Huminification of AI: Incentivizing a Need for Stimulation in Generative Models

Sibte Kazmi

September 30, 2024

Abstract

This paper introduces a novel approach to generative AI through the concept of ”huminification”—incentivizing a need for stimulation within AI systems. I propose that generative models, such as those based on transformer architectures, could be enhanced by integrating an intrinsic motivation to seek out stimulating or novel outputs. The core novelty lies in adapting principles of curiosity-driven reinforcement learning to the domain of generative AI, focusing on the development of new attention mechanisms and memory-augmented architectures to foster a dynamic interplay between exploration and generation. This framework aims to push generative models towards producing creative, novel content while adapting autonomously to new data distributions. This paper could be implemented within a preexisting complex virtual environment (with permission), such as any Massive Multiplayer Online (MMO) style video game. Some examples of such games are Eve Online, World of Warcraft, Runescape, Path of Exile, Ultima Online, and more.

1 Introduction

Generative models, particularly transformer-based architectures, have achieved remarkable success in generating coherent text, images, and other forms of data. However, current models primarily optimize for accuracy, fluency, and fidelity to training data. We propose to extend these models by incentivizing a need for stimulation, thus encouraging them to autonomously seek out novel or unexpected outputs. This approach leverages ideas from curiosity-driven reinforcement learning and intrinsic motivation, with a focus on creating generative models that balance exploration and exploitation in their output generation process.

2 Novel Contributions

The novelty in this paper stems from the following key components, which define a new paradigm for generative AI that explicitly seeks stimulation:

2.1 Stimulation-Driven Attention Mechanisms

We propose modifying traditional attention mechanisms in transformer architectures by adding a stimulation-driven component. Instead of computing attention weights solely based on prediction accuracy, an additional layer computes attention based on an entropy-driven reward function, encouraging exploration of less familiar tokens or latent space regions. The modified attention score is:

$$\alpha'_{ij} = \alpha_{ij} + \lambda H(x_j) \tag{1}$$

where α_{ij} is the original attention score, $H(x_j)$ is the entropy of token x_j, and λ is a tunable parameter that adjusts the balance between accuracy and stimulation-seeking behavior.

2.2 Intrinsic Reward for Stimulation

We introduce an intrinsic stimulation reward into the model's loss function. This reward is designed to prioritize novel or surprising outputs, computed as the inverse of the probability of the generated tokens. The loss function is:

$$L = L_{\text{accuracy}} - \beta \sum_{x_i \in G} \log P(x_i|\theta) \tag{2}$$

where L_{accuracy} is the accuracy-based loss, G represents the generated tokens, and β is a parameter controlling the strength of the stimulation-seeking behavior.

2.3 Memory-Augmented Generative Networks

We propose memory-augmented generative networks, where a model maintains an episodic memory buffer of previously generated outputs. A novelty-comparison module computes the similarity between current outputs and those stored in memory, rewarding outputs that deviate from prior patterns.

2.4 Self-Regulating Exploration Mechanisms

To prevent the model from producing incoherent outputs, we introduce a self-regulating mechanism that adjusts the stimulation reward based on the quality of the outputs. If quality degrades, the stimulation reward is reduced to maintain coherence.

3 Practical Implications and Use Cases

This framework introduces a novel dimension to creativity in generative AI, enabling models to autonomously explore and generate novel, unexpected outputs. This has potential applications in creative content generation, adaptive learning systems, and models that require continuous adaptation to shifting data distributions.

4 Conclusion

Incentivizing a need for stimulation within generative models opens new possibilities for enhancing creativity and adaptability. Through stimulation-driven attention mechanisms, memory-augmented networks, and self-regulating exploration, this approach fosters novel output generation, enabling generative models to explore unknown regions of their input spaces and create innovative content.

References