Towards Understanding How Knowledge Evolves in Large Vision-Language Models
— Supplementary Material —

Sudong Wang[1,2], Yunjian Zhang[3]*, Yao Zhu[3]*, Jianing Li[3]
Zizhe Wang[3], Yanwei Liu[1], Xiangyang Ji[3]
[1]Institute of Information Engeering, Chinese Academic of Sciences;
[2]Nanyang Technological University; [3]Tsinghua University
SWANG049@e.ntu.edu.sg, sdtczyj@gmail.com, ee_zhuy@zju.edu.cn

1. Settings

In our main article, we use the LLaVA-1.5-7b model [2] to analyze the evolution patterns of multimodal knowledge in Large Vision-Language Models (LVLMs) and observe various phenomena from the token probabilities level, probability distributions level, and feature encoding level. To validate the generality of these observations, we conduct additional experiments with the LLaVA-1.5-13b model [2]. This model operates similarly to the 7b version: multimodal inputs are first processed to extract features, then the image-text features are aligned through a linear projection, and finally, the aligned multimodal features are fed into a 40-layer Vicuna model [1] to generate the texts.

2. Token Probability Analyses

The prompt we use is "Please describe this image in detail", and early exit is applied to compute the probabilities for each token across different layers. An example with the probabilities of 24 tokens is shown in Figure 1. It can be observed that in the shallow layers, the probabilities of all tokens are very low, close to zero. Around the 20-th layer, the probabilities of some tokens suddenly increase, giving them a probabilistic advantage. The number of such tokens is extremely small, which means only a few during each token prediction process, while the probabilities of the remaining tokens consistently remain close to zero. Additionally, after gaining a probabilistic advantage, some tokens undergo another abrupt change in their probabilities. This indicates the emergence of new knowledge, altering the prediction probabilities of tokens. When this change is significant enough, it can lead to the creation of new tokens with probabilistic advantage, thereby influencing the output. Figure 2 presents the statistical results of the critical layers across different models, showing that while distribution of critical layer varies slightly between models, it is primarily concentrated in the 16th-19th layers.

3. Token Distribution Analyses

The probability distributions of four examples from the AMBER dataset [3] are shown in Figure 3 and Figure 4. **#1:** A noticeable shift in the JS divergence values emerges at approximately the 16-th layer. The early layers exhibit high divergence values, while beyond the 16-th layer, these values drop sharply to nearly zero and stay consistent up to the output layer. This indicates that before the 16-th layer, knowledge is in a phase of rapid evolution. After the 16-th layer, the knowledge derived from the input stabilizes, and the rate of evolution slows down. **#2:** During the next-token prediction process, the JSDs often undergo abrupt changes in the deeper layers, indicating that new information is being injected into the network, triggering a secondary evolution of knowledge within the model.

Figure 5 and Figure 6 show some experimental results of the skip connection setting: jumping from the critical layers to the mutation layers (**skip.1**), skipping only the mutation layers (**skip.2**), and jumping directly from the critical layers to the last five layers (**skip.3**). In the **skip.1** setting, the model's output is semantically similar to the original, whether it involves correct or hallucinated descriptions, indicating that knowledge evolves slowly during the stabilization stage. In the **skip.2** setting, the model not only preserves most of the original semantics but also corrects some hallucinations, suggesting that the information injected at the mutation layers is likely external and does not originate from the multimodal input. By skipping the mutation layers, knowledge from the shallow layers can be directly passed to the deeper layers, thereby ensuring faithfulness to the input. In the **skip.3** setting, the model's output quality deteriorates significantly, demonstrating the persistence of knowledge evolution in LVLMs, albeit at a relatively slow pace after the critical layers.

The image features two cows standing in a grassy field, enjoying the sunny day. One cow is located towards the left side of the field, while the other cow is located more towards the right side. Both cows are facing the camera, appearing to be curious about their surroundings.

The field is lush and green, providing a perfect environment for the cows to graze and relax. The cows are standing close to each other, possibly indicating a sense of companionship or social interaction between them.

