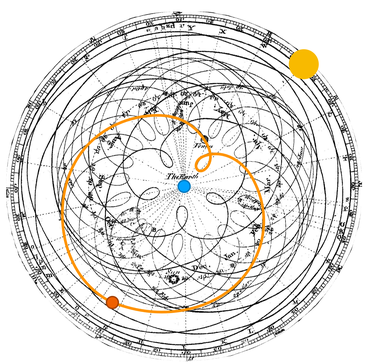
# The Pitfalls of Memorization: When Memorization Hurts Generalization

Reza Bayat[2,*], Mohammad Pezeshki[1,*]
Elvis Dohmatob[1], David Lopez-Paz[1], Pascal Vincent[1,2,3]

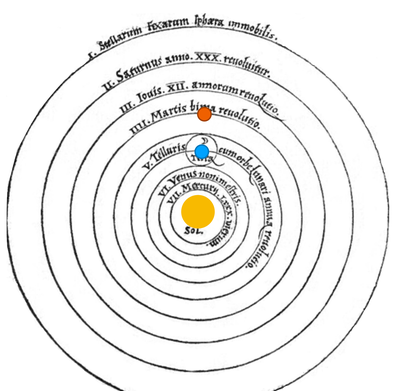
[1] FAIR at Meta
[2] Mila, Université de Montréal DIRO, [3] CIFAR

Scan for arXiv

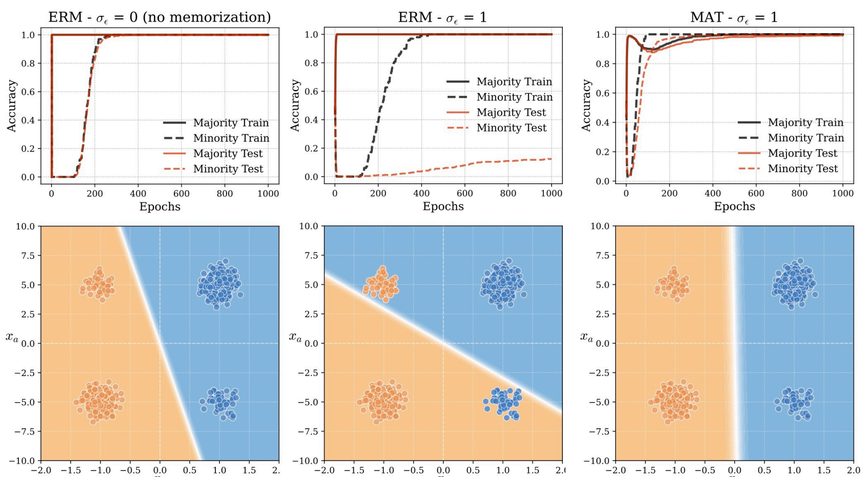
Neural networks can "simplify and memorize"- learning simple shortcuts and then memorizing exceptions.

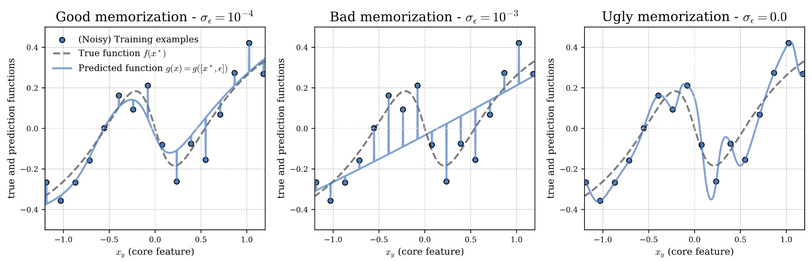



Earth at the center (170 AD)

Sun at the center (1543)




**The Good: when memorization benefits generalization:** Model learns the true function well and memorizes the residual noise towards zero training loss.

**The Bad: when memorization prevents generalization:** Model relies more on memorization than learning the true function to achieve zero training loss

**The Ugly: catastrophic overfitting:** With no example-specific features the model severely overfits.

## Memorization Aware Training (MAT)

**Insight:** The model shows poor held-out performance on memorized samples

**Solution:** Leveraging held-out predictions of a reference model to shift the logits, **prioritizing learning of examples with worse generalization performance.**

$$\mathcal{L}^{\text{MAT}} = \frac{1}{n}\sum_{i=1}^{n} l\big(\text{softmax}(f(\boldsymbol{x}_i;\boldsymbol{w}) + \log \overline{p}^{ho}(\cdot \mid \boldsymbol{x}_i)), y_i\big)$$

$$\overline{p}^{ho}(y \mid \boldsymbol{x}) := \sum_{y^{ho}} p(y \mid y^{ho}) p(y^{ho} \mid \boldsymbol{x})$$

$$p(y \mid y^{ho}) = \frac{p(y, y^{ho})}{\sum_{y'} p(y', y^{ho})}, \quad \text{where} \quad p(y, y^{ho}) = \frac{1}{n} \sum_{\{i : y_i = y\}} p(y^{ho} \mid \boldsymbol{x}_i)$$

| | | | Waterbirds | | CelebA | | MultiNLI | | CivilComments | |
|---|---|---|---|---|---|---|---|---|---|---|
| *tr* | *va* | | Avg | WGA | Avg | WGA | Avg | WGA | Avg | WGA |
| ✓ | ✓ | GroupDRO | 90.2± 0.3 | 86.5 ± 0.5 | 93.1 ± 0.3 | 88.3 ± 2.1 | 80.6 ± 0.4 | 73.4 ± 4.8 | 84.2 ± 0.2 | 73.8 ± 0.6 |
| ✗ | ✓ | ERM | 97.3 | 72.6 | 95.6 | 47.2 | 82.4 | 67.9 | 83.1 | 69.5 |
| | | LFF† | 91.2 | 78.0 | 85.1 | 77.2 | 80.8 | 70.2 | 68.2 | 50.3 |
| | | JTT† | 93.3 | 86.7 | 88.0 | 81.1 | 78.6 | 72.6 | 83.3 | 64.3 |
| | | LC† | - | 90.5 ± 1.1 | - | 88.1 ± 0.8 | - | - | - | 70.3 ± 1.2 |
| | | AFR† | 94.2 ± 1.2 | 90.4 ± 1.1 | 91.3 ± 0.3 | 82.0 ± 0.5 | 81.4 ± 0.2 | 73.4 ± 0.6 | 89.8 ± 0.6 | 68.7 ± 0.6 |
| | | MAT | 90.4 ± 0.7 | 88.1 ± 0.9 | 92.4 ± 0.4 | 90.5 ± 1.0 | 79.4 ± 0.4 | 74.6 ± 1.0 | 84.3 ± 0.3 | 74.0 ± 0.8 |
| ✗ | ✗ | ERM | 83.5 | 66.4 | 95.4 | 54.3 | 82.1 | 67.9 | 81.3 | 67.2 |
| | | uLA† | 91.5 ± 0.7 | 86.1 ± 1.5 | 93.9 ± 0.2 | 86.5 ± 3.7 | - | - | - | - |
| | | XRM† | 89.3 ± 0.6 | 88.1 ± 0.9 | 91.4 ± 0.5 | 89.1 ± 1.3 | 75.8 ± 1.2 | 72.1 ± 1.0 | 84.4 ± 0.6 | 72.2 ± 0.8 |
| | | MAT | 90.4 ± 0.7 | 88.1 ± 0.9 | 92.3 ± 0.3 | 89.9 ± 1.2 | 79.6 ± 0.2 | 73.0 ± 0.8 | 85.7 ± 0.1 | 68.1 ± 0.7 |