Generalised Mutual Information for Discriminative Clustering

Anonymous Author(s)
Affiliation
Address
email

Abstract

In the last decade, recent successes in deep clustering majorly involved the mutual information (MI) as an unsupervised objective for training neural networks with increasing regularisations. While the quality of the regularisations have been largely discussed for improvements, little attention has been dedicated to the relevance of MI as a clustering objective. In this paper, we first highlight how the maximisation of MI does not lead to satisfying clusters. We identified the Kullback-Leibler divergence as the main reason of this behaviour. Hence, we generalise the mutual information by changing its core distance, introducing the generalised mutual information (GEMINI): a set of metrics for unsupervised neural network training. Unlike MI, some GEMINIs do not require regularisations when training. Some of these metrics are geometry-aware thanks to distances or kernels in the data space. Finally, we highlight that GEMINIs can automatically select a relevant number of clusters, a property that has been little studied in deep clustering context where the number of clusters is a priori unknown.

1 Introduction

Clustering is a fundamental learning task which consists in separating data samples into several categories, each named cluster. This task hinges on two main questions concerning the assessment of correct clustering and the actual number of clusters that may be contained within the data distribution. However, this problem is ill-posed since a cluster lacks formal definitions which makes it a hard problem (Kleinberg, 2003).

Model-based algorithms make assumptions about the true distribution of the data as a result of some latent distribution of clusters (Bouveyron et al., 2019). These techniques are able to find the most likely cluster assignment to data points. These models are usually generative, exhibiting an explicit assumption of the prior knowledge on the data.

Early deep models to perform clustering first relied on autoencoders, based on the belief that an encoding space holds satisfactory properties (Xie et al., 2016; Ghasedi Dizaji et al., 2017; Ji et al., 2019). However, the drawback of these architectures is that they do not guarantee that data samples which should meaningfully be far apart remain so in the feature space. Early models that dropped decoders notably used the mutual information (MI) (Krause et al., 2010; Hu et al., 2017) as an objective to maximise. The MI can be written in two ways, either as measure of dependency between two variables x and y, e.g. data distribution $p(x)$ and cluster assignment $p(y)$:

$$\mathcal{I}(x;y) = D_{\mathrm{KL}}(p(x,y)||p(x)p(y)), \tag{1}$$

or as an expected distance between implied distributions and the overall data:

$$\mathcal{I}(x;y) = \mathbb{E}_{p(y)}[D_{\mathrm{KL}}(p(x|y)||p(x))], \tag{2}$$

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

with D_{KL} being the Kullback-Leibler (KL) divergence. Related works often relied on the notion of MI as a measure of coherence between cluster assignments and data distribution (Hjelm et al., 2019). Regularisation techniques were employed to leverage the potential of MI, mostly by specifying model invariances, for example with data augmentation (Ji et al., 2019).

The maximisation of MI thus gave way to contrastive learning objectives which aim at learning stable representations of data through such invariance specifications (Chen et al., 2020; Caron et al., 2020). The contrastive loss maximises the similarity between the features of a sample and its augmentation, while decreasing the similarity with any other sample. Clustering methods also benefited from recent successful deep architectures (Li et al., 2021; Tao et al., 2021; Huang et al., 2020) by encompassing regularisations in the architecture. These methods correspond to discriminative clustering where we seek to directly infer cluster given the data distribution. Initial methods also focused on alternate schemes, for example with curriculum learning (Chang et al., 2017) to iteratively select relevant data samples for training, by alternating K-means cluster assignment with supervised learning using the inferred labels (Caron et al., 2018), or by proceeding to multiple distinct training steps (Van Gansbeke et al., 2020; Dang et al., 2021; Park et al., 2021).

However, most of the methods above rarely discuss their robustness when the number of clusters to find is different from the amount of preexisting known classes. While previous work was essentially motivated by considering MI as a dependence measure, we explore in this paper the alternative definition of the MI as the expected distance between data distribution implied by the clusters and the entire data. We extend it to incorporate cluster-wise comparisons of implied distributions, and question the choice of the KL divergence with other possible statistical distances.

Throughout the introduction of the generalised mutual information (GEMINI), the contributions of this paper are:

- A demonstration of how the maxima of MI are not sufficient criteria for clustering. This extends the contribution of (Tschannen et al., 2019) to the discrete case.
- The introduction of a set of metrics called GEMINIs involving different distances between distributions which can incorporate prior knowledge on the geometry of the data. Some of these metrics do not require regularisations.
- A highlight of the implicit selection of clusters from GEMINIs which allows to select a relevant number of cluster during training.

2 Is mutual information a good clustering objective?

We consider in this section a dataset consisting in N unlabelled samples $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^N$. We distinguish two major use cases of the mutual information: one where we measure the dependence between two continuous variables, as is the case in representation learning, and one where the random variable is discrete. In representation learning, the goal is to construct a continuous representation $\boldsymbol{z}$ extracted from the data $\boldsymbol{x}$ using a learnable distribution of parameters θ. In clustering, samples $\boldsymbol{x}$ are assigned to the discrete variable y through another learnable distribution.

