Interpretation of Multimodal LLMs Trained for Medical Domain

Project Description

The objective of this project is to examine **pretrained** multimodal **LLAVA-Med-v1.5-mistral-7b** to enhance its interpretability and potential optimization. The project focuses on understanding internal representations such as **linearity**, **contextualization**, **anisotropy**, and **intrinsic dimensions** within the model. The project also includes studying **multimodal neurons** that integrate and process information from medical images and text, improving clinical decision-making reliability.

Most metric calculations were conducted using the dataset provided by the model's creators for evaluation purposes. To compute the metrics, we utilized various model features, including embeddings, attention maps, and specific layer weights. For contextualization, 50 distinct images paired with a single question were employed. For Grad-Cam we used a special prompt with images to get gradients and attention maps from meaningful token's prediction.

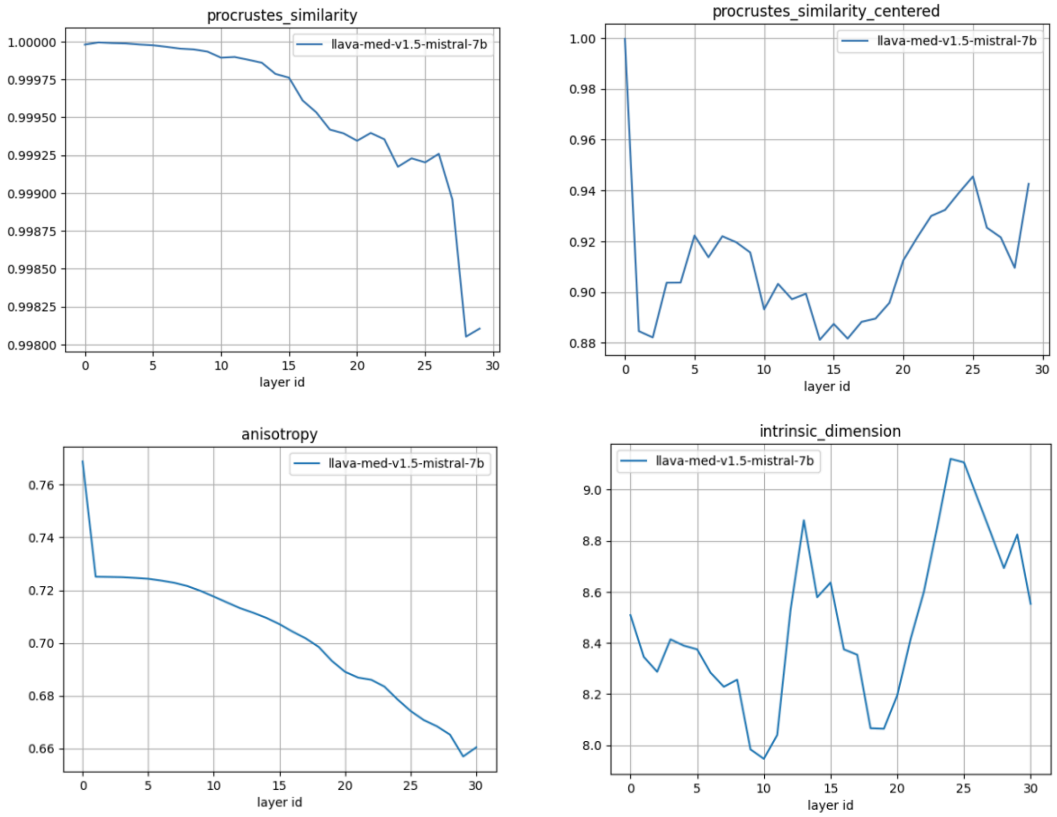
Results

1. Embedding Space and Properties Analysis

- **Procrustes Similarity**: The **Procrustes similarity** metric shows a gradual decline as layer depth increases. Although the linear relationship remains relatively high, the decrease towards the deeper layers suggests that embeddings become less linearly dependent on the original space, indicating increased complexity in the contextualization process.
- **Procrustes Similarity (Centered)**: A peak is observed at layer 25 in the **centered Procrustes similarity**. This suggests that the model accumulates the most information at this layer and slows down changes. The model's embeddings seem to capture the greatest amount of context on this layer before becoming more compressed in subsequent layers.
- **Intrinsic Dimension**: The **intrinsic dimension** metric mirrors the peak at layer 25 seen in the centered Procrustes similarity, reinforcing the idea that this layer captures the highest level of information without overfitting. After this peak, the dimension decreases, indicating that the embeddings become more compressed, possibly to optimize the prediction process.
- **Anisotropy**: The **anisotropy** metric steadily declines with layer depth, suggesting that the embedding space becomes more isotropic. This implies that

