

SEARCHING FOR MEMORIZATION IN VISUAL AUTOREGRESSIVE MODELS

Christopher Roßbach

Friedrich-Alexander-Universität Erlangen-Nürnberg

Erlangen, 91058, Germany

`christopher.rossbach@fau.de`

ABSTRACT

This work investigates memorization phenomena in visual autoregressive models (VAR), focusing on identifying and quantifying memorization at the unit (neuron/channel) level. By adapting the UnitMem methodology, I analyzed where memorization occurs within autoregressive models for images. More specifically, I analyzed at which depth and in which layer types of the self-attention blocks memorization occurs and which samples are most prone to be memorized.

1 METHODOLOGY

I applied the UnitMem methodology (Wang et al.) to visual autoregressive models (VAR) (Tian et al.). For this, I made a small modification to the original UnitMem formulation to adapt it to VAR models. The original UnitMem defines the mean activation of a unit u on a sample x as:

$$\mu_u(x) = \mathbb{E}_{x' \sim \text{Aug}(x)}[\text{activation}_u(x')], \quad (1)$$

where $\text{Aug}(x)$ is a probabilistic augmentation of x . I took the absolute value of the activation to account for the possible negative activations in VAR models and captured the excitement of a unit regardless of the sign and defined the mean activation as:

$$\mu_u(x) = \mathbb{E}_{x' \sim \text{Aug}(x)}[|\text{activation}_u(x')|]. \quad (2)$$

Further, I used the VAR model in the same way as during training. That meant I

1. supplied the model with $([s], e_1, e_2, \dots, e_{n-1})$, where $[s]$ is the start token obtained from the label and e_1, e_2, \dots, e_n are the upsampled image token embeddings of the image x obtained from the multi-scale token maps $(r_1, r_2, \dots, r_{n-1})$, and
2. ran the model with teacher-forced inputs.

2 EXPERIMENTS

The experiments were performed on the 16 blocks deep pretrained VAR model provided by Tian et al.¹, which is trained on ImageNet with 256x256 resolution. I obtained my dataset \mathcal{D}' by sampling 256 images from 32 different classes² of the ImageNet-1k training set. To approximate the expectation in equation 2, I used 8 augmentations per image.³ I restricted my search to the `attn.proj`, `ffn.fc1`, and `ffn.fc2` layers of each self-attention block, totaling $16 \cdot (1024 + 4096 + 1024) = 98,304$ units. For each of the 8192 images in \mathcal{D}' I calculated the mean activation for each unit in the selected layers and stored the result. From that data, I calculated the $\text{UnitMem}_{\mathcal{D}'}(u)$ for each unit u .

¹<https://github.com/FoundationVision/VAR>

²Experiments on the influence of the class on memorization may motivate a different split

³The only randomized augmentation used is random cropping from a 1.25x resized image, so 8 augmentations may be sufficient.

2.1 LOCALIZATION

I defined the set of the units with the highest 10% UnitMem values U_{10} as:

$$\mathcal{U}_{10} = \{u \in \mathcal{U} \mid \text{UnitMem}_{\mathcal{D}'}(u) \geq P_{90}(\{\text{UnitMem}_{\mathcal{D}'}(u) \mid u \in \mathcal{U}\})\}, \quad (3)$$

where P_{90} is the 90th percentile function and \mathcal{U} is the set of all units. For each $x \in \mathcal{D}'$ I defined the number of highly memorizing units maximally activated by x as:

$$\text{NumMemUnits}(x) = |\{u \in \mathcal{U}_{10} \mid \mu_u(x) = \max(\{\mu_u(x') \mid x' \in \mathcal{D}'\})\}|.$$

I found that highly memorizing units are distributed over all blocks, as shown in Figure 1. I noticed a strong decrease with increasing depth, except the last block, where memorization peaks again. Also, a wave pattern with local maxima at the block numbers 0, 5, 10, and 15 is visible. A very similar pattern can be observed when looking at the average UnitMem values over all units instead of the distribution of \mathcal{U}_{10} (see Figure 2).