2.1 Representation learning

Representation learning consists in finding high-level features $\boldsymbol{z}_i$ extracted from the data $\boldsymbol{x}_i$ in order to perform a *downstream task*, e.g. clustering or classification. Maximising the MI becomes then interesting to get coherent features (Hjelm et al., 2019). However, estimating correctly MI on two continuous domains is hard, thus lower bounds are preferred, e.g. variational estimators such as MINE (Belghazi et al., 2018), $\mathcal{I}_{\mathrm{NCE}}$(Van den Oord et al., 2018). Another common choice of loss function to train features are contrastive losses such as NT-XENT (Chen et al., 2020) where the similarity between the features $\boldsymbol{z}_i$ from data $\boldsymbol{x}_i$ is maximised with the features $\tilde{\boldsymbol{z}}$ from a data-augmented $\tilde{\boldsymbol{x}}_i$ against any other features $\boldsymbol{z}_j$. Recently, Do et al. (2021) achieved excellent performances in single-stage methods by highlighting the link between the $\mathcal{I}_{\mathrm{NCE}}$ estimator (Van den Oord et al., 2018) and contrastive learning losses. Representation learning therefore comes at the cost of a complex lower bound estimator on MI, which often requires data augmentation. Moreover, it was noticed that the MI is hardly predictive of downstream tasks (Tschannen et al., 2019) when the variable y is continuous, i.e. a high value of MI does not clarify whether the discovered representations are insightful with regards to the target of the downstream task.

(a) Good decision boundary (b) Misplaced decision boundary

Figure 1: Example of maximised MI for a Gaussian mixture $p(x) \sim \frac{1}{2}\mathcal{N}(\mu_0, \sigma^2) + \frac{1}{2}\mathcal{N}(\mu_1, \sigma^2)$. It is clear that Figure 1a presents the best decision boundary and posterior between the two Gaussian distributions. Yet, as the posterior turns sharper, the difference between both MIs converges to 0.

2.2 Discriminative clustering

The MI has been first used as an objective for learning discriminative clustering models (Bridle et al., 1992). Associated architectures went from simple logistic regression (Krause et al., 2010) to deeper architectures (Hu et al., 2017; Ji et al., 2019). Beyond architecture improvement, the MI maximisation was also carried with several regularisations. These regularisations include penalty terms such as weight decay (Krause et al., 2010) or Virtual Adversarial Training (VAT, Hu et al., 2017; Miyato et al., 2018). Data augmentation was further used to provide invariances in clustering, as well as specific architecture designs like auxiliary clustering heads (Ji et al., 2019). Rewriting the MI in terms of entropies:

$$\mathcal{I}(x; y) = \mathcal{H}(y) - \mathcal{H}(y|x) \tag{3}$$

highlights a requirement for balanced clusters, through the cluster entropy term $\mathcal{H}(y)$. Indeed, a uniform distribution maximises the entropy. This hints that an unregularised discrete mutual information for clustering can possibly produce uniformly distributed clusters among samples, regardless of how close they could be. We highlight this claim in section 2.3. As an example of regularisation impact: maximising the MI with ℓ_2 constraint can be equivalent to a soft and regularised K-Means in a feature space (Jabi et al., 2019). In clustering, the number of clusters to find is usually not known in advance. Therefore, an interesting clustering algorithm should be able to find a relevant number of clusters, i.e. perform model selection. However, model selection for parametric deep clustering models is expensive (Ronen et al., 2022). Cluster selection through MI maximisation has been little studied in related works, since experiments usually tasked models to find the (supervised) classes of datasets. Furthermore, the literature diverged towards deep learning methods focusing mainly on images, yet rarely on other type of data such as tabular data (Min et al., 2018).

2.3 Maximising the MI can lead to bad decision boundaries

Maximising the MI directly can be a poor objective: a high MI value is not necessarily predictive of the quality of the features regarding downstream tasks (Tschannen et al., 2019) when y is continuous. We support a similar argument for the case where the data x is a continuous random variable and the cluster assignment y a categorical variable. Indeed, the MI can be maximised by setting appropriately a sharp decision boundary which partitions evenly the data. This reasoning can be seen in the entropy-based formulation of the MI (Eq. 3): any sharp decision boundary minimises the negative conditional entropy, while ensuring balanced clusters maximises the entropy of cluster proportions. Consider for example Figure 1, where a mixture of Gaussian distributions with equal variances is separated by a sharp decision boundary. We highlight that both models will have the same mutual information on condition that the misplaced decision boundary of Figure 1b splits evenly the dataset (see Appendix A).

Globally, MI misses the intuitive idea in clustering that any two points close to one another, according to some chosen metric, may be in the same cluster. Hence regularisation to ensure this intuition is necessary. A sketch of these insights was mentionned by Bridle et al. (1992); Corduneanu and Jaakkola (2012).

3 Extending the mutual information to the generalised mutual information

Given the identified limitations of MI, we now describe the discriminative clustering framework based on our perception of the mutual information. We then detail the different statistical distances we can use to extend MI to the generalised mutual information (GEMINI).