the model's embeddings are distributed more evenly, which can be interpreted as the model making better use of the available representational space on deeper layers.

$$\text{linearity\_score} := 1 - \min_{A \in R^{d \times d}} ||\tilde{X}A - \tilde{Y}||_2^2$$



2. Analysis of Image vs. Text Embeddings

The graphs reveal that the embedding spaces for visual and text tokens evolve differently, which is expected given that the LLM part was pre-trained primarily on text data and only exposed to visual embeddings during fine-tuning.

- **Procrustes Similarity**: More significant changes are observed in the visual embeddings compared to the text embeddings. This can be attributed to the model dynamically learning to process visual data, while the text embeddings remain more structurally stable due to the model's extensive pre-training on text.
- **Intrinsic Dimension**: The intrinsic dimension for visual embeddings is consistently lower than for text embeddings. This suggests that the model captures more information in the text token space, which aligns with the model's original design as a language model. The intrinsic dimension behavior for text

tokens reflects the standard trend observed in language models, as discussed in related research. These models typically display a rise in dimensionality followed by stabilization, indicating efficient encoding of information at deeper layers.

- **Anisotropy**: The anisotropy of visual embeddings decreases with layer depth, suggesting that the model makes better use of the representational space for visual data at deeper layers. In contrast, the anisotropy for text embeddings remains relatively stable and significantly lower, further confirming that the model is already well-optimized for text processing.

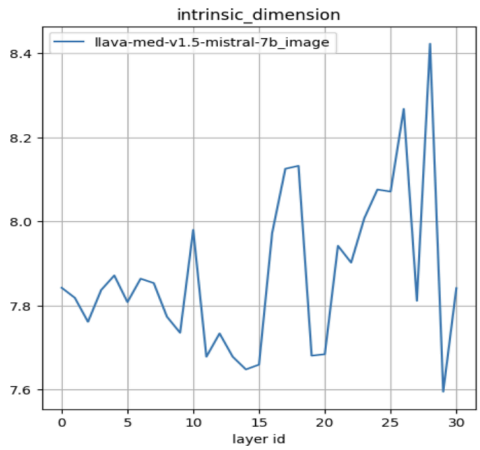
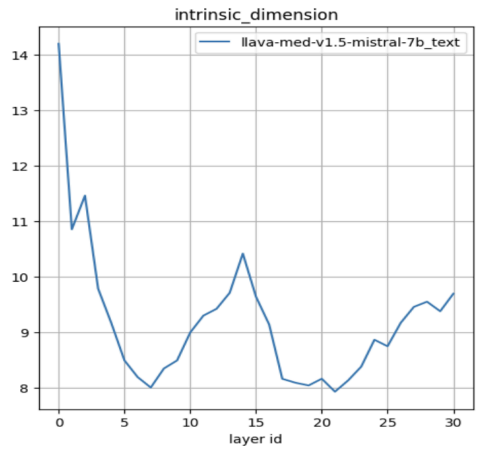