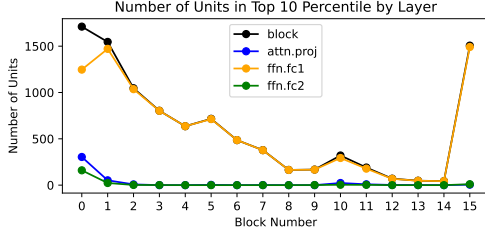


Figure 1: Number of highly memorizing units per block.

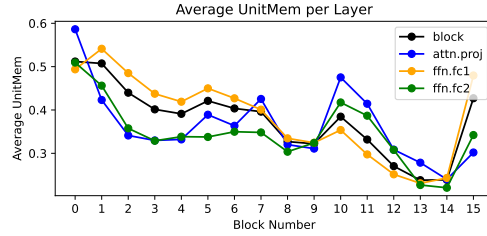


Figure 2: Average UnitMem values per block.

I also looked at the position *within* a block, i.e., the layer type (`attn.proj`, `ffn.fc1`, and `ffn.fc2`), and noted that the majority of highly memorizing units are located in the `ffn.fc1` layer, as shown in Figure 3. This trend even holds true when adjusting for the fact that most of the units are located in the `ffn.fc1` layer (4096 out of 6144 units per block). When looking at the average UnitMem values per layer type (see Figure 4), one can see a more even distribution, indicating that `ffn.fc1` layers also contain a lot of low-memorizing units. Figure 5 shows that the unit with maximal UnitMem value of each block is always found in the `ffn.fc1` layer.

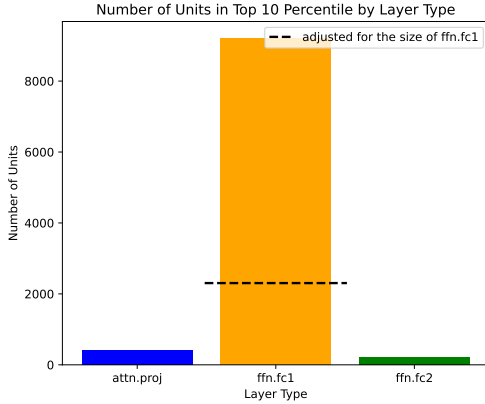


Figure 3: Number of highly memorizing units per layer type.

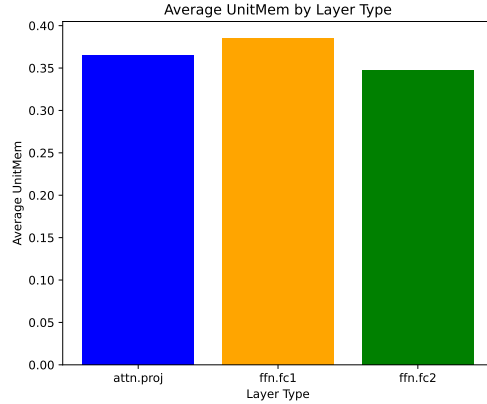


Figure 4: Average UnitMem values per layer type.

2.2 MEMORIZED SAMPLES

My experiments showed that only a relatively small number of samples are responsible for the high UnitMem values of the units in \mathcal{U}_{10} . Only about 27.2% of the samples in \mathcal{D}' are responsible for the

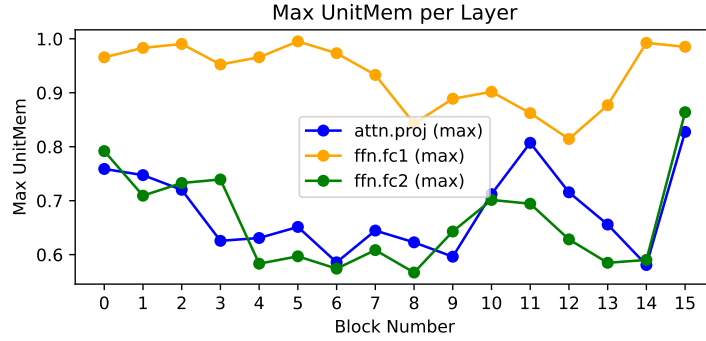


Figure 5: Layer type of the unit with maximal UnitMem value per block.