3.1 The discriminative clustering framework for GEMINIs

We change our view on the mutual information by seeing it as a discriminative clustering objective that aims at separating the data distribution given cluster assignments $p(x|y)$ from the data distribution $p(x)$ according to the KL divergence:

$$\mathcal{I}(\boldsymbol{x}; y) = \mathbb{E}_{y\sim p(y)}\left[D_{\mathrm{KL}}(p(\boldsymbol{x}|y)\|p(\boldsymbol{x}))\right]. \tag{4}$$

To highlight the discriminative clustering design, we explicitly remove our hypotheses on the data distribution by writing $p_{\text{data}}(\boldsymbol{x})$. The only part of the model that we design is a conditional distribution $p_\theta(y|\boldsymbol{x})$ that assigns a cluster y to a sample $\boldsymbol{x}$ using the parameters θ (Minka, 2005). This conditional distribution can typically be a neural network of adequate design regarding the data, e.g. a CNN, or a simple categorical distribution. Consequently, the cluster proportions are controlled by θ because $p_\theta(y) = \mathbb{E}[p_\theta(y|\boldsymbol{x})]$ and so is the conditional distribution $p_\theta(\boldsymbol{x}|y)$ even though intractable. This questions how Eq. (4) can be computed. Fortunately, well-known properties of MI can invert the distributions on which the KL divergence is computed (Bridle et al., 1992; Krause et al., 2010) via Bayes' theorem:

$$\mathcal{I}(\boldsymbol{x}; y) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[D_{\mathrm{KL}}(p_\theta(y|\boldsymbol{x})\|p_\theta(y))\right], \tag{5}$$

which is possible to estimateSince we highlighted earlier that the KL divergence in the MI can lead to inappropriate decision boundaries, we are interested in replacing it by other distances or divergences. However, changing it in Eq. (5) would focus on the separation of cluster assignments from the cluster proportions which may be irrelevant to the data distribution. We rather alter Eq. (4) to clearly show that we separate data distributions given clusters from the entire data distribution because it allows us to take into account the data space geometry.

3.2 The GEMINI

The goal of the GEMINI is to separate data distributions according to an arbitrary distance D, i.e. changing the KL divergence for another divergence or distance in the MI. Moreover, we question the evaluation of the distance between the distribution of the data given a cluster assumption $p_\theta(\boldsymbol{x}|y)$ and the entire data distribution $p_{\text{data}}(\boldsymbol{x})$. We argue that it is intuitive in clustering to compare the distribution of one cluster against the distribution of *another cluster* rather than the data distribution. This raises the definition of two GEMINIs, one named *one-vs-all* (OvA):

$$\mathcal{I}_D^{\text{OvA}}(\boldsymbol{x}; y) = \mathbb{E}_{y\sim p_\theta(y)}\left[D(p_\theta(\boldsymbol{x}|y)\|p_{\text{data}}(\boldsymbol{x}))\right], \tag{6}$$

which compares the cluster distributions to the data distribution, and the *one-vs-one* (OvO) in which we independently draw cluster assignments y_a and y_b:

$$\mathcal{I}_D^{\text{OvO}}(\boldsymbol{x}; y) = \mathbb{E}_{y_a, y_b\sim p_\theta(y)}\left[D(p_\theta(\boldsymbol{x}|y_a)\|p_\theta(\boldsymbol{x}|y_b))\right], \tag{7}$$

There exists other distances than the KL to measure how far two distributions p and q are one from the other. We can make a clear distinction between two types of distances, Csiszar's f-divergences (Csiszár, 1967) and integral probability metrics (IPM) (Sriperumbudur et al., 2009). However, unlike f-divergences, IPM-derived distances like the Wasserstein distance or the maximum mean discrepancy (MMD)(Gneiting and Raftery, 2007; Gretton et al., 2012) bring knowledge about the data throughout either a distance c or a kernel κ: these distances are geometry-aware.

f-divergence GEMINIs: These divergences involve a convex function $f : \mathbb{R}^+ \to \mathbb{R}$ such that $f(1) = 0$. This function is applied to evaluate the ratio between two distributions p and q, as in Eq. (8):

$$D_{\text{f-div}}(p, q) = \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[f\left(\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}\right)\right]. \tag{8}$$

We will focus on three f-divergences: the KL divergence, the total variation (TV) distance and the squared Hellinger distance. While the KL divergence is the usual divergence for the MI, the TV and

Table 1: Definition of the GEMINI for f-divergences, MMD and the Wasserstein distance. We directly write here the equation that can be optimised to train a discriminative model $p_\theta(y|\boldsymbol{x})$ via stochastic gradient descent since they are expectations over the data.

