Diffuse and disperse

A Preprint

Diffusion models have great perfomance in an area of generating, while still being disconnected from representation learning as diffusion mainly focuses on denoising and lacks explicit representation learning term. The method authors suggest requires no pretraining, no additional data or parameters. Briefly saying, Dispersive loss acts like contrastive loss without positive pairs(which means negative are used).

In image generation, REPA explores enhancing generative models through auxiliary representation learning. It aligns the intermediate representations of a generative model with those from a frozen, high-capacity, pre-trained encoder, which may be trained using external data and diverse objectives. SARA builds on this by introducing structural and adversarial representation alignment. In multimodal settings, SoftREPA extends REPA by aligning noisy image representations with soft text embeddings. While REPA and its extensions show substantial practical improvements, they require additional pre-training and depend on external sources of information, making it difficult to determine whether the gains stem from the alignment itself or from increased compute and data access.

The core idea of this paper is to use representation learning techiques, add regularization term in diffusion in order to lean model's inner representations by dispersing.

variant	contrastive	dispersive
InfoNCE	$\frac{D(z_i, z_i^+)}{\tau} + \log \sum_j \exp\left(-\frac{D(z_i, z_j)}{\tau}\right)$	$\log \mathbb{E}_{i,j}\left[\exp\left(-\frac{D(z_i, z_j)}{\tau}\right)\right]$
Hinge	$D(z_i, z_i^+)^2 + \mathbb{E}_j\left[\max(0, \epsilon - D(z_i, z_j))^2\right]$	$\mathbb{E}_{i,j}\left[\max(0, \epsilon - D(z_i, z_j))^2\right]$
Covariance	$(1 - \mathrm{Cov}_{mm})^2 + w \sum_{n \neq m} \mathrm{Cov}_{mn}^2$	$\sum_{m,n} \mathrm{Cov}_{mn}^2$

Таблица 1: Contrastive vs. dispersive objectives

Then the loss is defined as:

$$\mathcal{L}(x) = \mathbb{E}_x[L_{\text{diff}}] + \lambda L_{\text{disp}}$$

where x is in batch X. It was made due to the fact that diffusion processes one by one while contrastive learning is made on pairs which requires batch. Conceptually, Dispersive Loss can be derived from any existing contrastive loss by appropriately removing the positive terms. More importantly, all Dispersive Loss functions are applicable to a single-view batch, eleminating the need for multi-view augmentations which means that method can be used just as plug-in. As for experiments authors took DiT and SiT and ImageNet. Models were trained on $32 \times 32 \times 4$ VAE encoded latent space. Sampling is performed using the ODE-based Heun sampler with 250 steps.

Next, authors compare contrasitve and dispersive. To apply a contrastive loss, two views are sampled for each training example to form a positive pair.

Experiments show that overall method improves all models, especially big ones which means that it reduces overfitting. Also application to the MeanFlow framework and the achievement of new SOTA for one-step generation and table 7 in the paper explicitly demonstrates this, showing that "MeanFlow-XL/2 + Disp"achieves a better FID score than other state-of-the-art one-step models.

Figure 5: **Qualitative results.** We present curated samples generated from SiT-XL/2 with Dispersive Loss.

model	epochs	**baseline**	**dispersive**	Δ
MF-B/4	80	18.78	17.61	–6.23%
MF-B/2	80	9.77	8.97	–8.18%
MF-B/2	240	6.17	5.69	–7.77%
MF-XL/2	240	3.43	**3.21**	–6.41%

method	params	step	NFE	FID
iCT-XL/2 [37]	675M	1	1	34.24
Shortcut-XL/2 [11]	675M	1	1	10.60
IMM-XL/2 [43]	675M	1	2	7.77
MeanFlow-XL/2 [12]	676M	1	1	3.43
MeanFlow-XL/2 + **Disp**	676M	1	1	**3.21**

variant	**baseline**	**contrastive** (independent noise)	**contrastive** (restricted noise)	**dispersive**
none	36.49	–	–	–
InfoNCE, ℓ_2	–	43.66 (+19.65%)	36.57 (+0.22%)	**32.35** (–11.35%)
InfoNCE, cosine	–	41.62 (+14.06%)	34.83 (–4.55%)	34.33 (–5.92%)
Hinge	–	43.02 (+17.89%)	35.14 (–3.70%)	33.93 (–7.02%)
Covariance	–	37.85 (+3.73%)	35.87 (–1.70%)	35.82 (–1.84%)