iGPT

1 Introduction

Current generative language models process text sequentially, predicting one token at a time with autoregressive decoding. This approach results in quadratic computational complexity due to self-attention mechanisms, making it inefficient for long-context understanding and generation. However, this differs significantly from how humans process and generate language. Instead of constructing sentences token by token, human cognition forms structured thoughts before verbalizing them. This discrepancy suggests that current language modeling paradigms might not be optimally aligned with human cognitive mechanisms.

In this proposal, I aim to develop a new foundation model inspired by the way humans process and structure thoughts. Instead of relying solely on sequential token prediction, the model will first generate a condensed conceptual representation - an idea - before expanding it into linguistic units. This mirrors human cognition, where an abstract thought is progressively refined into words and sentences. Each idea is represented as an object with relationships to other ideas through spatial, contextual, or causal associations, forming a structured representation of knowledge. By iterating this process, the model can generate coherent language while maintaining a more efficient and interpretable structure. Moreover, new ideas emerge based on preceding ones, capturing a more human-like mode of reasoning and discourse generation.

2 Method

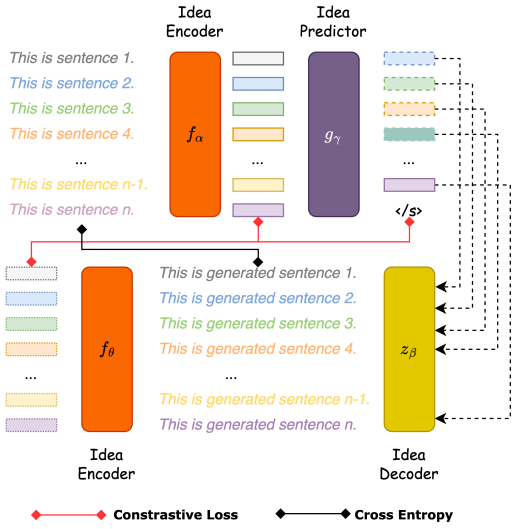

Figure 1: Enter Caption

Idea Encoder. Given an input text $\hat{y}$, we segment it into N non-overlapping sentences, and feed these through an idea encoder f_α to obtain a corresponding representation $\hat{i}_y = i_{y1}, ..., i_{yN}$, where $\hat{i}_{yk}$ is the representation associated with the k^{th} sentence.

Idea Predictor. The goal is to predict the next sentence representation and use this to reconstruct it into a sequence of tokens. To obtain the next representation, the representations output from idea encoder is passed into an idea predictor g_γ to predict the next representation in the sequence $\tilde{i}_{yk+1} = g_\gamma(i_{yk}), k \in \{1, ..., N\}$. The predicted representation is parameterized by a shared learnable vector with an added positional embedding.

Idea Decoder. Given the output of the idea predictor $\tilde{i}_y$, we wish to reconstruct the corresponding sentence using these representations. For a given position k^{th} corresponding to sentence k^{th}, the idea decoder $z_\beta(\cdot,\cdot)$ takes a k^{th} representation and a start of sequence token to regressively reconstruct the initial sentence $\tilde{y}_k = z_\beta(\tilde{i}_{yk}, < s >)$. The predicted token is parameterized by a shared learnable vector with an added positional embedding. These reconstructed sentences are then passed again into the idea encoder to get the corresponding representation for later loss calculation.

Loss. The has three parts as bellows:

- **Contrastive loss for sentence representation prediction.** This loss measures the distance between the predicted next-sentence representation from the idea predictor, denoted as $\tilde{i}_{yk}$, and the corresponding target representation from the idea encoder, $\hat{i}_{yk}$. The objective is to ensure that the predicted representation closely aligns with the actual encoded representation. This is achieved using a contrastive loss based on cosine similarity, formulated as:

$$\mathcal{L}^{predict}_{contrastive} = -log\frac{exp(sim(\tilde{i}_{yk}, \tilde{i}_{yk})/\tau)}{\sum_j exp(sim(\tilde{i}_{yk}, \tilde{i}_{yj}/\tau)}$$

  where $sim(a,b) = \frac{a \cdot b}{||a||||b||}$ represents cosine similarity, and τ is a temperature parameter that controls the separation between representations.

- **Cross-entropy loss for token-level reconstruction.** This loss measures the discrepancy between the reconstructed sentence tokens and the original target tokens. The reconstructed token, $\tilde{y}_k(i)$, should closely match the target sentence tokens, $\hat{y}_k(i)$. This loss ensures that the decoded sentence accurately reflects the original input and is computed using the standard cross-entropy formulation:

$$\mathcal{L}_{CE} = -\sum_i \hat{y}_k(i) log P(\tilde{y}_k(i))$$

  where $P(\tilde{y}_k(i))$ represents the predicted probability of the token at position i. By minimizing this loss, the model improves its ability to generate meaningful sentences from the idea representations.

- **Contrastive loss for sentence representation reconstruction.** It measures the alignment between the reconstructed sentence representation $\tilde{i}_{\tilde{y}}$ and the target sentence representation $\hat{i}_{yk}$. This loss ensures that after generating a sentence from an idea, the representation extracted from the reconstructed sentence remains faithful to the original. The formulation is similar to the first contrastive loss:

$$\mathcal{L}^{reconstruct}_{contrastive} = -log\frac{exp(sim(\tilde{i}_{\tilde{y}}, \hat{i}_{yk})/\tau)}{\sum_j exp(sim(\tilde{i}_{\tilde{y}}, \hat{i}_{yj}/\tau)}$$

Finally, the overall loss function is a weighted combination of these three components:

$$\mathcal{L} = \lambda_1 \mathcal{L}^{predict}_{contrastive} + \lambda_2 \mathcal{L}_{CE} + \lambda_3 \mathcal{L}^{reconstruct}_{contrastive}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters that balance the contributions of each loss term. These weights can be adjusted to optimize the model's performance, ensuring that it effectively learns meaningful representations, reconstructs sentences accurately, and maintains consistency between original and generated sentence embeddings.

3 Related Work

The network design is greatly inspired by Assran et al. (2023).

The loss function is from Sohn (2016)

4 Expected Contributions

5 Timeline

References

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.