Name	Equation
KL OvA/MI	$\mathbb{E}_{p_{\text{data}}(\boldsymbol{x})}\left[D_{\text{KL}}(p_\theta(y\|\boldsymbol{x})\|p_\theta(y))\right]$
KL OvO	$\mathbb{E}_{p_{\text{data}}(\boldsymbol{x})}[D_{\text{KL}}(p_\theta(y\|\boldsymbol{x})\|p_\theta(y))) + D_{\text{KL}}(p_\theta(y)\|p_\theta(y\|\boldsymbol{x})))]$
Squared Hellinger OvA	$1 - \mathbb{E}_{p_{\text{data}}(\boldsymbol{x})}[\mathbb{E}_{p_\theta(y)}[\sqrt{\frac{p_\theta(y\|\boldsymbol{x})}{p_\theta(y)}}]]$
Squared Hellinger OvO	$\mathbb{E}_{p_{\text{data}}(\boldsymbol{x})}[\mathbb{V}_{p_\theta(y)}[\sqrt{\frac{p_\theta(y\|\boldsymbol{x})}{p_\theta(y)}}]]$
TV OvA	$\mathbb{E}_{p_{\text{data}}(\boldsymbol{x})}[D_{\text{TV}}(p_\theta(y\|\boldsymbol{x})\|p_\theta(y))]$
TV OvO	$\frac{1}{2}\mathbb{E}_{p_{\text{data}}(\boldsymbol{x})}[\mathbb{E}_{y_a,y_b\sim p_\theta(y)}[\|\frac{p_\theta(y_a\|\boldsymbol{x})}{p_\theta(y_a)} - \frac{p_\theta(y_b\|\boldsymbol{x})}{p_\theta(y_b)}\|]]$
MMD OvA	$\mathbb{E}_{p_\theta(y)}\left[\mathbb{E}_{\boldsymbol{x}_a,\boldsymbol{x}_b\sim p_{\text{data}}(\boldsymbol{x})}\left[k(\boldsymbol{x}_a,\boldsymbol{x}_b)\left(\frac{p_\theta(y_a\|\boldsymbol{x})p_\theta(y_b\|\boldsymbol{x})}{p_\theta(y)^2} + 1 - 2\frac{p_\theta(y_a\|\boldsymbol{x})}{p_\theta(y)}\right)\right]^{\frac{1}{2}}\right]$
MMD OvO	$\mathbb{E}_{y_a,y_b\sim p_\theta(y)}\left[\mathbb{E}_{\boldsymbol{x}_a,\boldsymbol{x}_b\sim p_{\text{data}}(\boldsymbol{x})}\left[k(\boldsymbol{x}_a,\boldsymbol{x}_b)\left(\frac{p_\theta(y_a\|\boldsymbol{x}_b)p_\theta(y_a\|\boldsymbol{x}_b)}{p_\theta(y_a)^2} + \frac{p_\theta(y_b\|\boldsymbol{x}_a)p_\theta(y_b\|\boldsymbol{x}_b)}{p_\theta(y_b)^2} - 2\frac{p_\theta(y_a\|\boldsymbol{x}_a)p_\theta(y_b\|\boldsymbol{x}_b)}{p_\theta(y_a)p_\theta(y_b)}\right)\right]^{\frac{1}{2}}\right]$
Wasserstein OvA	$\mathbb{E}_{p_\theta(y)}\left[\mathcal{W}_c\left(\sum_{i=1}^N m_i^y\delta_{\boldsymbol{x}_i}, \sum_{i=1}^N \frac{1}{N}\delta_{\boldsymbol{x}_i}\right)\right]$
Wasserstein OvO	$\mathbb{E}_{y_a,y_b\sim p_\theta(y)}\left[\mathcal{W}_c\left(\sum_{i=1}^N m_i^{y_a}\delta_{\boldsymbol{x}_i}, \sum_{i=1}^N m_i^{y_b}\delta_{\boldsymbol{x}_i}\right)\right]$

the squared Hellinger distance present different advantages among f-divergences. First of all, both of them are bounded between 0 and 1. It is consequently easy to check when any GEMINI using those is maximised contrarily to the MI that is bounded by the minimum of the entropies of $\boldsymbol{x}$ and y (Gray and Shields, 1977). When used as distance between data conditional distribution $p_\theta(\boldsymbol{x}|y)$ and data distribution $p_{\text{data}}(\boldsymbol{x})$, we can apply Bayes' theorem in order to get an estimable equation to maximise, which only involves cluster assignment $p_\theta(y|\boldsymbol{x})$ and marginals $p_\theta(y)$.

MMD GEMINIs: The MMD corresponds to the distance between the respective expected embedding of samples from distribution p and distribution q in a reproducible kernel hilbert space (RKHS) $\mathcal{H}$:

$$\text{MMD}(p\|q) = \|\mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}[\varphi(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}[\varphi(\boldsymbol{z})]\|_{\mathcal{H}}, \tag{9}$$

where φ is the RKHS embedding. To compute this distance we can use the kernel trick (Gretton et al., 2012) by involving the kernel function $\kappa(\boldsymbol{a},\boldsymbol{b}) = \langle\varphi(\boldsymbol{a}),\varphi(\boldsymbol{b})\rangle$. We use Bayes' theorem to uncover a version of the MMD that can be estimated through Monte Carlo using only the predictions $p_\theta(y|\boldsymbol{x})$.

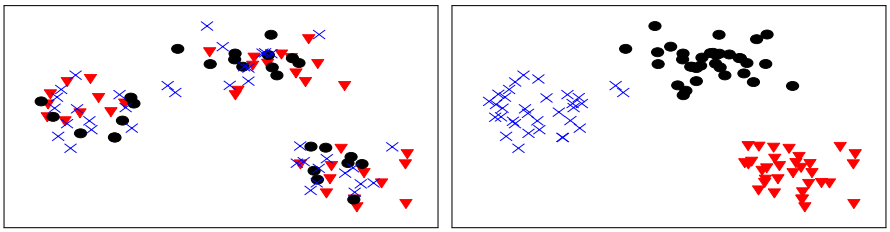
Wasserstein GEMINIs: This distance is an optimal transport distance, defined as:

$$\mathcal{W}_c(p,q) = \left(\inf_{\gamma\in\Gamma(p,q)} \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}\sim\gamma(\boldsymbol{x},\boldsymbol{z})}\left[c(\boldsymbol{x},\boldsymbol{z})\right]\right), \tag{10}$$

where $\Gamma(p,q)$ is the set of all couplings between p and q and c a distance function in $\mathcal{X}$. Computing the Wasserstein distance between two distributions $p_\theta(\boldsymbol{x}|y=k_1)$ and $p_\theta(\boldsymbol{x}|y=k_2)$ is difficult in our discriminative context because we only have access to a finite set of samples N. To achieve the Wasserstein-GEMINI, we instead use approximations of the distributions with weighted sums of Diracs:

$$p_\theta(\boldsymbol{x}|y=k) \approx \sum_{i=1}^N m_i^k\delta_{\boldsymbol{x}_i} = p_N^k, \quad \text{with} \quad m_i^k = \frac{p_\theta(y=k|\boldsymbol{x}_i)}{\sum_{j=1}^N p_\theta(y=k|\boldsymbol{x}_j)}, \tag{11}$$

where $\delta_{\boldsymbol{x}_i}$ is a Dirac located on sample location $\boldsymbol{x}_i \in \mathcal{X}$. The Wasserstein-OvA and -OvO applied to Dirac sums are compatible with the `emd2` function of the python optimal transport package (Flamary



