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, we analyze where memorization occurs within autoregressive models for images. More specifically, we analyse 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

We apply the UnitMem methodology (Wang et al.) to visual autoregressive models (VAR) (Tian et al.). For this, we 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:

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

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

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

Further we use the VAR model in the same way as during training. That means, we that we

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

2 EXPERIMENTS

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

[1] https://github.com/FoundationVision/VAR

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

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

2.1 Localization

We define the set of the units with the highest 10% UnitMem values U_{10} as:

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

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

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

We find that highly memorizing units are distributed over all blocks, as shown in Figure **??**. We notice a strong decrease with increasing depth, except the last block, where momorization peaks. 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 the distribution of $\mathcal{U}_{10}$ (see Figure **??**).

We 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 higly memorizing units are located in the `ffn.fc1` layer, as shown in Figure **??**. 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 **??**), we see a more even distribution, indicating that `ffn.fc1` layers also contain a lot of low memorizing units. Figure **??** shows that the unit with maximal `UnitMem` value of each block is always found in the `ffn.fc1` layer.

2.2 Memorized Samples

Our experiments show that only a relatively small number of samples are responsible for the high `UnitMem` values of the Units in $\mathcal{U}_{10}$. Only about 27.3% of the samples in $\mathcal{D}'$ are responsible for the 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]

We 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 assesment. We find that a huge portion of these samples show a spacially repeating high frequency pattern or huge areas of constant color (see Figure **??**).

We also (qualitatively) looked how the memorized samples differ depending on the block depth. For that we looked at the samples that are maximally activating the top 10 units (in terms of `UnitMem` value) in the `ffn.fc1` layer of blocks 0, 5, 10 and 15. While the samples for the 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 **??**.

2.2.1 A Look at the Validation Set

[6] For sanity check, we construct a dataset $\mathcal{D}_{val}$ of 256 images per class for the same 32 classes as above from the ImageNet-1k validation set. We use 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.

We expect to see less memorization, for samples in $\mathcal{D}_{val}$ as the model was not trained on these images.[7] Figure **??** 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 $\mathcal{D}_{all}$) we see that this trend increases.

[4] TODO: insert numbers from last experiments

[5] Additionally, not documented experiments suggest that the portion of memorized samples may be even smaller when looking at a larger dataset $\mathcal{D}'$.

[6] This section is not yet fully verified.

[7] maybe it was?

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-classes 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 2.2.1 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=R46HGlIjcG.

A Appendix

You may include other additional sections here.