On Diverse Asynchronous Activity Anticipation

He Zhao and Richard P. Wildes

York University, Toronto, Ontario, Canada
{zhufl, wildes}@cse.yorku.ca

1 Examination of the regularizer via a toy example

Work that has sought to mitigate mode collapse in the *continuous domain* (*e.g.* [4,5,1,6,2]) typically examines the proposed approach on some simple synthetic datasets to provide a sanity check. Popular test datasets for such experiments include a "Gaussian Grid" or "Gaussian Circle". Consideration of such datasets allows the effectiveness of the approach to be verified in relative isolation of other issues and thereby focus on the particular abilities to combat mode collapse. With the effectiveness demonstrated in these simplified scenarios, the approach is then applied to more complicated real-life datasets and tasks. To date, however, no analogous experiment has appeared for the *discrete domain*, especially not for discrete GANs.

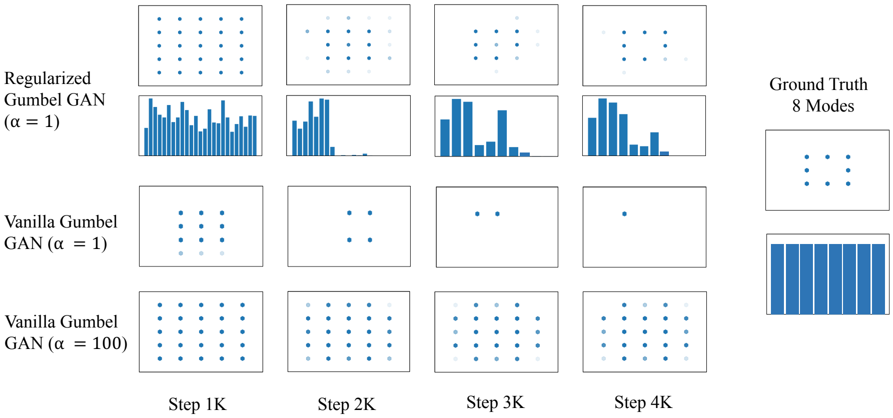

Fig. 1: Simplified demonstration of the quality *vs.* diversity trade-off with Gumbel softmax relaxation (3^{rd} and 4^{th} rows) and the improved result provided by the proposed regularizar (1^{st} and 2^{nd} rows). The first four columns show the evolution of results across training. The far right column shows ground truth. Each bar chart in the 2^{nd} row displays the numeric distribution of modes after every 1K training steps, which ideally should resemble the ground truth bar chart (uniformly distributed among 8 modes), last row of the far right column.

Here, we provide a novel experiment setting for verifying mode-collapse mitigation in the discrete domain and use it to evaluate our approach. Similar to above continuous Gaussian examples, we synthesize a dataset with 8 modes located on a (5×5) 2D grid at (x, y) positions, namely (1, 1), (1, 2), (1, 3), (2, 1), (2, 3), (3, 1), (3, 2), (3, 3). Each mode is given as a unit impulse shifted to one of the specified locations. The rightmost column of Fig. 1 "Ground Truth 8 Modes" provides an illustration.

To simulate a stochastic conditional generation environment, we construct a simple discrete GAN with a two-layer multilayer perceptron (MLP) as generator, that maps the input, $(0.5, 0.5)$, to two logit vectors, that represent the generated (x, y) coordinates. In complement, a separate two-layer MLP is used as a discriminator to provide an adversarial loss for optimization. Standard Gaussian noise is injected into the model during the generative process to provide stochastic capacity. During training, Gumbel-Softmax relaxation, is used to obtain the approximated one-hot vector. We compare results for optimization that solely works under a pure adversarial loss, *i.e.* using Eq. 15 from the main submission, *vs.* results for optimization that augments the advesarial loss with the normalized distance regularizer, *i.e.* using Eq. 16 from the main submission. Figure 1 shows three sets of experimental results to help understand the quality/diversity trade-off issue in discrete GANs and especially to demonstrate the effectiveness of the proposed regularizer in mitigating mode collapse.