(a) KL-OvA (MI) (b) MMD OvA with linear kernel

Figure 2: Clustering of a mixture of 3 Gaussian distributions with MI (left) and a GEMINI (right) using categorical distributions. The MI does not have insights on the data shape because of the model, and clusters points uniformly between the 3 clusters (black dots, red triangles and blue crosses) whereas the MMD is aware of the data shape through its kernel.

et al., 2021) which gracefully supports automatic differentation (see Appendix B for convergence). All GEMINIs are summarised in Table 1, (see Appendix D for derivations).

4 Experiments

For all experiments below, we report the adjusted rand index (ARI) (Hubert and Arabie, 1985), a common metric in clustering. This metric is external as it requires labels for evaluation. It ranges from 0, when labels are independent from cluster assignments, to 1, when labels are equivalent to cluster assignments up to permutations. An ARI close to 0 is equivalent to the best accuracy when voting constantly for the majority class, e.g. 10% on a balanced 10-class dataset. Regarding the MMD- and Wasserstein-GEMINIs, we used by default a linear kernel and the Euclidean distance unless specified otherwise. All discriminative models are trained using the Adam optimiser (Kingma and Ba, 2014). We estimate a total of 450 hours of GPU consumption. (See Appendix H for the details of python packages for experiments)

4.1 When the MI fails because of the modelling

We first took the most simple discriminative clustering model, where each cluster assignment according to the input datum follows a categorical distribution:

$$y|\boldsymbol{x} = \boldsymbol{x}_i \sim \text{Cat}(\theta_i^1, \theta_i^2, \cdots, \theta_i^K).$$

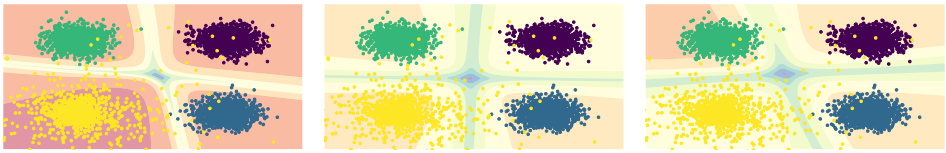
We generated $N = 100$ samples from a simple mixture of $K = 3$ Gaussian distributions. Each model thus only consists in NK parameters to optimise. This is a simplistic way of describing the most flexible deep neural network. We then maximised on the one hand the KL-OvA (MI) and on the other hand the MMD-OvA. Both clustering results can be seen in Figure 2. We concluded that without any function, e.g. a neural network, to link the parameters of the conditional distribution with $\boldsymbol{x}$, the MI struggles to find the correct decision boundaries. Indeed, the position of $\boldsymbol{x}$ in the 2D space plays no role and the decision boundary becomes only relevant with regards to cluster entropy maximisation: a uniform distribution between 3 clusters. However, it plays a major role in the kernel of the MMD-GEMINI thus solving correctly the problem.

4.2 Resistance to outliers

To prove the strength of using neural networks for clustering trained with GEMINI, we introduced extreme samples in Gaussian mixtures by replacing a Gaussian distribution with a Student-t distribution for which the degree of freedom ρ is small. We fixed $K = 4$ clusters, 3 being drawn from multivariate Gaussian distributions and the last one from a multivariate Student-t distribution in 2 dimensions for visualisation purposes with 1 degree of freedom (see AppendixE for other parameters and results). Thus, the Student-t distribution has an undefined expectation and produces samples that can be perceived as outliers regarding a Gaussian mixture owing to its heavy tail. We report the ARIs of multi-layered perceptron (MLP) trained 20 times with GEMINIs in Table 2. The presence of "outliers" leads K-Means and Gaussian Mixture models to fail at grasping the 4 distributions when tring to

Table 2: Mean ARI (std) of a MLP fitting a mixture of 3 Gaussian and 1 Student-t multivariate distributions compared with Gaussian Mixture Models and K-Means. The models try to find 4 at best and the Student-t distribution has ρ=1 degree of freedom. We write the ARI for the maximum a posteriori of an oracle aware of all parameters of the data.

K-Means	GMM		MMD		Wasserstein	
	full cov	diagonal cov	$\mathcal{I}^{\text{ova}}_{\text{MMD}}$	$\mathcal{I}^{\text{ovo}}_{\text{MMD}}$	$\mathcal{I}^{\text{ova}}_{\mathcal{W}}$	$\mathcal{I}^{\text{ova}}_{\mathcal{W}}$
0 (0)	0 (0)	0.024 (0.107)	0.922 (0.004)	0.921 (0.007)	0.470 (0.048)	0.922 (0.006)
Oracle	f-divergences					
	$\mathcal{I}^{\text{ova}}_{\text{KL}}$	$\mathcal{I}^{\text{ovo}}_{\text{KL}}$	$\mathcal{I}^{\text{ova}}_{\text{H}^2}$	$\mathcal{I}^{\text{ovo}}_{\text{H}^2}$	$\mathcal{I}^{\text{ova}}_{\text{TV}}$	$\mathcal{I}^{\text{ovo}}_{\text{TV}}$
0.989	**0.939 (0.006)**	0.723 (0.114)	0.906 (0.103)	0.858 (0.143)	0.904 (0.104)	**0.938 (0.005)**



