Context-Aware Variational Autoencoder for Grouped Partial Observations

Anonymous Authors[1]

Abstract

Group-instance disentanglement is the problem of learning separate representations for within-group and across-group variability. The current state-of-the-art methods for solving this problem are based on the Group Variational Autoencoder (GVAE). In this work, we create a synthetic dataset wherein the group and instance variables cannot be inferred accurately from a single datapoint. We use this dataset to show that models from the GVAE family are limited in how they 1) accumulate evidence for computing the group variable, 2) define the variational distribution of the instance variable, and 3) use adversarial training to prevent the instance encoder from encoding group information. We overcome this failure case by modifying the encoder and loss function of the GVAE.

1. Introduction

The task is to learn representations of grouped data (e.g. images grouped by author, temperature readings grouped by weather station, etc) that separate between within-group and across-group variation. We encode the former with a latent *instance* variable v and the latter with a latent *group* variable u.

Group-instance disentanglement is a version of weakly-supervised disentanglement which has attracted significant attention in recent years (Tschannen et al., 2018). State-of-the-art methods for group-instance disentanglement (Bouchacourt et al., 2018; Hosoya, 2019; Shu et al., 2020; Chen & Batmanghelich, 2020; Locatello et al., 2020) follow the Variational Autoencoder framework (Kingma & Welling, 2014; Rezende et al., 2014) whereby a generative model mapping latent variables to data observations is trained using variational inference. We call this family of models the Group Variational Autoencoder (GVAE).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

These methods promise a general approach for amortizing the cost of inference in nonlinear mixed models.

However, the existing models in the GVAE family have been designed and applied on data where only one observation is strictly necessary to accurately infer the group and instance variables. For example, in style-content disentanglement (a popular evaluation setting), the style and content of an image can be ascertained comfortably without any knowledge of the other observations in the group (Zhu et al., 2017; Kotovenko et al., 2019).

Unfortunately, as we will show in this work, these methods fail to learn disentangled representations in settings where the same observation could be present in different groups, and therefore have different underlying instance components. For example, assume the observation is a binary result of a *XOR* operation between the group and instance variables. Then, knowing the observation value gives no information about the instance in the absence of knowledge about the group. We call these settings *entangled*, and they occur abundantly in real-world problems, such as collaborative filtering, novel-view synthesis, and disease progression modelling.

To tackle this challenge, we propose several modifications to the standard GVAE which enable the model to infer the latent components more accurately. We implement the group encoder as a Deep Sets network (Zaheer et al., 2017), condition the instance variable on the inferred group variable (Li & Mandt, 2018), and introduce a novel regularization objective inspired by (Németh, 2020).

Our work comprises the following contributions:

1. We formalize the notion of "entangled" problems and create a simple synthetic dataset on which the current GVAE methods fail to learn disentangled representations.
2. We propose a new model, called the Context-Aware Variational Autoencoder (CxVAE) which can learn disentangled representations in "entangled" settings.
3. We perform extensive evaluation showing that our model outperforms the existing methods in holdout-data log-likelihood, unsupervised translation, and mutual information between the inferred latents and the

ground truth.

2. Generative Group-Instance

The Group Variational Autoencoder (Bouchacourt et al., 2018; Hosoya, 2019) is a family of models that use two latent variables to represent grouped data: one that captures the variation within groups, and one for the variation across groups.

Assume a dataset of the form $\{x_{ik}\}_{i\in 1:N,\ k\in 1:K_i}$ where N is the number of groups and K_i is the number of observations in group i. GVAE defines a generative model that maps a $\mathcal{N}(0,1)$ group latent variable u_i and a $\mathcal{N}(0,1)$ instance latent variable v_{ik} to a given data observation x_{ik}. In other words, the likelihood of a group is:

$$p(\mathbf{x}) = \mathbb{E}_{p(u)} \prod_{k=1}^{K} \mathbb{E}_{p(v_k)} \left[p(x_k|u, v_k)\right]$$

We omit the index of the group i for notational simplicity, since the groups are independent and identically distributed.

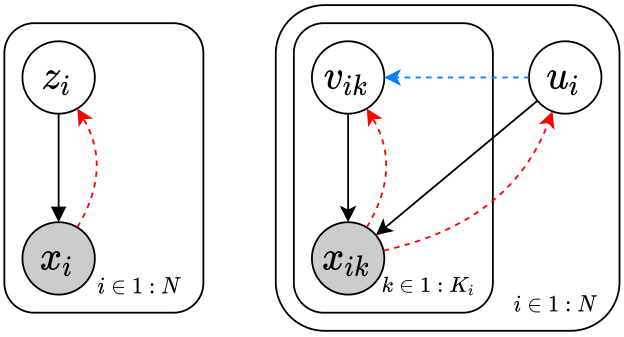

Figure 1. **Probabilistic graphical model of the VAE (left) and the GVAE (right).** The dotted arrows depict the variational latent posterior, and the blue arrow shows a dependency which is absent from the GVAE but present in our proposed model.