Vanilla Gumbel discrete GAN (Gumbel GAN without the normalized distance regularizer) suffers from low diversity with low temperature (*e.g.* for $\alpha = 1$, only a single mode is preserved) and low quality with high temperature (*e.g.* for $\alpha = 100$, we find the presence of false modes along with the correct ones). Overall, either choice leads to undesirable outcomes. In response, the majority of recent alternative approaches seek to find a reasonable balance between quality and diversity, which can be expensive and heuristic; see discussion in main text. Instead, we augment the vanilla approach with the normalized distance regularizer and find that it supports both high quality and diversity, while mitigating mode collapse. This ability is documented in the 4^{th} column of the 1^{st} row of Fig 1, where all 8 modes are captured much better than when the regularizer is lacking (even though not all modes are captured equally).

2 Estimation of loglikelihood via importance sampling

In this section, we describe how to obtain loglikelihood (LL) via **importance sampling** as mentioned in the protocol 2 evaluation of the main submission by following previous work [3].

LL estimation via sampling strategy is necessary when the probability density functions (pdf) of the datasets are unknown. We resort to a Monte-Carlo approach to repeatly sampling next-action labels from the same input obervation and sampled random variables. Thus, LL can be derived from computing $log(p(\theta; y_i)$ for each individual y_i in the test-set and then taking the average.

$$\mathcal{LL}(\theta; \boldsymbol{y}|x, z) = log(p(\theta; \boldsymbol{y}|x_i)) \approx \sum_{i=1}^{N} log(p(\theta; y_i|x_i, z) \tag{1}$$

where $p(\cdot, \theta)$ is our parameterized model for approximating the posterior and N is the number of data in each test set. The base for *log* operator is set as e.

The individual $p(\theta; y_i)$ is also expressed as an empirical mean weighted on the groundtruth using Monte Carlo procedure:

1. Draw M independent values φ_1, φ_2, ..., φ_m from the approximated posterior $p(\cdot; \theta)$

2. Estimate $p(\theta; y_i)$ with

$$p(\theta; y_i|x, z) = \frac{1}{M}\sum_{m=1}^{M} p(y_i|\varphi_m; \theta) = \frac{1}{M}\sum_{m=1}^{M} p(y_i|\varphi_m; \theta)\frac{p(\varphi_m|x, z; \theta)}{q(y_i|x, z)} \tag{2}$$

where $q(y_i|x, z)$ represent the corresponding groundtruth label from test set given input x and z is random variable from standard gaussian distribution.

As a concrete example, for input action *take_cup* and its ground-truth next action *pour_milk*, we sample $M = 16$ times by inputting the same input and sampled random variable, calculate the empirical mean ($e.g.$ 0.45) of the category *pour_milk* and therefore obtain the $\mathcal{LL}$ value ($e.g.$ **-0.7985**). The overall score is the average cross the test set.

Finally, we run above procedures 10 times and use the worst $\mathcal{LL}$ estimation as a lower bound for a fair comparison with variational methods [3].

3 l_1 distance for activity sequences

In this section, we further document how sample distances are calculated for the normalized distance regularizer, Eq. 13 in the main submission.

Let $<\hat{\mathsf{A}}_{\mathsf{n}+1:\tau}\hat{\mathsf{T}}_{\mathsf{n}+1:\tau}> |_{\mathsf{z}_1}$ and $<\hat{\mathsf{A}}_{\mathsf{n}+1:\tau}\hat{\mathsf{T}}_{\mathsf{n}+1:\tau}> |_{\mathsf{z}_2}$ be two samples generated by our generator, $\mathcal{G}$, from input $<\mathsf{A}_{1:n}, \mathsf{T}_{1:n}>$ with two different latent noise signals, z_1 and z_2, resp. Then, Eq. 13 of the main paper is realized by calculating normalized l_1 distances separately between the actions and temporal durations and then combining through averaging according to

