Visualizing Emergent Concepts in Transformer Hidden States

Kyle Lai
University of California, Merced
klai31@ucmerced.edu

Abstract

Large language models are becoming more capable and powerful but the complexity of their internal representations makes interpreting how they store and process information very difficult. In this study, we analyzed the hidden states of the language model DistilGPT2 to better visualize how representations of concepts evolve and emerge across layers. We constructed sets of prompts of various categories and concepts to pass into DistilGPT2 and extracted their corresponding activations and token information at each hidden state layer of the LLM. We applied three dimensionality reduction techniques: Principal Component Analysis (PCA), t-SNE, and UMAP in our experiments so we can visualize the clustering and structure of higher dimensional space but as a three dimensional representation. Our results show a consistent progression towards distinct separable clusters from the early layers to the later ones. We also examined that the different methods we used revealed different degrees of meaning and clarity which shows that the choice of technique is just as important for interpretability as the data itself. Overall, our findings show that while dimensionality reduction techniques are effective and can reveal very meaningful patterns and clusters, it also comes with information loss and an imperfect projection of its original representation. This emphasizes the need for recognizing the limitations and disadvantages of using one technique over another for interpretability research.

1 Introduction

Large Language Models have demonstrated surprising emergent capabilities and a capacity to perform complex natural language tasks. Despite continuous advancements that allow for training smarter and more capable models, we do not yet fully understand how LLMs work at a deeper level. Like humans, the LLM's internal processing of inputs remains a challenge to analyze and interpret. In the field of cognitive science, LLMs would be referred to as "black boxes" because of the current lack of transparency in transforming a prompt into an output. Uncovering how these models process information is important as LLMs are further developed and play an increasingly relevant role in our lives.

Like the firing of neurons in the human brain, the activations of tokens in each layer of an LLM form a major basis for current work on interpretability. High dimensional vectors attached to tokenized inputs are processed through multiple layers. Because of the vector's high dimensionality, a major challenge arises in analyzing such a large volume of data in a meaningful way. Additionally, the subtle intricacies of how meaning and concepts could be distributed across activations, tokens, and layers adds another layer of complexity. In addressing the problem of visualizing higher dimensional data, various techniques have been proposed such as PCA, t-SNE, and UMAP.

PCA, t-SNE, and UMAP are increasingly used techniques in the field of interpretability of LLMs but they are not perfect solutions. The outputs and conclusions of these techniques must be taken with the understanding of the limitations and advantages that each technique brings. While existing research presents findings based on visualizations using these techniques, it is often left unclear what the scope of that data is or whether certain subtle structures and representations might be lost as a drawback of one algorithm.

This research paper sets out to investigate how the three most used dimensionality reduction techniques of PCA, t-SNE, and UMAP transform and develop structures from the hidden states of LLMs. We conducted a series of tests using sets of prompts from different concepts. Then, we applied the three techniques to early, middle and late layers to observe how the prompt clustering evolves as it is processed further in the model. The behavior across

layers and techniques provided insight into how reliable these techniques are in bringing structure and meaning to a much lower dimension.

Our contributions are as follows:

- We evaluated the performance of PCA, t-SNE, and UMAP dimensionality reduction techniques on LLM hidden state representations across sets of prompts and different layers.
- We measured how much information on variance is lost during PCA across every layer and how the number of principal axes can increase the cumulated variance represented.
- We provided an analysis of the early, middle and late layers of DistilGPT2 to study how emergent representations form by using each dimensionality reduction technique.
- We examined the structure and relationships between clusters by varying the prompts to include abstract encompassing concepts as well as more similar concepts.

2 Related Work

2.1 Interpretability of Language Models

Understanding the internal representations of large language models is a large challenge in natural language processing. With the introduction of the groundbreaking paper, "Attention is All You Need," the modern model for creating hidden states was discovered (Vaswani et al., 2017). Now, these hidden states are the focus for work on interpretability.