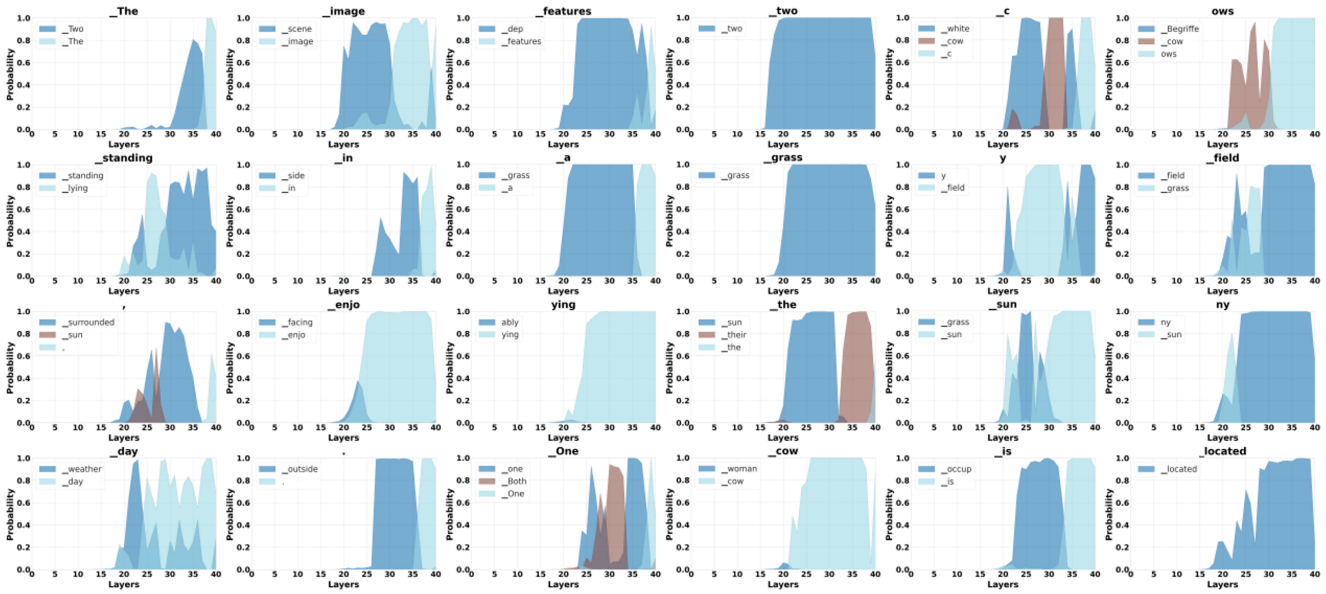


Figure 1. The probabilities of tokens across different layers during normal inference processes.