$$\begin{aligned}
&||\mathcal{G}(<\mathsf{A}_{1:n}, \mathsf{T}_{1:n}>, z_1) - \mathcal{G}(<\mathsf{A}_{1:n}, \mathsf{T}_{1:n}>, z_2)||_{l_1} \\
&= || <\hat{\mathsf{A}}_{\mathsf{n}+1:\tau}, \hat{\mathsf{T}}_{\mathsf{n}+1:\tau}> |_{z_1} - <\hat{\mathsf{A}}_{\mathsf{n}+1:\tau}, \hat{\mathsf{T}}_{\mathsf{n}+1:\tau}> |_{z_2})||_{l_1} \\
&= \frac{1}{2}\Big(||\hat{\mathsf{A}}_{\mathsf{n}+1:\tau}|_{z_1} - \hat{\mathsf{A}}_{\mathsf{n}+1:\tau}|_{z_2}||_{l_1} + ||\hat{\mathsf{T}}_{\mathsf{n}+1:\tau}|_{z_1} - \hat{\mathsf{T}}_{\mathsf{n}+1:\tau}|_{z_2}||_{l_1} \Big) \\
&= \frac{1}{2}\Big(\frac{1}{\tau - n}\sum_{i=1}^{\tau-n} ||\hat{\mathsf{a}}_i|_{z_1} - \hat{\mathsf{a}}_i|_{z_2}||_{l_1} \\
&+ \frac{1}{\tau - n}\sum_{i=1}^{\tau-n} ||\hat{\mathsf{t}}_i|_{z_1} - \hat{\mathsf{t}}_i|_{z_2}||_{l_1} \Big).
\end{aligned}$$

4 Additional visualization results

In this section, we provide additional visualization results that compare our full approach *vs.* our approach lacking normalized distance regularization, analogous to Figure 5 of the main submission. Figure 2 shows sample sequences generated from an initial observation of Take_cup. For sequences generated without the regularizer, it is seen that generated samples either lack diversity (*e.g.* $\alpha = 1$ always yields the same result) or lack quality (*e.g.* $\alpha = 10$ and $\alpha = 100$ yield sequences that do not follow naturally from the initial observation). In constrast, inclusion of the normalized distance regularizer, *i.e.* our full approach, yields samples that are both diverse and naturalistic.

Figure 3 shows examples generated from an initial observation (Cut_fruit, Peel_fruit), that belongs to a video of preparing salads. This particular category of activity sequence contains a large portion of action pairs that are repetitive and interchangable, *e.g.* (Cut_fruit, Peel_fruit) and (Peel_fruit, Cut_fruit). Therefore, the low temperature model without regularizer (*e.g.* $\alpha = 1$) generates stationery oscillation between Cut_fruit and Peel_fruit, thus losing important modes *e.g.* Put_fruit2bowl or Stir_fruit. For higher temperature cases *e.g.* $\alpha = 10$ and $\alpha = 100$, we find generated examples become more random and unrealistic, similar to the previous two visualization results in Fig 2 of this supplement as well as Fig. 5 of the main submission. Again, our full approach generates reasonable outputs that are close to real samples and includes a wider range of valid modes.

