ANALYZING MEMORIZATION IN THE ENCODER OF A VISUAL AUTOREGRESSIVE MODEL

Mohammad Jafari
Sharif University of Technology

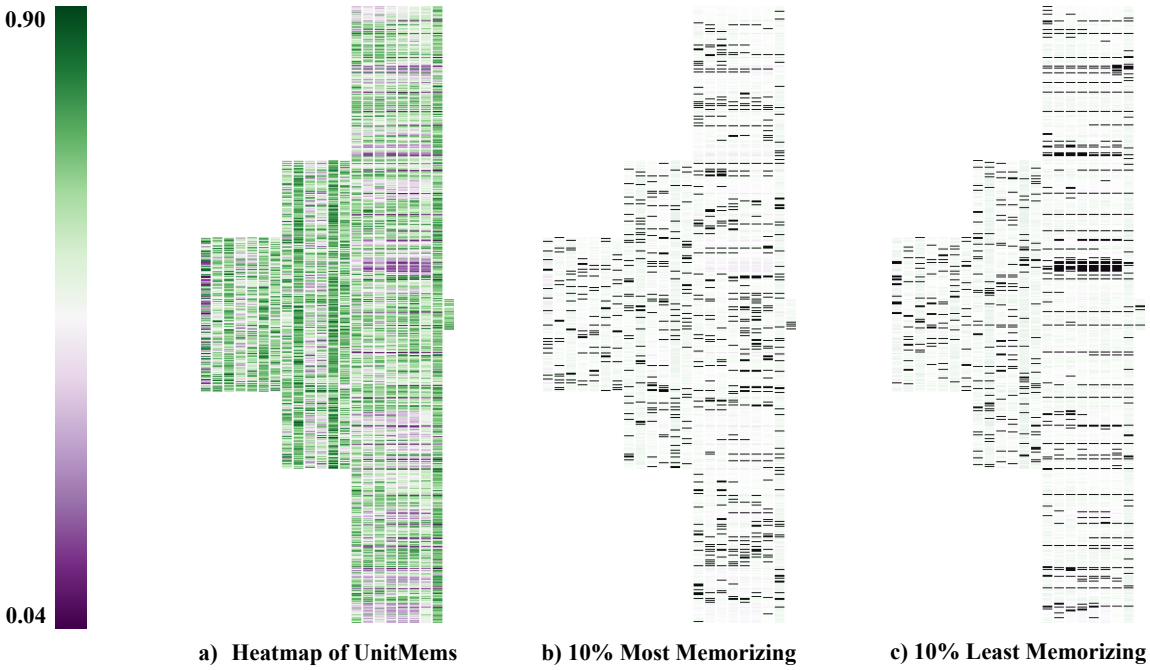


Figure 1: Visualization of UnitMem values across different layers of a neural network. Each column represents a specific layer (as referenced in Section 2.2; the leftmost layer is `conv_in`, and the rightmost layer is `conv_out`), and each row within a column's box displays the UnitMem values for individual channels in that layer. (a) Heatmap of all UnitMems, showing the distribution of memorization across all units. (b) The 10% of units with the highest memorization scores, highlighting the most influential neurons for memorization. (c) The 10% of units with the lowest memorization scores, indicating the least influential neurons in terms of memorization.

1 INTRODUCTION

In this report, we analyze the memorization phenomenon in the encoder of the Visual Autoregressive Model (VAR) Tian et al. (2024). We employ the UnitMem method Wang et al. (2024), to quantify memorization at the unit level. For the VAR encoder, this translates to analyzing memorization across the feature maps of convolutional layers, resulting in a memorization metric for each channel within the model.

2 EXPERIMENTAL SETUP

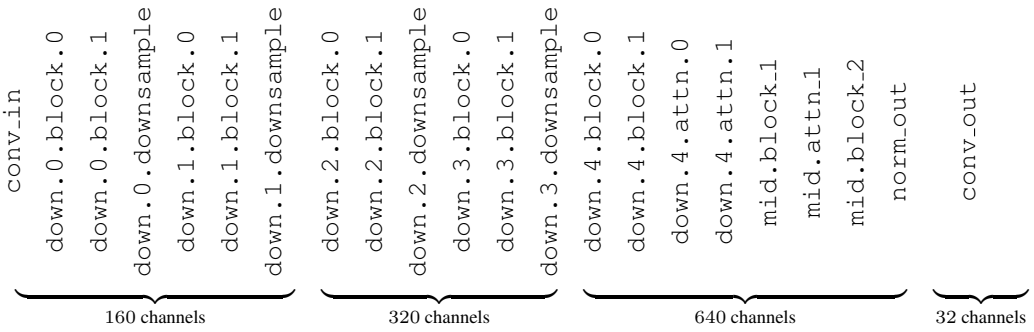
2.1 MODEL AND DATA

We utilize the smallest pretrained encoder from the VAR paper (VAR-d16), which is based on a VQVAE architecture van den Oord et al. (2018). The VQVAE consists of an encoder, a quantized codebook (lookup table), and a decoder. Our analysis focuses solely on the encoder component.