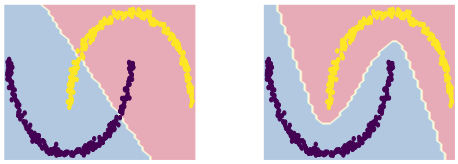
(a) KL-OvA (MI) Entropy Map (b) MMD-OvO Entropy Map (c) Wasserstein-OvO Entropy Map

Figure 3: Entropy maps of the predictions of each MLP trained using a GEMINI or the MI. The bottom-right distribution (yellow) is a Student-t distribution with 1 degree of freedom that produces samples far from the origin. The Rényi entropy of prediction is highlighted from lowest (red background) to highest (blue background). MI on the left has the most confident predictions overall and the smallest uncertainty around the decision boundary, i.e. high entropy variations.

find 4 clusters. Meanwhile, GEMINIs perform better. Note that the MMD and Wasserstein-GEMINI present lower standard deviation for high scores compared to f-divergence GEMINIs. We attribute these performances to both the MLP that tries to find separating hyperplanes in the data space and the absence of hypotheses regarding the data. Moreover, as mentioned in section 2.3, the usual MI is best maximised when its decision boundary presents little entropy $\mathcal{H}(p_\theta(y|\boldsymbol{x}))$. As neural networks can be overconfident (Guo et al., 2017), MI is likely to yield overconfident clustering by minimizing the conditional entropy. We highlight such behaviour in Figure 3 where the Rényi entropy (Rényi, 1961) associated to each sample in the MI (Figure 3a) is much lower, if not 0, compared to MMD-OvO and Wasserstein-OvO (figures 3b and 3c). We conclude that Wasserstein- and MMD-GEMINIs train neural networks not to be overconfident, hence yielding more moderate distributions $p_\theta(y|\boldsymbol{x})$.

4.3 Leveraging a manifold geometry

We highlighted that MI can be maximised without requiring to find the suitable decision boundary. Here, we show how the provided distance to the Wasserstein-OvO GEMINI can leverage appropriate clustering when we have a good a priori on the data.



(a) $\mathcal{I}^{\text{ova}}_{\text{kl}}$ for 2 clusters (MI) (b) $\mathcal{I}^{\text{ovo}}_{\mathcal{W}}$ for 2 clusters

Figure 4: Fitting hand-generated moons using the GEMINI on top of an MLP for different amount of clusters.

The importance of the distance c: We generated a dataset consisting of two facing moons on which we trained a MLP using either the MI or the Wasserstein-OvO GEMINI. To construct a distance c for the Wasserstein distance, we derived a distance from the Floyd-Warshall algorithm (Warshall, 1962; Roy, 1959) on a sparse graph describing neighborhoods of samples. This distance describes how many neighbors are in between two samples, further details are provided in appendix G. We report the different decision boundaries in Figure 4.

Table 3: ARI for deep neural network trained with GEMINIs on MNIST for 500 epochs. Models were trained either with either 10 clusters to find or 15. We indicate in parentheses the number of used clusters by the model after training.

GEMINI		10 clusters		15 clusters	
		MLP	LeNet-5	MLP	LeNet-5
KL	OvA	0.320 (10)	0.138 (8)	0.271 (15)	0.136 (12)
	OvO	0.348 (7)	0.123 (4)	0.333 (8)	0.104 (4)
Squared Hellinger	OvA	0.301 (10)	0.207 (6)	0.224 (13)	0.162 (7)
	OvO	0.287 (10)	0.161 (6)	0.305 (13)	0.157 (7)
TV	OvA	0.299 (10)	0.171 (6)	0.277 (15)	0.140 (6)
	OvO	0.422 (10)	0.161 (9)	0.330 (15)	0.182 (14)
MMD	OvA	0.373 (10)	0.382 (10)	0.345 (15)	0.381 (15)
	OvO	0.361 (10)	0.373 (10)	0.364 (15)	0.379 (15)
Wasserstein	OvA	0.135 (2)	0.127 (4)	0.114 (2)	0.118 (2)
	OvO	**0.450** (10)	**0.383** (10)	**0.415** (15)	**0.414** (15)
K-Means		0.367		0.385	

We observe that the insight on the neighborhood provided by our distance c helped the MLP to converge to the correct solution with an appropriate decision boundary unlike the MI. Note that the usual Euclidean distance in the Wasserstein metric would have converged to a solution similar to the MI. Indeed for 2 clusters, the optimal transport plan has a larger value, using a distribution similar to Figure 4a, than in Figure 4b. This toy example shows how an insightful metric provided to the Wasserstein distance in GEMINIs can lead to correct decision boundaries while only designing a discriminative distribution $p_\theta(y|\boldsymbol{x})$ and a distance c.

Not using all clusters In addition, we highlight an interesting behaviour of all GEMINIs. During optimisation, it is possible that the model converges to using fewer clusters than the number to find. For example in Figure 5, for 5 clusters, the model can converge to 4 balanced clusters and 1 empty cluster (Figure 5b) unlike MI that produced 5 misplaced clusters (Figure 5a). Indeed, the entropy on the cluster proportion in the MI forces to use the maximum number of clusters.