maximal activation of at least one unit in \mathcal{U}_{10} (i.e., have a NumMemUnits value greater than 0). At the same time, already 1% of the samples in \mathcal{D}' are responsible for the maximal activation of 49% of the units in \mathcal{U}_{10} . 10% of the samples in \mathcal{D}' are responsible for the maximal activation of 83% of the units in \mathcal{U}_{10} .^{4 5}

I also looked at the samples that maximally activate the highest number of highly memorizing units, i.e., the samples with the highest NumMemUnits value, and performed a qualitative assessment. I found that a large portion of these samples show a spatially repeating high-frequency pattern or huge areas of constant color (see Figure 7).

I also (qualitatively) looked at how the memorized samples differ depending on the block depth. For that, I looked at the samples that maximally activate the top 10 units (in terms of UnitMem value) in the `ff1n.fc1` layer of blocks 0, 5, 10, and 15. While the samples for blocks 0 and 5 are quite diverse, in block 10, 7 out of 10 samples stem from the same class, and in block 15, all top 10 UnitMem scores originate from the same image, as can be seen in Figure 8.

3 FUTURE WORK

1. The last observation (less diversity in memorized samples with increasing depth) is very interesting and should be further investigated in a quantitative manner. The per-class as well as the per-image distribution are interesting aspects to look at.
2. The same experiments should be performed on different model depths.
3. The influence of the class on memorization may be interesting to explore.
4. Other relevant layers like `head` may also be interesting to look at.
5. The findings in A.1 are unexpected and should be further investigated (they may not even be correct).

REFERENCES

- Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. URL <https://openreview.net/forum?id=gojL67CfS8>.
- Wenhao Wang, Adam Dziedzic, Michael Backes, and Franziska Boenisch. Localizing Memorization in SSL Vision Encoders. URL <https://openreview.net/forum?id=R46HG1IjcG>.

⁴TODO: insert numbers from last experiments

⁵Additionally, undocumented experiments suggested that the portion of memorized samples may be even smaller when looking at a larger dataset \mathcal{D}' .

A APPENDIX

You may include other additional sections here.

A.1 A LOOK AT THE VALIDATION SET

⁶ For sanity check, I constructed a dataset \mathcal{D}_{val} of 256 images per class for the same 32 classes as above from the ImageNet-1k validation set. I used the data set $\mathcal{D}_{all} = \mathcal{D}_{val} \cup \mathcal{D}'$ consisting of 16,384 images from the validation set and the images from the training set.

I expected to see less memorization for samples in \mathcal{D}_{val} as the model was not trained on these images.⁷ Figure 6 shows that the number of UnitMem scores caused by samples from the validation set is slightly higher than for samples from the training set. Restricting the analysis to only the units in $\mathcal{U}_{10, \mathcal{D}_{all}}$ (obtained as in equation 3 but on \mathcal{D}_{all}), I saw that this trend gets stronger.

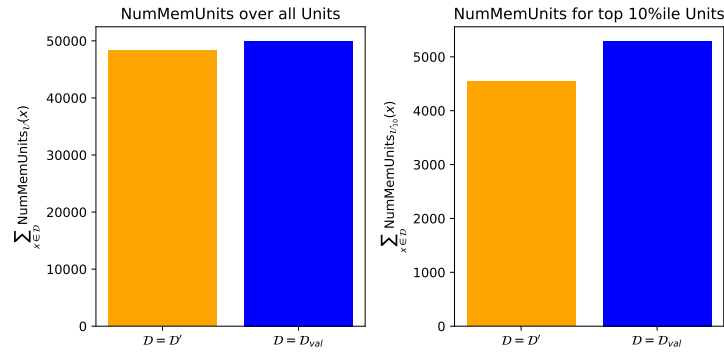


Figure 6: Number of UnitMem scores caused by samples from the training and validation set.

⁶This section is not yet fully verified.

⁷maybe it was?

