CS475 HW4: Make Your Own Benchmark

Gaussian Semantic Alignment Benchmark (GSAM)

Team 20

December 14, 2024

1 Introduction

This report introduces the **Gaussian Semantic Alignment Metric (GSAM)**, a new method for evaluating Large Language Models (LLMs) from a structural, representation-level perspective. Unlike conventional benchmarks that rely on task-specific outputs (e.g., accuracy, BLEU scores, human ratings), GSAM evaluates the internal consistency and isotropy of semantic representations within an LLM's hidden states. By focusing on the distributional properties of token-level embeddings, GSAM provides a task-agnostic measure of how well a model organizes meaning internally.

2 What Capability Does GSAM Measure?

The Gaussian Semantic Alignment Metric (GSAM) evaluates the isotropy and semantic coherence of a language model's internal representations. It measures how evenly and effectively the model activates its representational dimensions across various inputs, ensuring the model does not over-rely on a narrow subset of features. A high GSAM score reflects the model's ability to utilize its parameters efficiently, maintaining a balanced, contextually rich semantic space. This metric provides insights into the structural quality of a model's representational system, beyond just output accuracy.

3 Motivation and Importance

Team 20 investigated the semantic space of LLMs using Stacked Autoencoders (SAEs) and found that hidden state vectors resemble a multivariate Gaussian distribution. High likelihood under this model indicates an isotropic and well-structured representation, reflecting efficient parameter utilization. Conversely, low likelihood suggests uneven use of dimensions and underutilization of capacity. The proposed metric highlights the internal structural quality of models, guiding the design of more efficient architectures.

4 Data Collection and Annotation

We constructed a corpus of 300 diverse English sentences drawn from *sst2* in *glue*. The goal was to sample a wide variety of topics and syntactic structures, ensuring that the internal representation space is thoroughly probed. Since GSAM does not require labeled data, no explicit annotation was performed. Our dataset merely consists of raw text, carefully preprocessed for length uniformity (10–30 tokens per sentence). The absence of annotation reduces subjectivity and allows the benchmark to focus solely on distributional characteristics.

5 Evaluation Metrics

5.1 GSAM Metric

We define GSAM as measuring how closely the model's activations follow a single multivariate Gaussian distribution. The process is as follows:

1. Extract token-level hidden states from a chosen layer of the model.
2. Fit a multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$ to the activations.
3. Compute the GSAM score using either KL divergence as the underlying metric.

5.2 KL Divergence-Based GSAM

When the metric is set to `'kl_divergence'`, we first define the KL divergence between the empirical data distribution P_{data} (uniform over N samples) and the fitted Gaussian distribution P_{gaussian} as:

$$\text{KL}(P_{\text{data}}||P_{\text{gaussian}}) = -\log(N) - \frac{1}{N}\sum_{i=1}^{N}\log p(z_i),$$

where $p_{\text{gaussian}}(x_i)$ is the probability density of the i-th embedding x_i under the estimated multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$. Here, N is the number of samples and each $x_i \in \mathbb{R}^d$.

Since we want a score that increases as the embeddings become more Gaussian-like, we take the negative of the KL divergence:

$$\text{GSAM} = -\text{KL}(P_{\text{data}}||P_{\text{gaussian}}) = \log(N) + \frac{1}{N}\sum_{i=1}^{N}\log p(x_i).$$

A higher GSAM in this formulation implies a lower KL divergence, thus indicating a better alignment to a Gaussian distribution.

5.3 Interpretation

The GSAM score ultimately relates to the average log-probability of the embeddings under the fitted Gaussian. The GSAM score reflects how well the distribution of the embeddings approximates the Gaussian distribution, with lower KL divergence values leading to higher GSAM scores. Thus, a larger GSAM value consistently indicates a closer alignment of the model's embedding space to an isotropic Gaussian distribution.

6 Evaluation Results

We extract the GSAM scores of 7 models, *gpt-neo-125M*, *gpt-neo-1.3B*, *gpt-neo-2.7B*, *llama2-7b*, *falcon-7b*, *mpt-7b*, *pythia-7b*. *Falcon-7* model reports the highest score with 24625.55, which means Falcon-7 is the best model fits with Gaussian distribution. On the other hand, *gpt-neo-125M* model reports the lowest score with 2441.50, which is the worst fit model with Gaussian distribution. We compared same model *gpt-neo* with different number of parameters. With larger number of parameters, higher score reported. In general, large number of parameters can learn complex patterns. Our result adds up to the relationship between the size of model parameters and performance.

Model	Metric	GSAM Score
gpt-neo-125M	KL Divergence	2441.50
gpt-neo-1.3B	KL Divergence	9868.11
gpt-neo-2.7B	KL Divergence	12886.67
llama2-7B	KL Divergence	22019.19
falcon-7B	KL Divergence	24625.55
mpt-7B	KL Divergence	21952.11
pythia-7B	KL Divergence	21983.34

Table 1: GSAM Scores for Evaluated Models

7 Discussion

7.1 Model Parameters and GSAM

GSAM measures how well a model's embeddings align with a Gaussian distribution without directly factoring in the number of parameters. It computes an average log-likelihood, ensuring that the number of tokens does not inflate the score. However, larger models, when properly designed and trained, can achieve higher GSAM scores due to their ability to better distribute representations. Conversely, poorly trained large models may still deviate significantly from Gaussian-like embeddings. While GSAM does not explicitly depend on parameters, it reflects the effects of model complexity, size, and training quality, making it a useful tool for evaluating whether a model's resources are effectively utilized to create coherent, isotropic embeddings.

7.2 Conclusion

In this work, we suggest GSAM, a novel metric for evaluating the isotropy and semantic coherence of LLM hidden spaces. Unlike conventional task-based benchmarks, GSAM directly examines the statistical properties of internal embeddings without relying on explicit annotations or tasks. Preliminary experiments suggest that well-trained models exhibit higher GSAM scores, indicating more uniform and efficiently utilized semantic representations. While GSAM is not a direct function of model size, larger models may leverage their additional capacity to produce more isotropic embeddings. Ultimately, GSAM opens up new avenues for understanding and improving the internal structural quality of large-scale language models.

8 Team Contributions

- Park Seung Jun: Ideation, Model inference pipeline, hidden state extraction, GSAM computation, mathematical derives, and report writing.
- Park Hyeong Jun: Data collection, Processing, Writing test, Collect evalution results, and report writing.
- Ahmet Ilhan Balcik: Overall code feedbacks, Participate in discussion.
- Anurag Yadav: Overall report feedbacks, Participate in discussion.