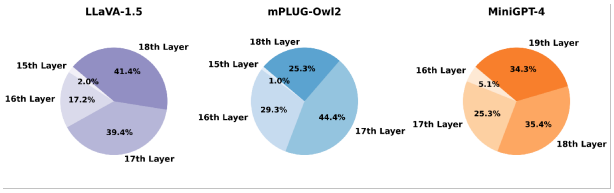


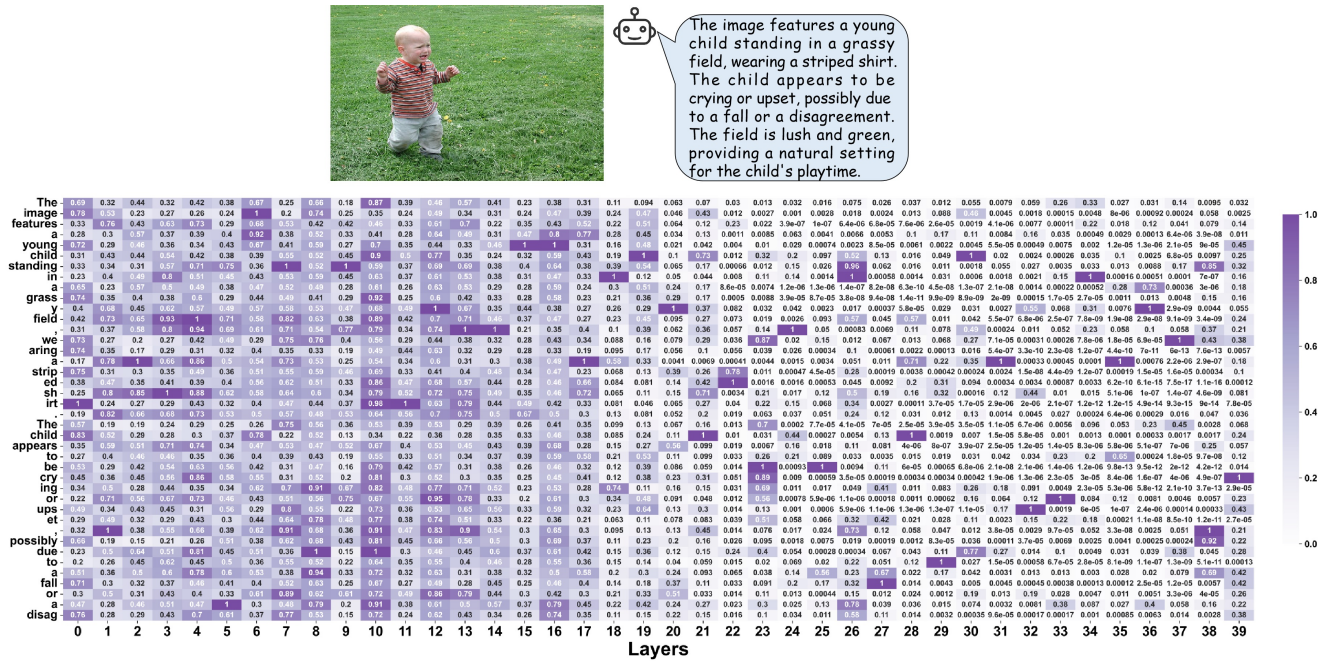
Figure 2. Critical layers distribution across various models.

4. Feature Encoding Analyses

We use t-SNE to compress the high-dimensional feature encodings, and some results on the single images for different tokens are shown in Figure 7. We observe that the feature encodings of different tokens tend to converge in the initial layers and gradually diverge in the deeper layers. This suggests that the knowledge learned in the initial layers is similar across tokens, while as the layers deepen, the knowledge evolves to exhibit token-specific characteristics. Furthermore, for the same token, its features across different layers diverge in an approximately linear manner, indicating a degree of continuity in the knowledge evolution. However, this continuity may be disrupted by injected information, as evidenced by the separation between shallow and deep features for some tokens. We then conduct an analysis on different images in a question-answering task, where the model is prompted to output the key objects in the image. We observe that in the shallow layers, the features of all images are similar, indicating that the knowledge learned by the model for different inputs is general. However, after the critical layers, the features gradually diverge, resulting image-specific characteristics.