A.2 MEMORIZED SAMPLES

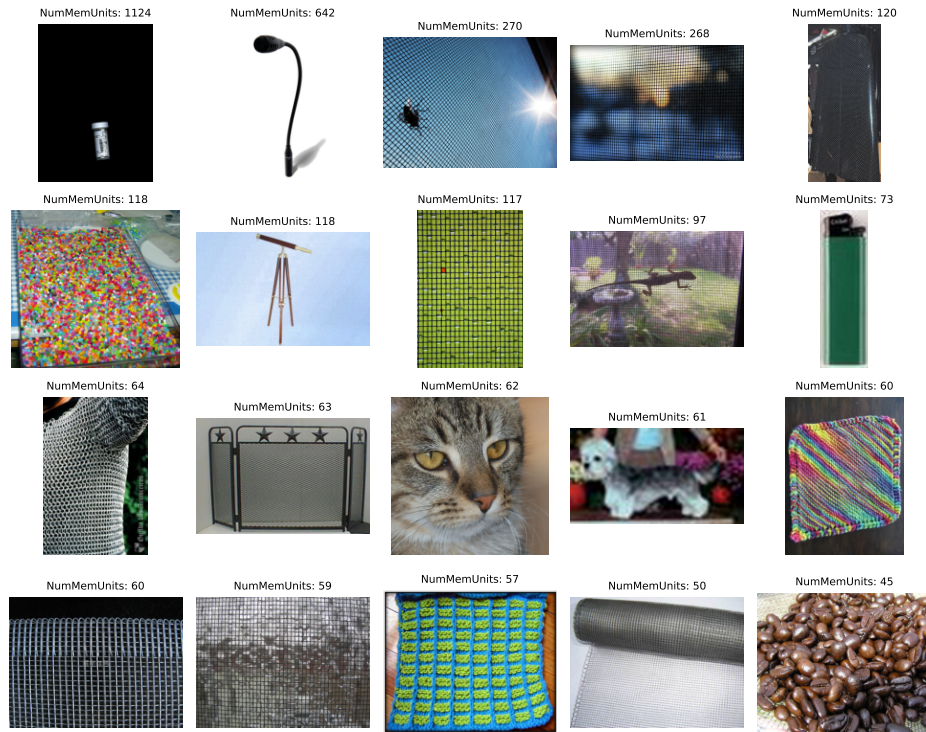


Figure 7: Samples that maximally activate the highest number of highly memorizing units.

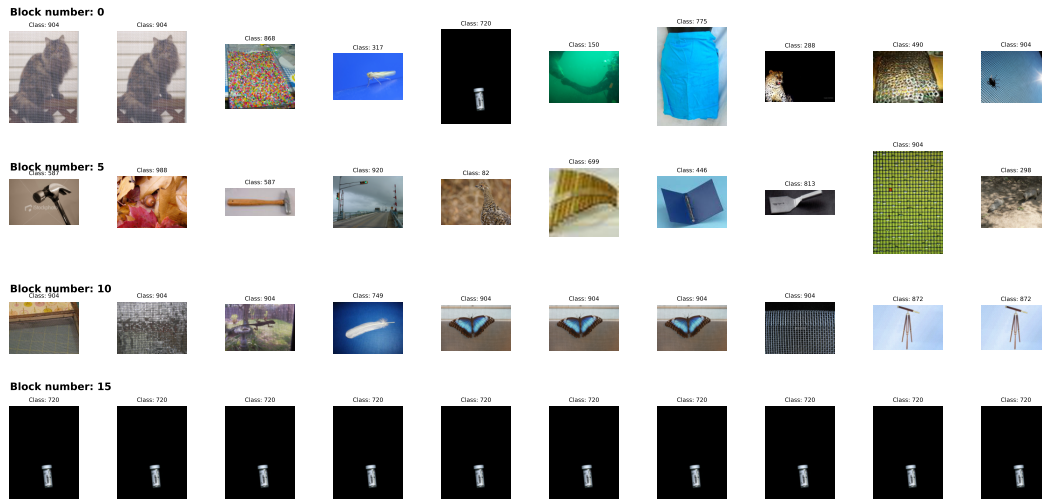


Figure 8: Samples maximally activating the top 10 units in the ff1.fc1 layer of blocks 0, 5, 10, and 15.