Trimblr - Reading Exercise

1 Chosen paper

I choose to read **Your Diffusion Model is Secretly a Zero-Shot Classifier** (Li et al. 2023) available on ArchiX at `https://arxiv.org/abs/2303.16203` and presented at ICCV conference.

2 Paper summary

The authors propose **Diffusion Classifier**, a novel method to exploit the modelization of underlying data distribution made by diffusion models to perform classification without any additional training (zero-shot context).

The authors point out that discriminative and generative approaches are two different paradigms. Previous works proposed the use of generative models to learn data representations and use them to perform discriminative tasks. More recent works show that generative models could learn representation for discriminative tasks. But those works mostly use joint learning for the generative and discriminative tasks or are based on fine-tuning. Using generative models directly for discriminative tasks, such as image classification, is less studied.

The authors propose to use diffusion models, which are currently popular for creating and editing content, as image classifiers. They want to output their zero-shot image classification capabilities residing in their abilities to learn robust representations and generalize them.

Their approach consists of calculating class conditional density estimates. It exploits the training objective of diffusion models (minimizing the variational lower bound -ELBO- of the log-likelihood) and Bayes theorem to approach the prediction of $p(c|x)$ (probability that sample x belongs to class c). A Monte Carlo estimate of each expectation -class- is computed by sampling fixed pairs of diffusion model's timesteps and Gaussian noise. Using those pairs, errors are calculated for each possible class c. The errors are calculated using either L1 or L2 norm. The class c with the minimal mean error over the timesteps/noise pairs is returned.

The authors tried using different timestep sampling strategies, with the best accuracy obtained when sampling uniformly from the full range of timesteps.

After splitting the evaluation process into different stages, they propose to "eliminate" the classes with the highest average errors in each stage to improve their approach efficiency. The next stages of the evaluation will only be performed on the remaining classes, and the total number of trials performed on those remaining classes can be augmented through different stages.

The authors mentioned that their method could be improved to be more computationally efficient, as it can still be considered slow/a bottleneck when realizing inference.

Methods such as pruning/weak discriminative models, parallelization, or different architectures are evoked in the appendix and conclusion.

For zero-shot classification evaluation, a Diffusion Classifier was implemented with Stable Diffusion (SD) 2.0 trained on a filtered subset of LAION-5B in the experiments. The authors compared their approach with CLIP ResNET-50 and OpenCLIP viT-H/14 discriminative zero-shot models. It is also compared with more similar models (dataset and architecture-wise):

- a ResNet-50 classifier trained on SD synthetic data using class names as prompts.
- a ResNet-50 classifier built on top of SD features (mid-layer) using labeled dataset images/classnames (not zero-shot).

Multiple datasets, such as Food-101, CIFAR-10, ImageNet, etc., were used.

Their approach outperforms the SD synthetic data baseline (likely because this model is trained on features that do not transfer to real data) and CLIP ResNet-50 on all datasets as well as SD features baseline on most datasets. Remember, their approach doesn't need additional training. It is also competitive compared to OpenCLIP viT-H/14. The authors notice that the datasets used for those evaluations may be out-of-distribution for the SD model used in their approach. The authors note that using larger diffusion models trained on more data could improve zero-shot classification results.

The compositional reasoning of their approach is also evaluated on the Winoground dataset and compared to previously used CLIP models. The Winoground dataset is composed of pairs of images associated to captions with the same sets of words in a different order and tests the model's ability to match captions to the corresponding image. Diffusion Classifier outperforms both models on this dataset. This seems linked to a better cross-model binding of concepts to images in the SD model.

For supervised classification, a Diffusion Classifier was built on top of a Diffusion Transformer DiT-XL/2 trained only on ImageNet-1k. It is compared to multiple ResNet, ViT-L, and ViT-B models trained on the same dataset. Their classification ability on in-distribution (ImageNet) and out-of-distribution (variations of ImageNet) datasets are evaluated.

On ImageNet-A, the Diffusion Classifier outperforms the others models, most models on ImageNet and ImageNet-V2 but far less models on ObjectNet. Diffusion Classifier appears to have stronger out-of-distribution performances than most discriminative approaches. Contrary to discriminative approaches, Diffusion Classifier doesn't follow the linear relationship between their in-distribution and out-of-distribution accuracy and achieves state-of-the-art effective robustness to distribution shift.

The authors suggest that their approach could be a solution for scaling up training to the largest models without overfitting or numerical instability (NaNs).

3 Why did it interest me?

Nowadays, diffusion models are gaining increasing popularity due to their ability to generate diverse content. A growing number of diffusion models, with remarkable capabilities, are becoming accessible.

While certain strategies have been suggested for leveraging pre-trained diffusion models in a discriminative context, they often require additional training of these pre-trained models (for example, through fine-tuning). In contrast, this innovative approach proposes the utilization of pre-trained diffusion models without any additional training, relying on their inherent capacity to represent the data on which they were originally trained.

This novel approach has the potential to open the way for the development of new and improved image classifiers and may change the way we currently perform image classification.