The model was pretrained on the ImageNet dataset Deng et al. (2009). Ideally, memorization analysis would be performed using the full ImageNet training set. However, due to resource constraints, we use a subset of 10,000 ImageNet images. A significant challenge is the storage requirement for activations from each data point, which becomes substantial for large datasets.

2.2 ENCODER ARCHITECTURE

We identified the following layers as the primary components of the VAR-16 encoder:



2.3 ACTIVATION COLLECTION AND AUGMENTATION

We registered a forward hook on each of the listed layers to store the activation outputs during inference. Following the methodology of the UnitMem paper, we averaged the activations across 10 different augmentations for each image to improve the robustness of the analysis. The augmentations used are:

```
import torchvision.transforms as transforms

augmentations = transforms.Compose([
    transforms.RandomApply([transforms.RandomHorizontalFlip()], p=0.5),
    transforms.RandomApply([transforms.ColorJitter(0.9 * s, 0.9 * s, 0.9 * s,
    0.1 * s)], p=0.9),  # s needs to be defined (e.g., s=1)
    transforms.RandomGrayscale(p=0.1),
    transforms.Resize((256, 256)), # Resize *before* normalization
    transforms.Lambda(lambda x: x * 2 - 1)  # Normalize to [-1, 1]
])
```

Listing 1: Data Augmentations

All images were resized to 256x256 pixels, matching the input size used during the model's training. The normalization $\mathbf{x}_{norm} = 2\mathbf{x} - 1$ (where $\mathbf{x}$ is the original image tensor and $\mathbf{x}_{norm}$ is the normalized tensor) was applied, consistent with the original VAR model training.

2.4 UNITMEM CALCULATION MODIFICATION

The original UnitMem method relies on non-negative activations resulting from ReLU activation functions. However, the VQVAE encoder does not exclusively use ReLUs, and the normalization to [-1, 1] introduces negative activations. To address this, we used the absolute value of the activations when calculating UnitMem. This decision is justified by the observation that an activation with a large magnitude, regardless of its sign, has a significant impact. Mathematically, for a weight matrix $\mathbf{W}$ and activation f:

$$f \cdot \mathbf{W} = -f \cdot (-\mathbf{W}) \tag{1}$$

This indicates that a negative activation with large magnitude can be equivalent to a positive activation if the corresponding weights in the subsequent layer are negated. We calculated UnitMem using both the mean and median, but observed no substantial differences in the results.

3 EXPERIMENTS AND RESULTS

3.1 LOCATION OF HIGHLY MEMORIZING NEURONS

We investigated the distribution of the top 10% most memorizing neurons across the encoder layers. Figure 1 shows that these neurons are scattered across various layers and channels. No clear pattern emerges indicating a strong correlation between layer depth and memorization.

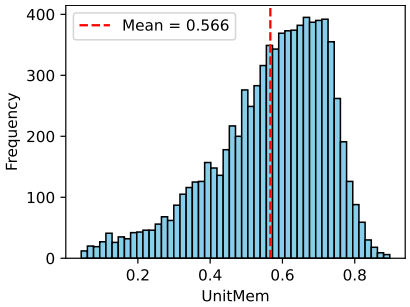


Figure 2: Histogram of UnitMem values across all units. The overall average UnitMem is approximately 57%.

Table 1 presents the 20 most memorizing units, sorted by their mean UnitMem score.

The table reveals that a significant portion of the highest memorization occurs in the `down.3.block.1` layer. Notably, the deeper `down.4.block` layers are absent from the top 20.

Table 2 shows the mean UnitMem for the top 10% and bottom 10% of units within each layer, along with a t-test comparing the two groups.

The t-tests consistently show highly significant differences ($p < 0.05$) between the most and least memorizing units within each layer, confirming that the observed differences are unlikely due to chance.

Figure 2 presents a histogram of UnitMem values across all units, revealing an overall average UnitMem of approximately 57%.

3.2 CHARACTERISTICS OF HIGHLY MEMORIZED DATA POINTS

We examined the data points that most strongly influenced the top 10% and bottom 10% of memorizing units (Figure 3). Preliminary observations suggest the following:

Highly Memorized Images: These images often exhibit high-frequency patterns, such as grids, and frequently contain images of insects. **Least Memorized Images:** These images tend to lack a clear separation between foreground and background, making it difficult to identify the primary subject. Many of these images contain green, plant-like areas.