Image Tokens



Text Tokens

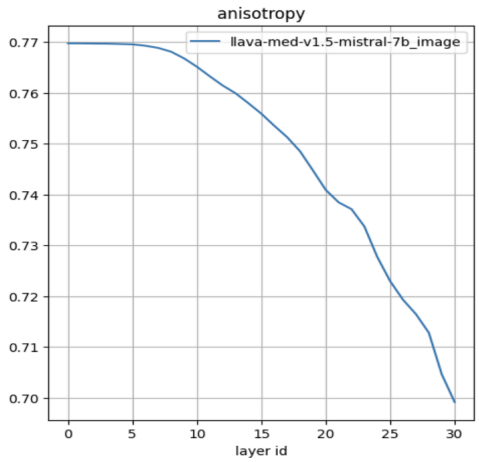
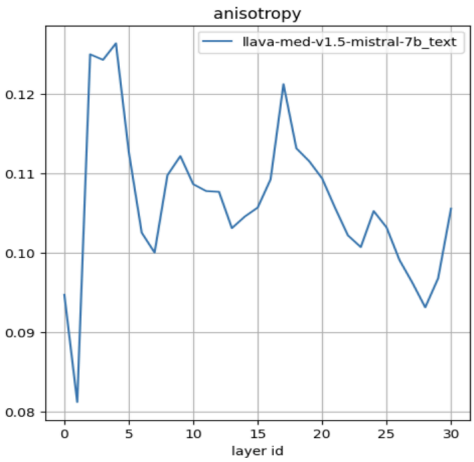


Image Tokens

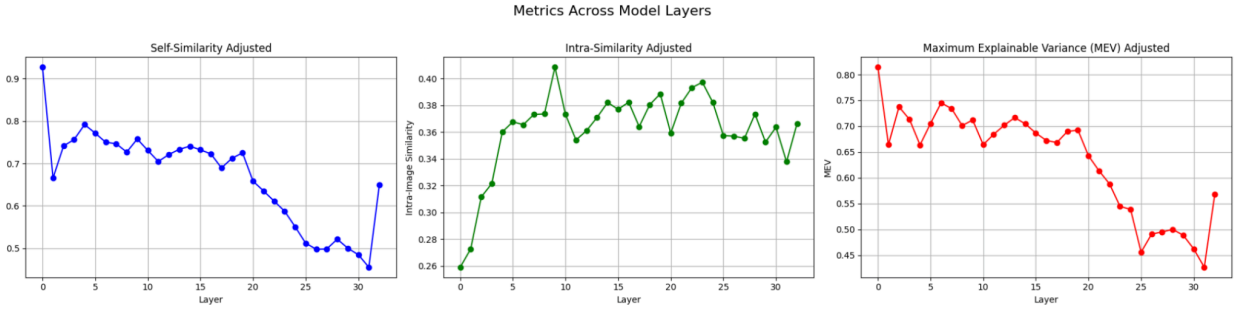


Text Tokens

4. Contextualization Analysis

- To assess contextualization, three metrics were calculated for the text tokens at each model layer: **self-similarity**, **intra-similarity**, and **MEV**. To ensure accurate evaluation, anisotropy was computed at each layer using randomly sampled pairs of embeddings from the test set, accounting for the average similarity of embeddings across layers. Additionally, MEV was calculated for randomly selected pairs to adjust the metrics, providing a more reliable interpretation of the contextualization process. These computed values were then subtracted from the respective metrics to adjust for baseline similarity, providing a more reliable interpretation of the contextualization process.

The first graph shows a gradual decrease in **self-similarity** with increasing layer depth, indicating reduced similarity of the same token's embeddings across different visual contexts. The second graph illustrates an increase in **intra-similarity**, reflecting greater internal consistency of embeddings within a sentence as depth increases. The third graph reveals a decline in the proportion of variance explained by the first principal component (**MEV**) as the layer depth grows, highlighting the increasing complexity of the contextualized embedding space.



$$SelfSim_\ell(w) = \frac{1}{n^2 - n} \sum_j \sum_{k \neq j} \cos(f_\ell(s_j, i_j), f_\ell(s_k, i_k))$$

$$IntraSim_\ell(s) = \frac{1}{n} \sum_i \cos(\vec{s}_\ell, f_\ell(s, i))$$

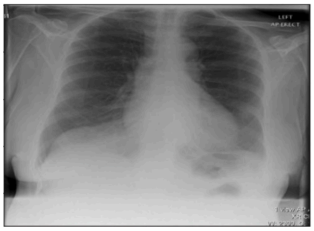
$$\text{where } \vec{s}_\ell = \frac{1}{n} \sum_i f_\ell(s, i)$$

$$MEV_\ell(w) = \frac{\sigma_1^2}{\sum_i \sigma_i^2}$$

5. Image Tokens Occurrence

- The top ten tokens by occurrence were evaluated across the 5th, 10th, 15th, 20th, 25th, and 30th layers. Notably, in the 5th layer, the most frequently occurring tokens appeared to be domain-specific, often related to medical terminology. However, as the analysis progressed through subsequent layers, token representation became increasingly generalized, which may indicate the model's shift toward higher-level feature abstraction. Additionally, a significant amount of recurring tokens was observed, suggesting the model may be developing an internal approach for image interpretation, potentially refining representations through repeated pattern recognition.

