The Diversity Coefficient: a Data Quality Metric for Machine Learning shows that LLM are Pretrained on formally Diverse Data

Alycia Lee [1], Brando Miranda [1], Patrick Yu [2], and Sanmi Koyejo [1]

[1]Computer Science, Stanford University [2]University of Illinois Urbana-Champaign

Summary

Contributions

- Develop a **data quality metric** of the pre-training data of Large Language Models (LLLMs)
- Our focus is on a formal **quantitative** definition of **data diversity**

Background

Task2Vec method (Achille, Lam, et al 2019) computes vectorial representation of task τ using probe network.

- Definition: Task2Vec embedding of τ = diagonal entries of Fisher information matrix (FIM) after fine-tuning final layer of probe network on task dataset D_τ:

$$\hat{F}(D_\tau, f_w) = \mathbb{E}_{x,y\sim\hat{p}(x|\tau)\hat{p}(y|x,f_w)}\left[\nabla_w \log p(y|x,f_w)\nabla_w \log p(y|x,f_w)^T\right]$$

- Intuition: FIM = measure of information in probe network parameter about $p_w(x,y)$, i.e. importance of parameter for network to perform a task.

Methods

Task2Vec-based diversity coefficient (Miranda, Yu, et al 2022) approximately measures the intrinsic variability of tasks in a few-shot learning benchmark B.

- Ground Truth Diversity Coefficient: expected distance between pairs of tasks τ_1, τ_2:

$$\hat{div}(B) = \mathbb{E}_{\tau_1,\tau_2\sim\hat{p}(\tau|B):\tau_1\neq\tau_2}\left[d(p(x_1,y_1 \mid \tau_1), p(x_2,y_2 \mid \tau_2))\right]$$

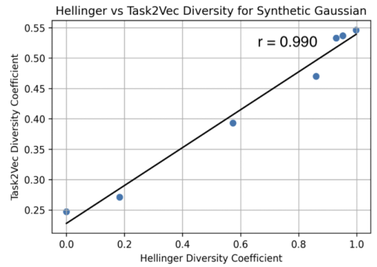
- Definition: expected distance between pairs of tasks τ_1, τ_2 as Task2Vec embeddings:

$$\hat{div}(B) = \mathbb{E}_{\tau_1,\tau_2\sim\hat{p}(\tau|B):\tau_1\neq\tau_2}\mathbb{E}_{D_1\sim\hat{p}(x_1,y_1|\tau_1),D_2\sim\hat{p}(x_2,y_2|\tau_2)}\left[d(\mathrm{diag}(\hat{F}_{D_1,f_w}), \mathrm{diag}(\hat{F}_{D_2,f_w})\right]$$

- Fixed probe network f_w required for fair comparison of tasks.
- Intuition: diversity $\sim$ information content in data.

Experiments & Results

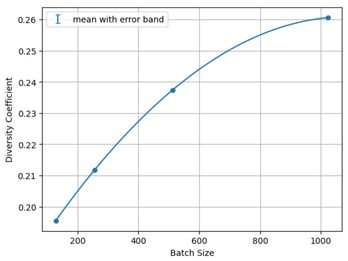
Task2Vec Diversity Coefficient Correlates with Ground Truth Diversity



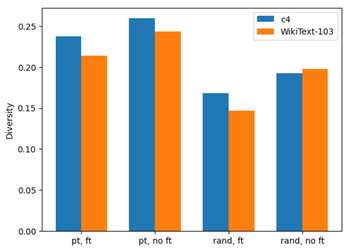
Diversity of c4, Wikitext-103, and The Pile are twice as high vs. benchmarks evaluated in Miranda, Yu, et al (2022)

Dataset	Probe Network	Diversity Coeff
MiniImagenet	Resnet18	0.117 ±2.098e-5
Cifar-fs	Resnet18	0.100 ±2.18e-5
c4	GPT-2	**0.2374 ±2.785e-5**
WikiText-103	GPT-2	**0.2140 ±7.93e-5**
The Pile	GPT-2	**0.2463 ±3.034-5**

Batch size has positive correlation with diversity, but diminishing returns

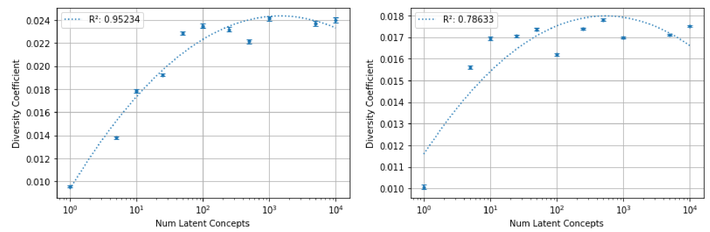


Random probe underestimates diversity, none fine-tuned overestimates diversity

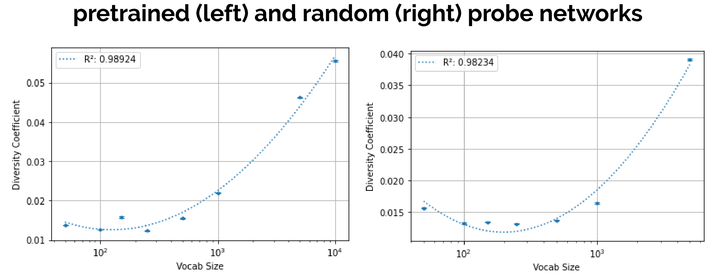


Experiments & Results (cont.)

Diversity coefficient correlates with latent concepts in GINC data, using pretrained (left) and random (right) probe networks



Diversity coefficient correlates with vocab size in GINC data, using pretrained (left) and random (right) probe networks



Takeaways

- Diversity coeff's of three LLM pretraining datasets – c4, WikiText-103, and The Pile – are high compared to previous work on vision datasets.
- Diversity coeff is sensitive to changes in task batch size and probe network.
- Diversity coeff value is underestimated when using random probe network vs. overestimated when using none fine-tuned network.
- Diversity coefficient correlates individually with latent concepts and vocab size in GINC datasets.