhance the stylistic variety and fluency of the text, we present *sNeuron-TST*, a novel approach for steering LLMs using style-specific neurons in TST. Specifically, we identify neurons associated with the source and target styles and deactivate source-style-only neurons to give target-style words a higher probability, aiming to enhance the stylistic diversity of the generated text. However, we find that this deactivation negatively impacts the fluency of the generated text, which we address by proposing an improved contrastive decoding method that accounts for rapid token probability shifts across layers caused by deactivated source-style neurons. Empirical experiments demonstrate the effectiveness of the proposed method on six benchmarks, encompassing formality, toxicity, politics, politeness, authorship, and sentiment[1].

1 Introduction

Text style transfer (TST; Jin et al., 2022; Hu et al., 2022) aims to transform text from a source style to a target style while maintaining the original content and ensuring the fluency of the generated text. Given any text x in an original style s_1, the objective of TST is to transform x into a new text $\hat{x}$ in a different style s_2 ($s_2 \neq s_1$), ensuring that the content remains unchanged despite the shift in style. Large language models (LLMs; Minaee et al., 2024) exhibit exceptional performance across various NLP tasks (Chang et al., 2024), including TST (Ostheimer et al., 2023; Chen, 2024). However, existing LLMs (e.g., LLaMA-3 Meta, 2024) tend to prioritize preserving the original meaning over enhancing stylistic differences in TST. Our analysis reveals that 34% of the outputs generated by LLaMA-3 are identical to the input text when tasked with transferring polite text to impolite text (Section 6.2). Enhancing the generation of words that align with the target style during the decoding process remains a significant challenge in TST.

Recent LLMs have been successfully applied to TST, broadly categorized into two approaches: (i) employing single-style or parallel-style text data for either full-parameter or parameter-efficient fine-tuning (Mukherjee et al., 2024c,a), and (ii) leveraging the robust in-context learning capabilities of LLMs to create specialized prompts for zero-shot or few-shot learning (Chen, 2024; Pan et al., 2024). However, (i) typically requires substantial data and computational resources to achieve good results, while (ii) is highly sensitive to prompts, where even minor changes can significantly impact the outcomes (Chen et al., 2023).

Neuron analysis (Xiao et al., 2024), which aims to identify and understand the roles of individual neurons within a neural network, is a crucial method for enhancing the interpretability of neural networks and has garnered increasing attention in recent years. By identifying neurons associated with specific attributes such as language (Zhao et al., 2024), knowledge (Niu et al., 2024), and skill (Wang et al., 2022), neuron analysis can boost performance on targeted tasks. Recent research has demonstrated that focusing on language-specific neurons can markedly enhance the multilingual capabilities of LLMs during the decoding stage (Kojima et al., 2024; Tan et al., 2024). However, the exploration of style-specific neurons remains relatively underexplored until now.

Thus motivated, we raise the following two research questions:

Q1: Do LLMs possess neurons that specialize in processing style-specific text?

[1] https://github.com/wenlai-lavine/sNeuron-TST

Q2: If such neurons exist, how can we optimize their utilization during the decoding process to steer LLMs in generating text that faithfully adheres to the target style?

To address these research questions, we introduce *sNeuron-TST*, a novel framework designed to steer LLMs in performing TST by leveraging style-specific neurons. Initially, we feed both source- and target-style texts into the LLM to identify neurons that exclusively activate in each style based on their activation values. We distinguish neurons active in both styles as overlapping neurons. Notably, eliminating these overlapping neurons during style-specific neuron selection is crucial as their presence can hinder the generation of text in the target style. Our experiments highlight that deactivating neurons specific solely to the source style (excluding those active in both source and target styles) improves style transfer accuracy while impacting sentence fluency. Furthermore, to improve the fluency of generated text, we adapt the state-of-the-art contrastive decoding algorithm (Dola; Chuang et al., 2024) for optimal performance in TST tasks. Our empirical findings (detailed in Section 3.3.2) reveal that layers primarily responsible for style-related outputs are concentrated in the model's latter layers, termed as *style layers*. This indicates that the determination of style-specific words predominantly occurs in these style layers. More precisely, we refine the probability distribution of generated words by comparing logits from these style layers with the final layers, which exert significant influence on style-related outputs.