There is existing interpretability work that has explored how models encode syntax and semantic information. Using probing methods that train classifiers on hidden states, this paper, "A Structural Probe for Finding Syntax in Word Representations" has shown that information can be recognized within the hidden state vectors (Hewitt and Manning, 2019; Tenney et al., 2019). In contrast, this work is focused on revealing that representations in hidden states could be visualized geometrically.

Another similar field in interpretability is mechanistic interpretability which identifies components in transformer models that correspond to behaviors. This approach traces the flow of information through layers. One downside is that it is also often specific to one model's architecture. There are also surveys of interpretability methods that show the diverse range of approaches researchers have taken in understanding language models (Belinkov and Glass, 2019; Rogers et al., 2020).

2.2 Visualization of Hidden Representations

This paper's approach to interpretability involves visualizing hidden states in lower-dimensional spaces. This would not be possible without the work of creating the dimensionality reduction techniques of, principal component analysis (PCA) (Jolliffe and Cadima, 2016), t-SNE (van der Maaten and Hinton, 2008), and UMAP (McInnes et al., 2018).

Prior studies have experimented with using hidden states from transformer models in creating meaningful geometric structures. They noticed that similar inputs often clustered together. Visualization has been used to explore words with multiple meanings, similar contexts, and how representations evolved across layers (Reif et al., 2019). However, these analyses rely heavily on visual interpretation. In particular, the paper, "Visualizing and Measuring the Geometry of BERT," uses PCA to plot BERT's 1024-dimensional hidden states to study the geometric structures and syntax trees. In another paper, "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings," researchers used quantitative analysis to reveal that 95% of the variance of hidden states is because of context rather than the original token embeddings.(Ethayarajh, 2019) This paper builds off this idea by creating sets of prompts with the same context and seeing their geometric representations cluster through hidden states.

2.3 Layer-wise Representation Analysis

As early as 2019, researchers have examined how representations evolve across layers in transformer models. Earlier layers seemed to relate to low-level syntactic information, while the later layers reflected high-level concepts (Tenney et al., 2019; Ethayarajh, 2019). In the paper, "BERT Rediscovers the Classical NLP Pipeline," the researchers showed that BERT's behavior resembled the classical NLP pipeline, which described how language was processed before deep learning.

2.4 Our Position

While current research has relied on dimensionality reduction as a visualization tool, we want to center our focus on how the choice of projection method is a large factor in interpretability. We analyzed

the differences in the structures revealed by PCA, t-SNE, and UMAP on the same data. We provided our analysis of how these differences affect conclusions about clustering.

3 Methodology

3.1 Model and Representation Extraction

We analyzed the hidden state representations from the language model DistilGPT2. When we pass an input prompt into DistilGPT2, the model produces a sequence of hidden states for each of its 6 different transformer layers. Each token is represented as a high-dimensional vector of activations. The amount of information per layer could be represented with a matrix of variable size n×m, where n is the number of tokens in the input sequence and m is the number of dimensions. To obtain a representation for each prompt and any given layer, we chose the last token within the layer because of its access to every preceding token before it. DistilGPT2 is an autoregressive model which means it processes tokens one at a time and bases the new token on all preceding tokens. Since the last token attends to the entire sequence at that layer as well as everything else in previous layers, it serves as a summary of each layer with the advantage of being more compact and standardized throughout our experiments. Now the vectors are fixed to the dimensionality of the LLM used. In our case, DistilGPT2 has 768 dimensional activations for each token.

3.2 Preprocessing

Before proceeding further, the variation in each vector may be more pronounced in some dimensions than in others. This poses the challenge because we don't want smaller dimensions to be overshadowed by others that might have naturally high mean values. Therefore we must both normalize the mean and z score. Now, each dimension has zero mean and standardized variance. This step equalizes scale differences across features and improves the stability of our final projections.