Table 1: Top 20 Most Memorizing Units

Layer Name	Unit Index	UnitMem (Mean)
down.2.block.1	271	0.897
down.1.block.1	51	0.893
down.3.block.1	167	0.888
down.3.block.1	304	0.883
down.3.block.1	247	0.881
down.3.block.1	287	0.881
norm_out	379	0.878
down.3.block.1	267	0.877
down.2.block.0	85	0.873
down.2.block.1	71	0.871
down.3.block.1	47	0.867
down.3.block.1	314	0.866
down.2.block.1	209	0.864
down.0.block.1	142	0.863
conv_in	105	0.863
conv_in	147	0.861
down.2.downsample	239	0.859
down.2.block.1	85	0.858
down.0.block.0	87	0.857
norm_out	13	0.857

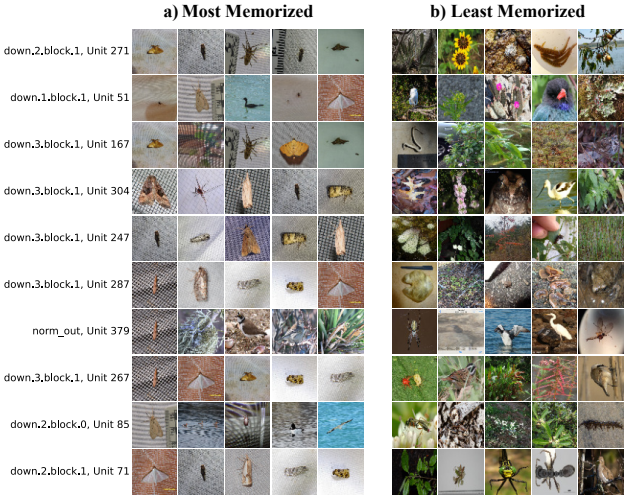


Figure 3: Samples of the Top 10 Memorizing Units

Figure 4: Top 20 Memorizing Units and Corresponding Image Samples for the Top 10 Memorizing Units

Table 2: Per-Layer Memorization Comparison

Layer Name	Mean UnitMem (Top 10%)	Mean UnitMem (Bottom 10%)	t-test (t/p)
conv_in	0.837	0.126	54.14 / 2.314e-22
down.0.block.0	0.783	0.387	29.27 / 1.249e-22
down.0.block.1	0.801	0.476	20.00 / 8.318e-15
down.0.downsample	0.742	0.292	41.51 / 6.013e-23
down.1.block.0	0.779	0.352	33.40 / 1.000e-21
down.1.block.1	0.796	0.459	27.47 / 5.351e-22
down.1.downsample	0.787	0.398	32.51 / 2.131e-24
down.2.block.0	0.791	0.366	41.56 / 2.117e-44
down.2.block.1	0.825	0.480	25.16 / 6.139e-25
down.2.downsample	0.767	0.256	50.26 / 3.546e-51
down.3.block.0	0.768	0.269	49.65 / 1.700e-46
down.3.block.1	0.837	0.520	23.39 / 4.365e-24
down.3.downsample	0.788	0.424	43.18 / 8.42e-40
down.4.block.0	0.759	0.281	52.60 / 2.102e-64
down.4.block.1	0.750	0.199	65.13 / 4.844e-74
down.4.attn.0	0.746	0.259	54.32 / 7.226e-63
down.4.attn.1	0.739	0.180	67.03 / 1.214e-78
mid.block_1	0.744	0.165	61.10 / 4.550e-72
mid.attn_1	0.749	0.169	69.44 / 6.973e-75
mid.block_2	0.722	0.200	66.87 / 1.966e-96
norm_out	0.778	0.237	27.20 / 1.329e-37
conv_out	0.731	0.421	4.45 / 0.04575

These observations are preliminary and require further investigation to confirm and quantify these patterns. A more rigorous analysis would involve clustering or other techniques to identify common features among the most and least memorized images.

4 CONCLUSION

This report provides an initial analysis of memorization in the VAR-16 encoder using the UnitMem method. We found that highly memorizing neurons are distributed across various layers, with no strong evidence for a depth-memorization correlation. The `down.3.block.1` layer appears to contain a disproportionate number of highly memorizing units. Preliminary analysis of the most and least memorized images suggests potential patterns related to image content and clarity, but further investigation is needed. Future work should include a more comprehensive analysis of image features, potentially using techniques like clustering, to identify common characteristics of highly memorized data points. Additionally, exploring the relationship between UnitMem and other metrics, such as image complexity or frequency content, could provide further insights.

REFERENCES

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction, 2024. URL `https://arxiv.org/abs/2404.02905`.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. URL `https://arxiv.org/abs/1711.00937`.

Wenhao Wang, Adam Dziedzic, Michael Backes, and Franziska Boenisch. Localizing memorization in ssl vision encoders, 2024. URL `https://arxiv.org/abs/2409.19069`.