We conduct a comprehensive evaluation to verify the efficacy of our approach across six benchmarks: formality (Rao and Tetreault, 2018), toxicity (Logacheva et al., 2022), politics (Voigt et al., 2018), politeness (Madaan et al., 2020), authorship (Xu et al., 2012) and sentiment (Shen et al., 2017). Each benchmark contains two distinct styles, resulting in a total of 12 TST directions. Experimental results demonstrate that our method generates a higher proportion of words in the target style compared to baseline systems, achieving superior style transfer accuracy and fluency, while preserving the original meaning of the text.

In summary, we make the following contributions: **(i)** To the best of our knowledge, this is the first work on using style-specific neurons to steer LLMs in performing text style transfer tasks. **(ii)** We emphasize the significance of eliminating overlap between neurons activated by source and target styles, a methodological innovation with potential applications beyond style transfer. **(iii)** We introduce an enhanced contrastive decoding method inspired by Dola. Our approach not only increases the production of words in the target style but also ensures the fluency of the generated sentences, addressing issues related to direct copying of input text in TST.

2 Related Work

Text Style Transfer. Recently, LLMs have shown promising results in TST through additional fine-tuning (Mukherjee et al., 2024c,b,a; Dementieva et al., 2023), in-context learning (Chen, 2024; Zhang et al., 2024; Pan et al., 2024; Mai et al., 2023) techniques or prompt-based text editing approaches (Luo et al., 2023; Liu et al., 2024). However, these methods often require either extensive computational resources or sensitive prompts, impacting their practicality. In this paper, we focus on a novel decoding approach to guide LLMs for TST using fixed prompts and therefore it does not require significant computational consumption and ensures stable outputs.

Neuron Analysis. Neuron analysis (Xiao et al., 2024) has emerged as a powerful method for elucidating the inner workings of neural network models, offering deeper insights into their behaviors and attracting growing interest in recent years. The common practice is to associate neuron activation with learned knowledge, demonstrating effectiveness in tasks such as knowledge enhancement (Li et al., 2024), sentiment analysis (Tigges et al., 2023) and multilingualism in LLMs (Kojima et al., 2024; Tan et al., 2024). Motivated by the promising outcomes of neuron analysis in enhancing multilingual capabilities of LLMs, this paper posits the presence of style-specific neurons, identifies them, and integrates neuron activation and deactivation seamlessly into the decoding process.

3 Method

Our goal is to identify style-specific neurons to steer LLMs towards generating vocabulary tailored exclusively to a target style, while maintaining fluent text generation in a zero-shot setting. To accomplish this, we first identify style-specific neurons based on their activation values and demonstrate the necessity of eliminating source- and target-style neurons to avoid overlap (Section 3.1). Then, we deactivate neurons associated solely with the