(a) $\mathcal{L}_{\text{kl}}^{\text{ova}}$ for 5 clusters (MI) (b) $\mathcal{L}_{\mathcal{W}}^{\text{ova}}$ for 5 clusters

Figure 5: The Wasserstein-ovo model with 5 clusters eventually found 4 unlike the MI that maintained 5 clusters.

4.4 Fitting MNIST

We trained a neural network using either MI or GEMINIs. Following Hu et al. (2017), we first tried with a MLP with one single hidden layer of dimension 1200. To further illustrate the robustness of the method and its adaptability to other architectures, we also experimented using a LeNet-5 architecture (LeCun et al., 1998) since it is adequate to the MNIST dataset. We report our results in Table 3. Since we are dealing with a clustering method, we may not know the number of clusters a priori in a dataset. The only thing that can be said about MNIST is that there are *at least* 10 clusters, one per digit. Indeed, writings of digits could differ leading to more clusters than the number of classes. That is why we further tested the same method with 15 clusters to find in Table 3. We first see that the scores of the MMD and Wasserstein GEMINIs are greater than the MI, except for the Wasserstein-OvA which systematically converged to using only 2 clusters. We cannot explain why this behaviour was observed in all following experiments. Therefore we stopped experimenting with it for the remainder of this paper. We also observe that no f-divergence-GEMINI always yield best ARIs. Nonetheless, we observe better performances in the case of the TV GEMINIs owing to its bounded gradient. This results in controlled stepsize when doing gradient descent contrarily to KL- and squared Hellinger-GEMINIs. Notice that the change of architecture from a MLP to a LeNet-5

Table 4: ARI score of models trained for 200 epochs on CIFAR10 with different architectures using GEMINIs. The kernel for the MMD is either a linear kernel or the dot product between features extracted from a pretrained SIMCLR model. Both the Euclidean norm between images and SIMCLR features are considered for the Wasserstein metric. We report the ARI of related works when not using data augmentation for comparison.*: scores reported from Li et al., (2021)

Architecture	No kernel	Linear kernel / ℓ_2 norm			SIMCLR (Chen et al., 2020)		
	$\mathcal{I}_{\text{KL}}^{\text{ova}}$	$\mathcal{I}_{\text{MMD}}^{\text{ova}}$	$\mathcal{I}_{\text{MMD}}^{\text{ovo}}$	$\mathcal{I}_{\mathcal{W}}^{\text{ovo}}$	$\mathcal{I}_{\text{MMD}}^{\text{ova}}$	$\mathcal{I}_{\text{MMD}}^{\text{ovo}}$	$\mathcal{I}_{\mathcal{W}}^{\text{ovo}}$
LeNet-5	0.026	0.049	0.048	0.041	**0.157**	0.145	0.138
Resnet-18	0.008	0.047	0.044	0.036	0.122	**0.145**	0.080
KMeans (images / SIMCLR)		0.041	0.147	CC (Li et al., 2021)		0.030	
IDFD (Tao et al., 2021)		0.060		JULE (Yang et al., 2016)*		0.138	

unexpectedly halves the scores for the f-divergences. We believe this drop is due to the change of notion of neighborhood implied by the network architecture.

4.5 Cifar10 clustering using a SIMCLR-derived kernel

To further illustrate the benefits of the kernel or distance provided to GEMINIs, we continue the same experiment as in section 4.4. However, we focus this time on the CIFAR10 dataset. As improved distance, we chose a linear kernel and ℓ_2 norm between features extracted from a pretrained SIMCLR model (Chen et al., 2020). We provide results for two different architectures: LeNet-5 and ResNet-18 both trained from scratch on raw images, the latter being a common choice of models in deep clustering literature (Van Gansbeke et al., 2020; Tao et al., 2021). We report the results in Table 4 and provide the baseline of MI. We also write the baselines from related works when not using data augmentations to make a fair comparison. Indeed, models trained with GEMINIs do not use data augmentation: only the architecture and the kernel or distance function in the data space plays a role. We observe here that the choice of kernel or distance can be critical in GEMINIs. Indeed, while the Euclidean norm between images does not provide insights on how images of cats and dogs are far as shown by K-Means, features derived from SIMCLR carry much more insight on the proximity of images. This shows that the performances of GEMINIs depend on the quality of distance functions. Interestingly, we observe that for the Resnet-18 using SIMCLR features to guide GEMINIs was not as successful as it has been on the LeNet-5. We believe that the ability of this network to draw any decision boundary makes it equivalent to a categorical distribution model as in Sec. 4.1. Finally, to the best of our knowledge, we are the first to train from scratch a standard discriminative neural network on CIFAR raw images without using labels or direct data augmentations, while getting sensible clustering results. However, other recent methods achieve best scores using data augmentations which we do not (Park et al., 2021).

5 Conclusion

We highlighted that the choice of distance at the core of MI can alter the performances of deep learning models when used as an objective for a deep discriminative clustering. We first showed that MI maximisation does not necessarily reflect the best decision boundary in clustering. We introduced the GEMINI, a method which only needs the specification of a neural network and a kernel or distance in the data space. Moreover, we showed how the notion of neighborhood built by the neural network can affect the clustering, especially for MI. To the best of our knowledge, this is the first method that trains single-stage neural networks from scratch using neither data augmentations nor regularisations, yet achieving good clustering performances. We emphasised that GEMINIs are only searching for a maximum number of clusters: after convergence some may be empty. However, we do not have insights to explain this convergence which is part of future work. Finally, we introduced several versions of GEMINIs and would encourage the MMD-OvA as a default choice, since it proves to both incorporate knowledge from the data using a kernel while remaining the less complex than MMD-OvO and Wasserstein-OvO in time and memory. OvO versions could be privileged for fine-tuning steps. Future works could include an optimisation of the time performances of the Wasserstein-OvO to make it more competitive.