References

[1] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%*



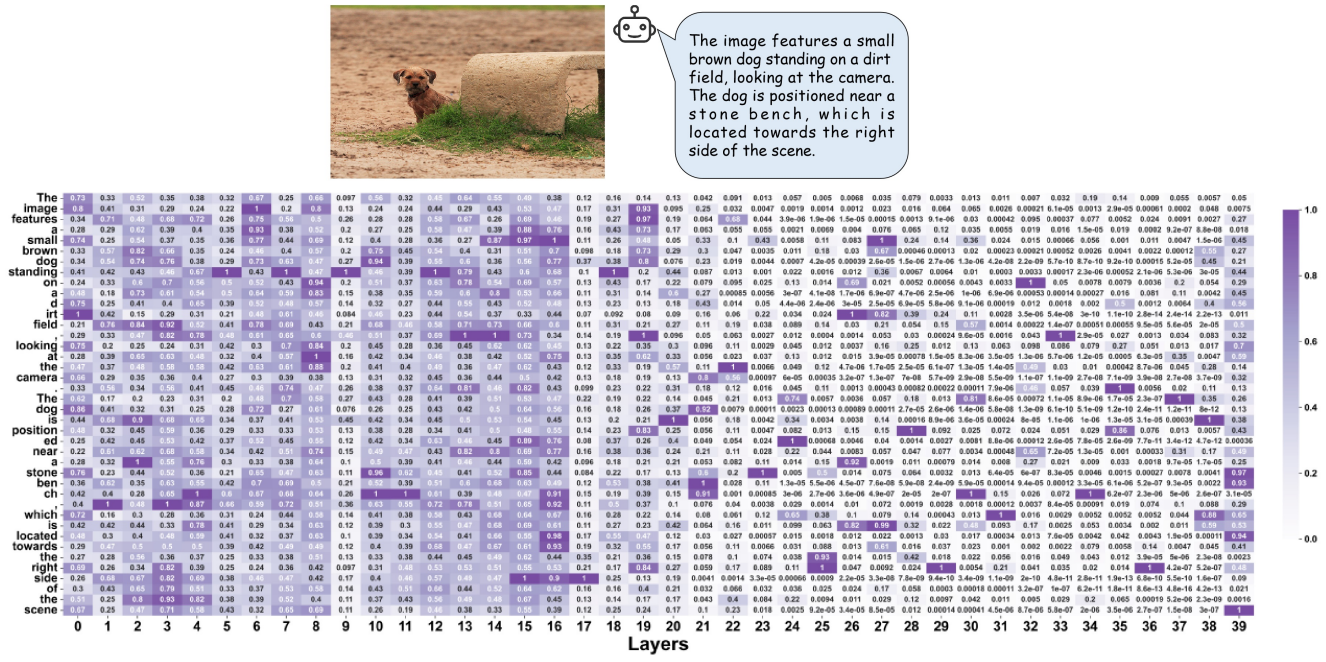


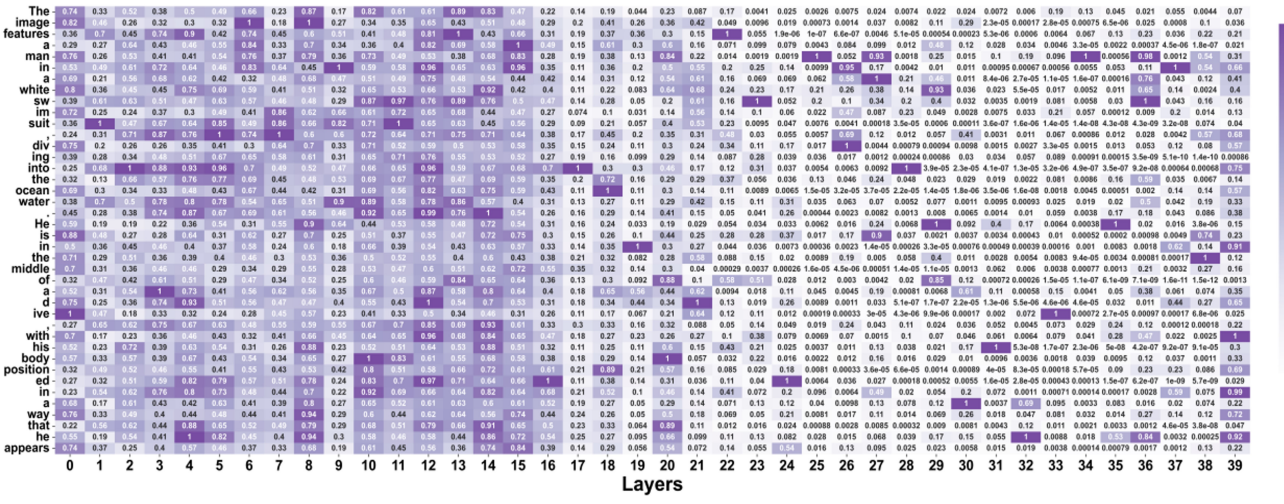
Figure 3. The JS divergences of token probability distributions across adjacent layers during normal inference processes.

chatgpt quality, 2023.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[3] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

The image features a man in a white swimsuit, diving into the ocean water. He is in the middle of a dive, with his body positioned in a way that he appears to be upside down. The man is wearing swim trunks, and his feet are visible as he makes the dive.



The image features a man riding a horse across a vast, open field with a mountain range in the background. The man is positioned towards the left side of the scene, while the horse is located closer to the center. The field appears to be dry and grassy.