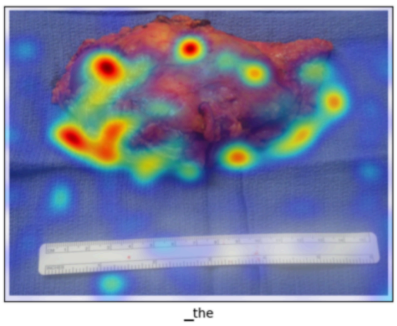
Images

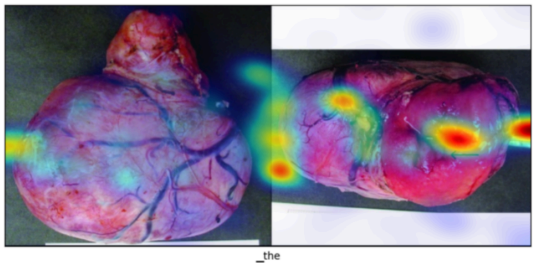


5th Layer	30th Layer		5th Layer	30th Layer
most_frequent_words	most_frequent_words		most_frequent_words	most_frequent_words
[('_kennis', 440), ('_/******/', 410), ('_‘', 233), ('_ingår', 170), ('_[...]', 133), ('_sep', 66), ('_scan', 35), ('_database', 26), ('_FBI', 25), ('_cards', 24)]	[('_/******/', 1071), ('_', 302), ('_bekan', 292), ('_Fuß', 216), ('_kennis', 170), ('_Hug', 168), ('_question', 154), ('_PARTICULAR', 105) ('_the', 90), ('_hint', 79)]	Top 10 Tokens by Occurrence	[('_kennis', 522), ('_/******/', 337), ('_ingår', 296), ('_‘', 153), ('_sep', 122), ('_[...]', 116), ('_therefore', 93), ('_surgery', 70), ('_Nah', 55), ('_patients', 48)]	[('_/******/', 1012) ('_bekan', 255), ('_Fuß', 224), ('_', 216), ('_kennis', 165), ('_Hug', 145), ('_IC', 115), ('_question', 110), ('_PARTICULAR', 73) ('_Pred', 66)]

6. Grad-CAM Interpretation

- Grad-CAM analysis was performed on the 15th layer, selected as an intermediate layer where the model is expected to begin recognizing general patterns. Statistical indicators suggested potential significance at this layer. To validate our hypothesis, we experimented with various layers, confirming that the 15th layer offered valuable insights.





Conclusion

- **Linearity Scores:** The model retains a relatively high level of linearity across layers, confirming that its embedding transformations are consistent. Image embeddings undergo more complex transformations compared to text embeddings.

- **Anisotropy and Intrinsic Dimension:** The model effectively reduces the embedding space to a lower-dimensional form, with stable intrinsic dimensionality across layers, indicating efficient information compression.
- **Contextualization:** The model's ability to adapt embeddings based on visual context improves with depth, particularly in the way it processes complex medical data.
- **Layer-wise token analysis:** Early layers capture domain-specific tokens, while later layers generalize them, suggesting a hierarchical feature learning process that enhances the model's interpretability and abstraction capabilities.
- **Grad-CAM Visualizations**: Grad-CAM analysis shows that the model consistently focuses on critical areas, such as regions of pathology, indicating effective learning of relevant visual cues for medical image interpretation.
- The approaches used in this study provide a good starting point for interpreting multimodal LLMs in medical applications, but deeper research into multimodal neurons and embedding dynamics is recommended for improved transparency in clinical settings.