3.3 Dimensionality Reduction

We applied three dimensionality reduction techniques:

- Principal Component Analysis (PCA)
- t-SNE
- UMAP

PCA is a linear method. It first identifies an axis in the high dimensional space with all the prompts that captures the greatest variance. The computed axis forms the first principal component. PCA then can compute a new set of axes orthogonal to the first and project all the data points to the space created by these principal components. This results in a lower dimensional space where projected data points preserve as much global variance from the original space as possible.

Unlike PCA, t-SNE is a nonlinear method. It focuses on preserving local relationships. It begins by measuring the distances to every other point in the high-dimensional space. These distances form a basis for how it positions points in a lower dimensional space. A perplexity score determines the model's emphasis on how many neighbors to consider. T-SNE then places points in the low-dimensional space such that nearby points remain close together, while distant points are pushed apart. The result is good local clusters and separation but this may distort the global structure.

Like t-SNE, UMAP is another nonlinear method that forms a graph using the local relationships between points as a basis. Unlike t-SNE, UMAP also tries to retain the global structure by taking into account the connectivity of clusters. This produces a nice result that reflects both local clustering and broad relationships between clusters.

4 Experiments

4.1 Model Selection

Our experiments could be conducted on any LLM. DistilGPT2 was used as the base model for all experiments because it was most suited to our experiment's needs. Its small number of layers can capture a larger number of processes that would have spanned many more layers in larger models like GPT-3 which has up to 93 layers. We chose to analyze layer 1 for the early layers, layer 3 for the middle layers, and layer 5 for the late layers.

4.2 Prompt Construction

We constructed sets of prompts corresponding to four conceptual categories. We created four classes: dog, cat, firetruck, and vehicle. Each class contains 26 prompts with varied phrasing but maintains the semantic meaning of its class. Some prompts use different vocabulary and don't necessarily say the word dog or cat explicitly like in the following prompt: "A feline watched from the window." We

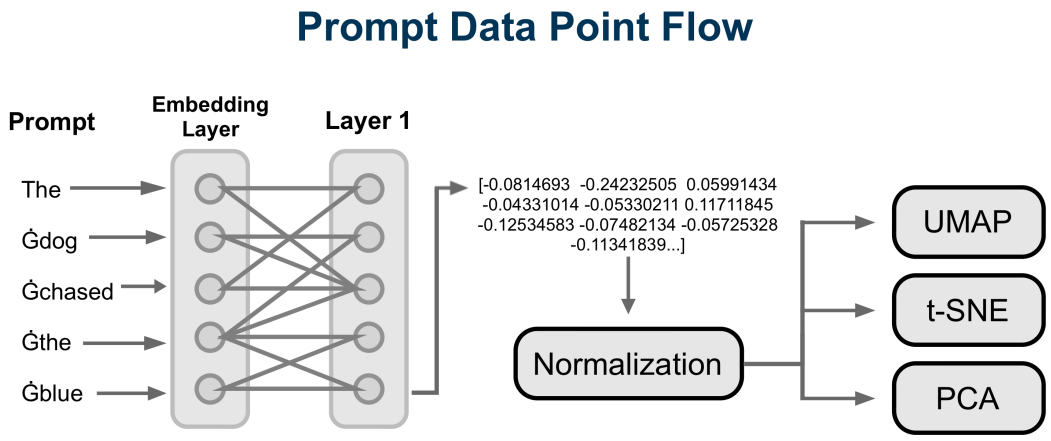

Figure 1: Data flow from prompt to visualization.

chose the classes, dog and cat because of their similarity and associativity as common household pets. We chose firetruck and vehicle for similar reasons but also to test if there will be a visual indication that firetruck is a vehicle since the vehicle class should encompass firetrucks as well as many other vehicles. This setup allows us to examine multiple points of interest as the clustering behavior can reveal insight into any one of these subtleties.

4.3 Dimensionality Reduction Settings

For PCA, no additional parameters need to be set as this technique as explained in the methodology section projects data onto the components that capture maximum variance. For t-SNE, we used a perplexity value of 7. For UMAP we set the number of neighbors considered to 15 which is larger than our t-SNE value and should be able to capture global structure better. We also set the minimum distance on the diagram to 0.1. For each layer and method, we evaluated the projected representations and analyze the behaviors of clusters across each prompt category.