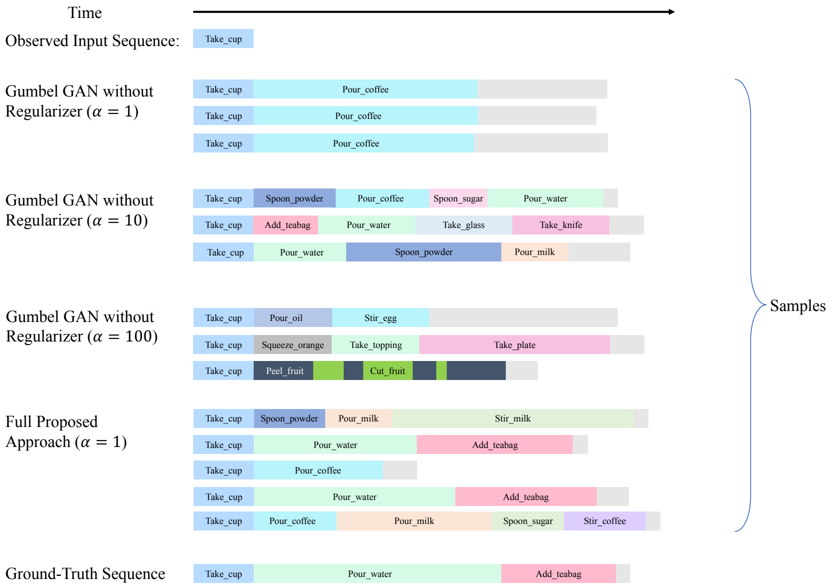

Fig. 2: Visualization results from the Take_cup initial observation. Comparisons include samples from Gumbel GAN without the normalized distance regularizer while temperature ranges $\alpha \in \{1, 10, 100\}$, as well as samples from our full approach that includes the regularizer.

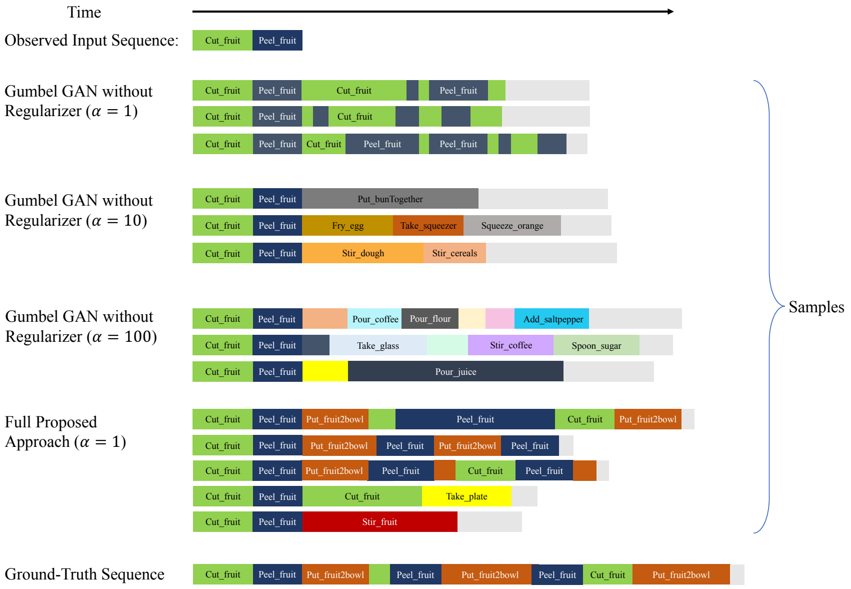

Fig. 3: Results from the (Cut_fruit, Peel_fruit) initial observation. Comparisons include samples from Gumbel GAN without Regularizer with temperature permuting $\alpha \in \{1, 10, 100\}$, as well as samples from our full approach that includes the regularizer.

References

1. Lin, Z., Khetan, A., Fanti, G., Oh, S.: Pacgan: The power of two samples in generative adversarial networks. In: NIPS (2018)
2. Liu, S., Zhang, X., Wangni, J., Shi, J.: Normalized diversification. In: CVPR (2019)
3. Mehrasa, N., Jyothi, A.A., Durand, T., He, J., Sigal, L., Mori, G.: A variational auto-encoder model for stochastic point processes. In: CVPR (2019)
4. Srivastava, A., Valkov, L., Russell, C., Gutmann, M.U., Sutton, C.: Veegan: Reducing mode collapse in gans using implicit variational learning. In: NIPS (2017)
5. Xiao, C., Zhong, P., Zheng, C.: Bourgan: Generative networks with metric embeddings. In: NIPS (2018)
6. Yang, D., Hong, S., Jang, Y., Zhao, T., Lee, H.: Diversity-sensitive conditional generative adversarial networks. In: ICLR (2019)