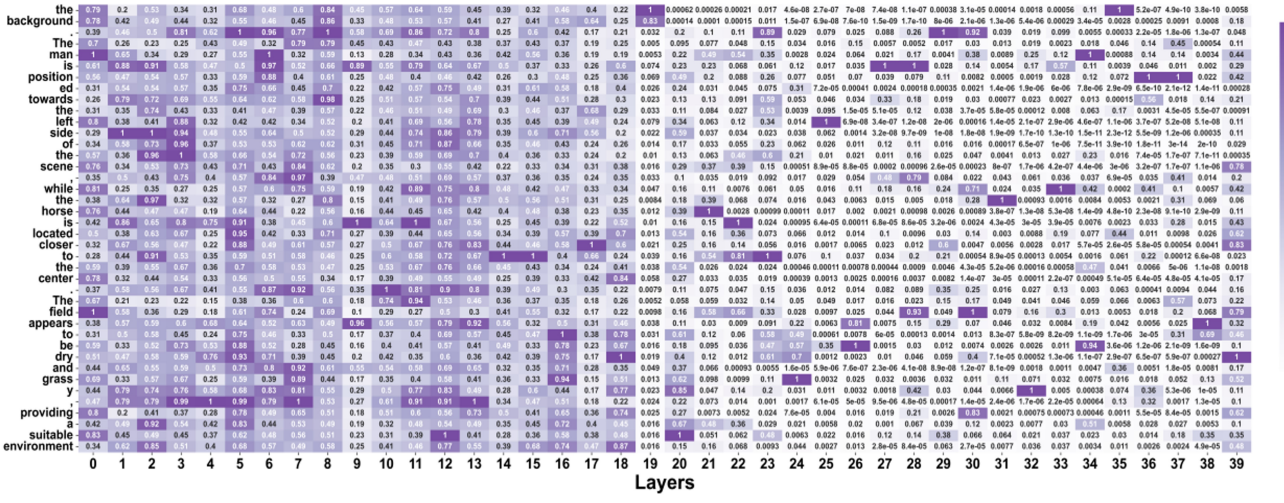
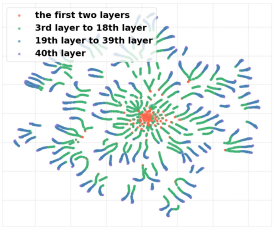


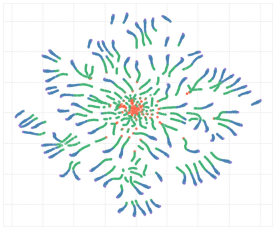
Figure 4. The JS divergences of token probability distributions across adjacent layers during normal inference processes.

Figure 5. The effect of skip connections on model's output. From left to right: the original image, the descriptions from the original model, the descriptions when skipping from the critical layers to the mutation layers, the descriptions when only skipping the mutation layers, and the descriptions when skipping from the critical layers to the last few layer (as the layers near the output contain linguistic priors, we retain the final 5 layers). Hallucinated tokens are marked in red, and corrected tokens are marked in green.

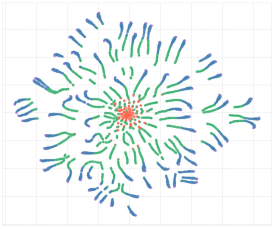
Figure 6. The effect of skip connections on model's output. From left to right: the original image, the descriptions from the original model, the descriptions when skipping from the critical layers to the mutation layers, the descriptions when only skipping the mutation layers, and the descriptions when skipping from the critical layers to the last few layer (as the layers near the output contain linguistic priors, we retain the final 5 layers). Hallucinated tokens are marked in red, and corrected tokens are marked in green.



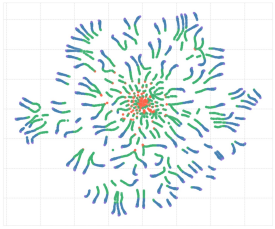
The image features a man riding a bicycle down a curvy road. He is wearing a helmet for safety while enjoying his ride. The road is surrounded by a lush green hillside, creating a picturesque scene.



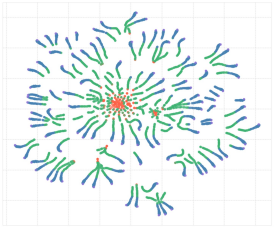
The image features a young child standing in a grassy field, wearing a red hat. The child appears to be pointing at a bird, possibly a goose, that is walking in the grass nearby. The bird is located towards the right side of the image, closer to the center.