4.4 Results

Across all methods of PCA, t-SNE, and UMAP, we observed a consistent progression of the data points which began as unstructured in early layers to become increasingly organized in later layers. Besides this, the clarity and geometry of the clusters still varied significantly by method.

PCA As shown in Figure 2, the structure that is produced with PCA is spread out across all of its layers. In its early layers, we can observe minimal clustering with many prompts overlapping. Since layer 1 is directly based off of the embedding layer, it would be most affected by positional and word-based information. Still, animal and cat prompts are directionally distinct with dog and cat prompts on the left side of the graph and firetruck and vehicle on the right. This transition is still gradual and includes enough overlap to be hard to separate.

In the middle layers represented by layer three, this separation becomes slightly more apparent with a small gap forming between the two pairs of prompts. Also, the combination of dog and cat prompts start forming two clusters yet are almost entirely intermixed. By the later layers, this pattern is maintained with three overall clusters present. The subsections that emerged from cat and dog prompts could suggest an additional factor beyond our categorizations of dog and cat. It is also notable that the variance captured in each of these three graphs varies from 35.3% in layer 3 to 46.4% in layer 5. This number signifies that half of the information on variance is lost which could have a large influence on the true structure of these concepts.

Variance in PCA To provide a quantitative analysis of information loss of the linear projection method, PCA, we analyzed the cumulative variance of PCA based on the number of principal components (Figure 3). We analyzed every layer

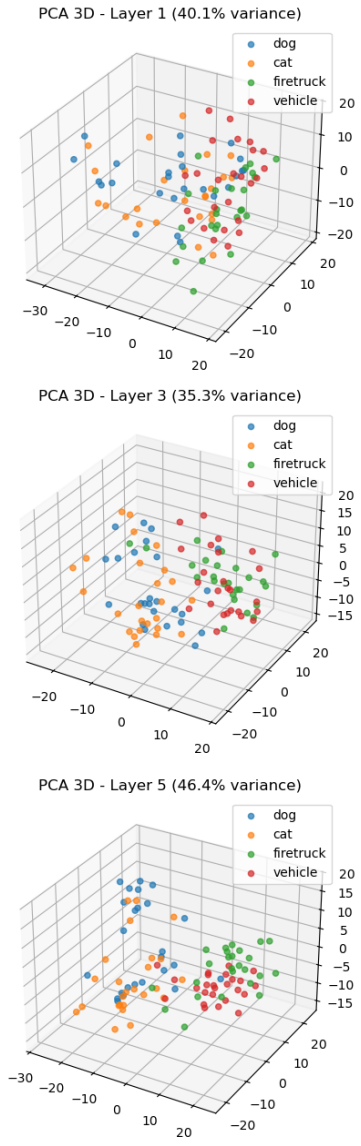

Figure 2: PCA projection of hidden states showing clustering structure across prompt categories.

and their corresponding cumulations. Layer 0 is also included but it only includes information about embeddings and position which explains its deviation from other layers as it is the only unprocessed layer and is thus easier to represent. Less than 7 principal axes are necessary for projecting layer 0's variance information. With the first three axes, 35-46% of the projected variance is present in our graph like shown in Figure 2. These results show that with just a few axes out of 768 dimensions PCA is able to capture much information when projected. But also, there is a significant portion of original information that is discarded. Even when you increase the number of principal axes used to 40, around 10% of variance is left unaccounted for

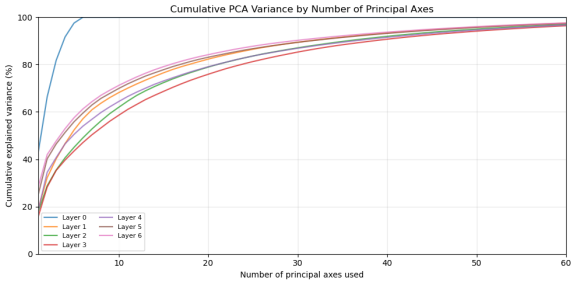

Figure 3: Cumulative PCA Variance by Number of Principal Axes.