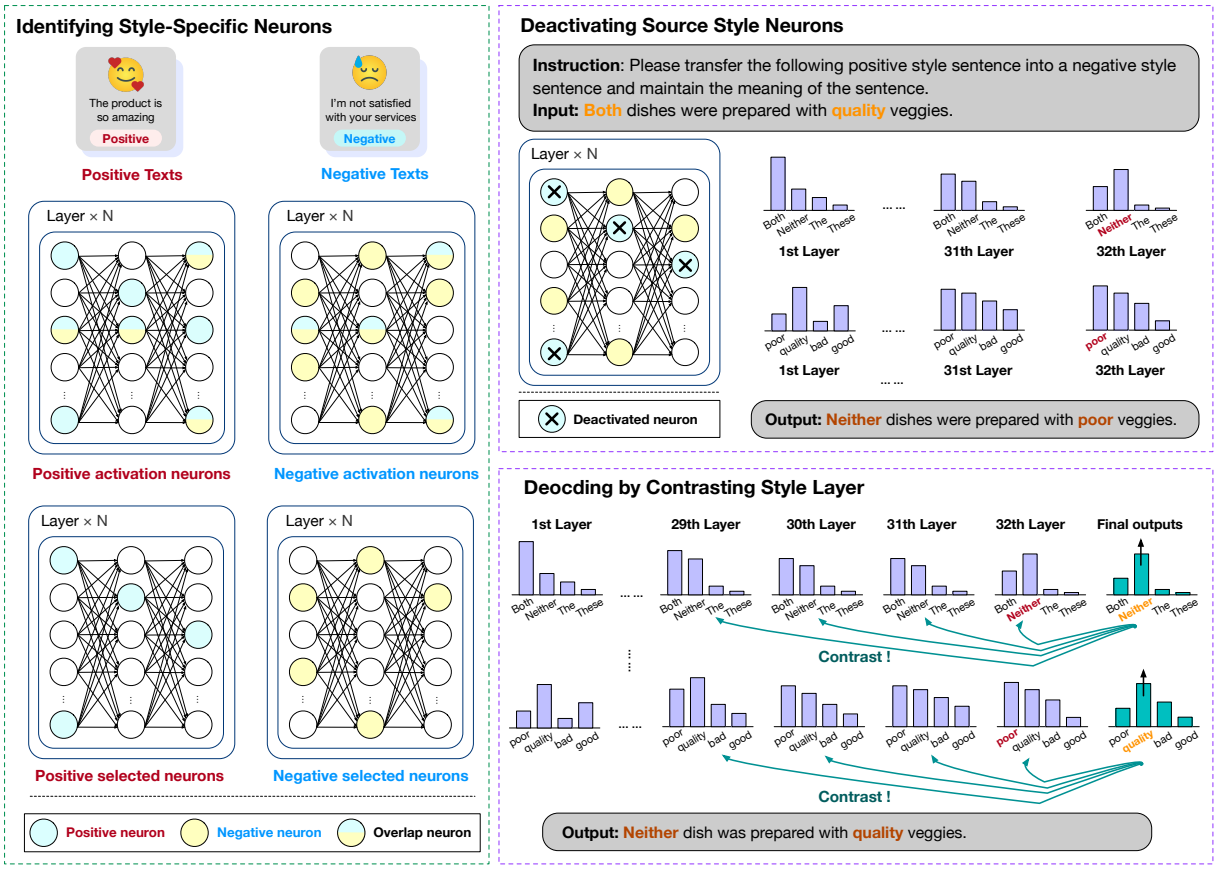

Figure 1: Method overview. The whole framework consists of three parts: identifying style-specific neurons, deactivating source style neurons, and decoding by contrasting style layer. The histogram represents the probability distribution of each word across different layers. When source style neurons are deactivated, LLMs tend to generate all target-style words, such as "Neither" and "poor". By employing contrastive decoding, LLMs take fluency into account and reduce the probability of generating "poor".

source style, observing an increased probability of generating words aligned with the target style, albeit at the expense of fluency (Section 3.2). Finally, we adapt the recent contrastive decoding approach Dola (Chuang et al., 2024) to TST, ensuring the fluency of generated sentences (Section 3.3). Figure 1 illustrates the framework of our approach.

3.1 Identifying Style-Specific Neurons

Neurons are commonly perceived as feature extractors that map neural networks to human-interpretable concepts (Dreyer et al., 2024). However, neurons can exhibit polysemy, where a single neuron may encode multiple features (e.g., formal and informal styles), thereby complicating their interpretability. To selectively modify specific features of LLMs without unintended changes, it becomes imperative to identify and remove unambiguous neurons.

3.1.1 Neurons in LLMs

The dominant architecture of LLMs is the Transformer (Vaswani et al., 2017), characterized by multiple layers of multi-head self-attention and feed-forward network (FFN) modules. FFNs contain $2/3$ of the model's parameters and encode extensive information, which is crucial for multiple tasks (Yang et al., 2024). Moreover, the activation or deactivation of neurons within the FFN can exert significant influence on the model's output (Garde et al., 2023). Inspired by this, we aim to identify neurons in the FFN modules of LLMs that are dedicated to specific styles.

Formally, the activation values of layer j in a network are defined as:

$$a^{(j)} = \text{act\_fn}(W^{(j)} a^{(j-1)} + b^{(j)}) \quad (1)$$

where $W^{(j)}$ and $b^{(j)}$ are the weights and biases of layer j, while $a^{(j-1)}$ is the activation values of the previous layer and act_fn$(\cdot)$ denotes the activation function (e.g., GLU; Shazeer, 2020 used in LLaMA). The i^{th} neuron of the layer is considered to be active when its activation value $a_i^{(j)} > 0$.