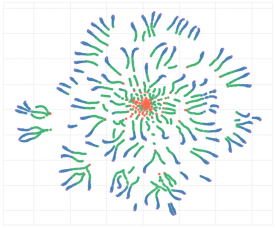
The image features a young boy lying on the grass, playing with a football. The boy appears to be enjoying his time outdoors, possibly taking a break from playing or just relaxing. The scene captures a moment of leisure and fun for the child.



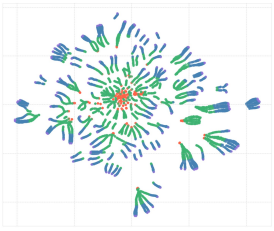
The image captures a beautiful sunset scene at a golf course, with two men standing on the green. The men are positioned close to each other, with one slightly to the left and the other to the right.



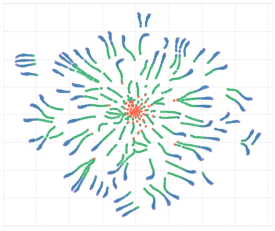
The image features a small boat resting on a rocky beach near the ocean. The boat is positioned towards the left side of the scene, with the ocean stretching out in the background. The beach is covered in rocks, giving it a rugged appearance.



The image features a grassy field with two zebras standing and laying down. One zebra is standing on the left side of the field, while the other is laying down on the right side. They are positioned close to each other, possibly enjoying each other's company.

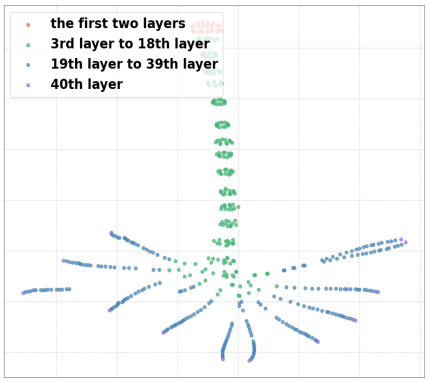


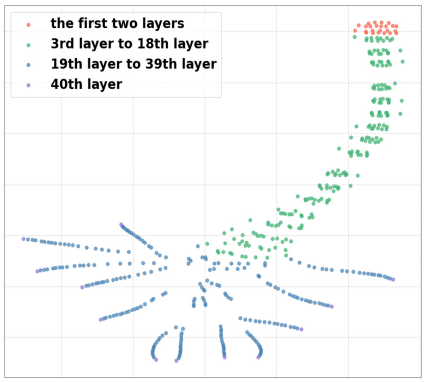
The image features a kitchen counter with a microwave placed on top of a shelf. The microwave is positioned in the center of the counter, and a coffee cup is located to the right of the microwave. There are also two coffee mugs on the counter, one near the left edge and the other near the right edge.



The image features a dog standing on a rocky beach near the water. The dog is positioned towards the center of the scene, with the ocean in the background. The beach appears to be wet, possibly due to the tide or waves.

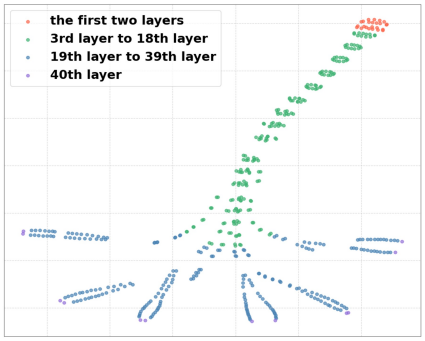
Figure 7. Feature encodings across layers in the same image for different tokens.

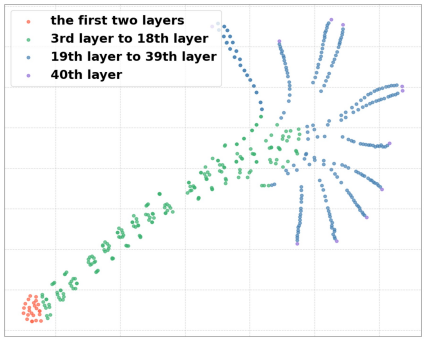




What is the person doing in the image?

What is the vegetable shown in the image?





What is the musical instrument shown in the image?

What is the fruit shown in the image?

Figure 8. Features encodings across layers for different images.