which shows the limitation of linear techniques like PCA.

t-SNE Figure 4 shows a very gradual transition from a mixed structure to a separated representation. In the early layers, the points are a uniform, spherical shape where no clear patterns between prompts could be seen. In the middle layers, the prompts begin to shift directionally with dog and cat prompts shifting to the lower right and firetruck and vehicle prompts shifting to the top left.

In the late layers, represented by layer 5, the separation which started in the middle layers is even more pronounced. There are two regions: one for cat and dog prompts and another for firetruck and vehicle prompts. The structure resembles two flat disks. Despite the separation of these very different prompt sets, like with PCA, cat and dog prompts remain mixed with no discernable patterns emerging. The uniformness and even spacing of the prompts remains consistent through every layer which is likely influenced by the low perplexity setting of 7. With a low perplexity, local neighbors are prioritized over global distances which could potentially reshape and normalize the original representation's spacing.

UMAP The results for UMAP were the most visually significant. All layers revealed strong clustering patterns. Unlike the visuals shown by PCA and t-SNE, UMAP already showed a directional separation between the two pairs or prompts in layer 1. A vertical structure formed with cat and dog related prompts towards the top and firetruck and vehicle related prompts towards the bottom. In this stage, they are still not distinct clusters as there still remains a gradual transition.

In the middle layers, the shape of the structure transforms with the two animal-related prompts elongating into a capital J shape and the two other prompt categories forming a flower shape with its

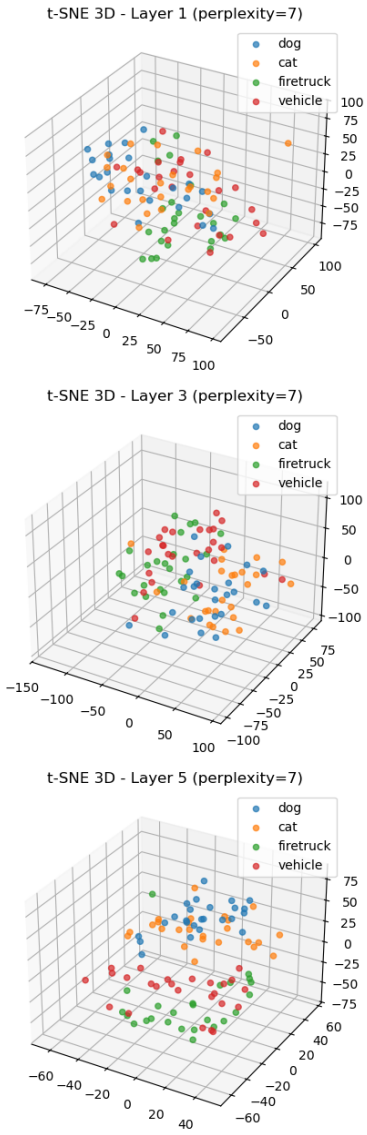

Figure 4: t-SNE projection of hidden states showing clustering structure across prompt categories.