References

M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press, 2019.

J. Bridle, A. Heading, and D. MacKay. Unsupervised classifiers, mutual information and 'phantom targets. In J. Moody, S. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1992. URL `https://proceedings.neurips.cc/paper/1991/file/a8abb4bb284b5b27aa7cb790dc20f80b-Paper.pdf`.

M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

J. Chang, L. Wang, G. Meng, S. Xiang, and C. Pan. Deep Adaptive Image Clustering. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5880–5888, 2017. doi: 10.1109/ICCV.2017.626.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

A. Corduneanu and T. S. Jaakkola. On information regularization. *arXiv preprint arXiv:1212.2466*, 2012.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

Z. Dang, C. Deng, X. Yang, K. Wei, and H. Huang. Nearest Neighbor Matching for Deep Clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13693–13702, 2021.

K. Do, T. Tran, and S. Venkatesh. Clustering by maximizing mutual information across views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9928–9938, 2021.

R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL `http://jmlr.org/papers/v22/20-451.html`.

K. Ghasedi Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *Proceedings of the IEEE international conference on computer vision*, pages 5736–5745, 2017.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

R. M. Gray and P. C. Shields. The maximum mutual information between two random processes. *Information and Control*, 33(4):273–280, 1977.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. Publisher: JMLR. org.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bklr3j0cKX.

W. Hu, T. Miyato, S. Tokui, E. Matsumoto, and M. Sugiyama. Learning Discrete Representations via Information Maximizing Self-Augmented Training. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1558–1567, International Convention Centre, Sydney, Australia, Aug. 2017. PMLR. URL http://proceedings.mlr.press/v70/hu17b.html.

J. Huang, S. Gong, and X. Zhu. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8849–8858, 2020.

L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

M. Jabi, M. Pedersoli, A. Mitiche, and I. B. Ayed. Deep clustering: On the link between discriminative models and K-means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. doi: 10.1109/TPAMI.2019.2962683.

X. Ji, J. F. Henriques, and A. Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

J. Kleinberg. An Impossibility Theorem for Clustering. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2003. URL https://proceedings.neurips.cc/paper/2002/file/43e4e6a6f341e00671e123714de019a8-Paper.pdf.

A. Krause, P. Perona, and R. Gomes. Discriminative Clustering by Regularized Information Maximization. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/42998cf32d552343bc8e460416382dca-Paper.pdf.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng. Contrastive clustering. In *2021 AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long. A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture. *IEEE Access*, PP:1–1, 2018. doi: 10.1109/ACCESS.2018.2855437.

T. Minka. Discriminative models, not discriminative training. Technical report, Technical Report MSR-TR-2005-144, Microsoft Research, 2005.

T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. Publisher: IEEE.

S. Park, S. Han, S. Kim, D. Kim, S. Park, S. Hong, and M. Cha. Improving Unsupervised Image Clustering With Robust Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12278–12287, 2021.

A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press, 1961.

M. Ronen, S. E. Finder, and O. Freifeld. Deepdpm: Deep clustering with an unknown number of clusters, 2022. URL https://arxiv.org/abs/2203.14309.

B. Roy. Transitivité et connexité. *Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences*, 249(2):216–218, 1959.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet. On integral probability metrics,$\backslash$phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Y. Tao, K. Takagi, and K. Nakata. Clustering-friendly Representation Learning via Instance Discrimination and Feature Decorrelation. *arXiv preprint arXiv:2106.00131*, 2021.

M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

A. Van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018.

W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*, pages 268–285. Springer, 2020.

S. Warshall. A theorem on boolean matrices. *Journal of the ACM (JACM)*, 9(1):11–12, 1962.

J. Xie, R. Girshick, and A. Farhadi. Unsupervised Deep Embedding for Clustering Analysis. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 478–487, New York, New York, USA, June 2016. PMLR. URL http://proceedings.mlr.press/v48/xieb16.html.

J. Yang, D. Parikh, and D. Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5147–5156, 2016.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We clearly state that we focus on mutual information for clustering and how we improve it.
   (b) Did you describe the limitations of your work? [Yes] We describe shortcomings in Sec. 4.5,5 and provide more in App. F.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Yes, we confirm having read the ethics review guidelines and conform to it.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We clearly highlighted that we compute the GEMINI between a continuous variable, namely the data, and a discrete variable, namely the cluster assignment. The framework was detailed precisely in Sec. 3.1 and all distances have been introduced.
   (b) Did you include complete proofs of all theoretical results? [Yes] Please refer to appendices A, B or D for proofs of our claims.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Link to github will be available upon release of paper and a sample is zipped aside in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Data generation processes are reported in supplementary materials and main parameters are described at the head of Sec. 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] When applicable due to multiple runs in table form, e.g. Table 3.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We mention it at the beginning of Sec. 4
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] All packages for the experiments are mentionned in Appendix H
   (b) Did you mention the license of the assets? [Yes] We are under licenses of MIT (NumPy, POT,), BSD (torch,pandas, scikit-learn) and the python software foundation (matplotlib, python).
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The code once released will contain the specific file `losses.py` that will contain only the definitions of GEMINIs, alongside all configurations to reproduce the experiments.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]