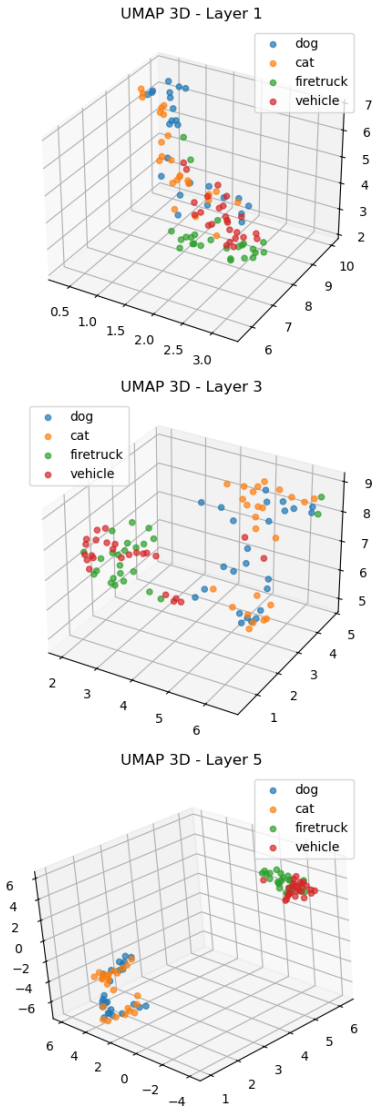

Figure 5: UMAP projection of hidden states showing clustering structure across prompt categories.

stem joining with the tip of the J shape. In the late layers, the clustering is much more compact. A very clear boundary has formed between the two opposing categories. The cluster of cat and dog prompts remain mixed but its structure has transformed into a U shape. The cluster for vehicle and firetruck has a local separation as well with firetruck on the left side and vehicle on the right. Overall, UMAP produced very visually separable clusters under our experimental conditions.

4.5 Analysis

The results of our experiments indicate that there is a clear progression towards structure across the layers. Early layers show an overlap with other prompts and are spread out which suggests that they are most influenced by positional and lexical encoding. After propagation through transformer layers, information on the category of these prompts is increasingly present and structured and meaningful organization emerges. In the final layers, the representations reflect the higher level abstractions that originally guided the creation of the prompts and produce clear distinct separations.

The proximity of dog and cat prompts no matter the layer or technique suggests that the model captures semantic features and information that encapsulates both closely-related concepts. At the same time, the consistent separation of animal related and vehicle related prompts in later layers

reflects a high level distinction for broadly different concepts. The appearance of firetrucks being in a separate subcluster than vehicle indicates that even though a vehicle should encompass firetrucks, general and specific concepts may be represented in the same manner in the LLM.

5 Conclusion

Overall, we were able to investigate the general structure of internal representations of the language model DistilGPT2 by using dimension reduction techniques. We followed the structure of prompt categories and how they changed across multiple layers.

Our results indicate that the later layers encode more abstract semantic information. We also found that the choice of technique used for dimensionality reduction will have a significant impact on the resulting structure. Using a nonlinear method like UMAP or t-SNE will produce visualizations with clearer clustering. Using a linear method like PCA has the downside of potentially missing important non-linear relationship information.

Future work can build upon our analysis by introducing a larger and more diverse dataset. Conducting these tests with different models and using more measurable methods to analyze the resulting visuals could further the reliability of the techniques we focused on in this study. Hidden state representations still remain as a topic LLM researchers have yet to fully explore and our research on existing interpretability tools is an important step towards a deeper understanding of LLMs.

6 Limitations

The first limitation is that our study is not as quantitative. When assessing the clustering and behavior of t-SNE, and UMAP visualizations, visual inspection is our primary basis for analysis and not any quantitative metrics. This inevitably limits our ability to create comparisons between methods or develop concrete metrics on how much of the observed patterns reflects the original variance like we can with PCA methods.

The second limitation is that our experiments were conducted through a single language model, DistilGPT2 for consistency and simplicity. This model is much smaller than the modern large language models which have demonstrated even more powerful reasoning abilities. It is hard to say if our findings are generalizable to a larger architecture model with many more layers and dimensions.

The third limitation is that our prompt dataset is limited in both size and scope. Our manually constructed examples may not fully represent the full structure for a concept and our limited categories are not enough to display a large diversity of the natural language or relationships that is stored in LLMs.

Finally, our decision to use the final token's hidden state for each layer only provides one perspective of a position in the sequence of tokens. Other representations like mean pooling may reveal an entirely new insight into the structure of hidden states.

References

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL*.

Ian T. Jolliffe and Jorge Cadima. 2016. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.