2.1. Variational Inference

Because the exact likelihood is intractable, the Variational Autodencoder (Kingma & Welling, 2014; Rezende et al., 2014) performs optimization by introducting a variational latent posterior $q(u, \mathbf{v}|\mathbf{x})$ and maximizing the Evidence Lower Bound (Jordan et al., 2004):

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(u,\mathbf{v}|\mathbf{x})}[\log p(\mathbf{x}|u, \mathbf{v})] - \mathrm{KL}[q(u, \mathbf{v}|\mathbf{x})||p(u, \mathbf{v})]$$

The models in the GVAE family use a class of variational distributions that assume independence between the latent variables in a group when conditioned on the data.

$$q(u, \mathbf{v}|\mathbf{x}) = q(u|\mathbf{x}) \prod_{k=1}^{K} q(v_k|x_k)$$

In our work, we show that this assumption hinders disentanglement when the generative model is entangled.

2.2. Group Encoder

The variational group posterior is realised as normal density with μ, Σ computed with an encoder network. The way to implement the encoder network is not obvious, since the number of inputs K varies across groups. Hosoya (2019) encode separately each observation in the group using the same encoder E_u to produce μ_k, Σ_k and then averages the outputs.

$$\mu = \frac{1}{N}\sum_{k=1}^{K} \mu_k,\ \Sigma = \frac{1}{N}\sum_{k=1}^{K} \Sigma_k$$

Bouchacourt et al. (2018) also encode each observation individually and then accumulate the evidence through a product of normal densities, computed using the following equations:

$$\Sigma^{-1} = \sum_{k=1}^{K} \Sigma_k^{-1},\ \mu^T\Sigma^{-1} = \sum_{k=1}^{K} \mu_k^T\Sigma_k^{-1}$$

They justify that such a product of normals produces a valid evidence accumulation using the following result:

$$q(u|\mathbf{x}) \propto \prod_{k=1}^{K} q(u|x_k)$$

However, the above is not a universal property, since

$$\begin{aligned} q(u|\mathbf{x}) &= \frac{\prod_{k=1}^{K} q(x_k)}{q(\mathbf{x})} \frac{1}{q(u)^{K-1}} \prod_{k=1}^{K} q(u|x_k) \\ &\propto \frac{1}{q(u)^{K-1}} \prod_{k=1}^{K} q(u|x_k) \end{aligned} \tag{1}$$

In fact, by using a product of normals to accumulate evidence, the authors implicitly assume that the marginal distribution of the inferred group variable is a uniform. This has the effect of sampling u values which are less representative of the current group and more skewed towards the marginal distribution of u.

In our work, we propose a more general approach to encoding u by using a Deep Sets network (Zaheer et al., 2017) to

encode the whole set of observations instead of encoding each observation separately.

2.3. Regularization

In certain cases, the model might learn to encode both kinds of variation (within- and across-group) in the instance variable, effectively turning the model into a standard VAE. In this eventuality, the group variable becomes irrelevant and disentanglement is not achieved.

Such behaviour has been identified by Hosoya (2019) and Németh (2020) to occur when the instance code too high-dimensional, the instance encoder too expressive, or group sizes too small. One solution is to limit the dimensionality of the instance code (Hosoya, 2019), with the downside of hindering the overall model performance.

As a more targeted solution, Hosoya (2019) propose an adversarial loss minimizing the mutual information between an observation and the instance variable inferred from the other observations in the group:

$$I_r(x, v) = \mathrm{KL}[r(x, v)||r(x)r(v)]$$

where $r(x, v) = r(v|x)r(x)$ is the joint distribution of an observation and the instance variable inferred from any of the other observations in the group and

$$r(v|x_k) = \frac{1}{K-1} \sum_{l=1,\ l\neq k}^{K} q(v|x_l)$$

The mutual information is approximated empirically using the results of Belghazi et al. (2018).

$$I_r(x, v) \approx \max_T \mathbb{E}_{r(x,v)}[T(x, v)] - \log \mathbb{E}_{r(x)r(v)}[\exp T(x, v)] \tag{2}$$

T is a neural network and the expectation terms are computed by sampling.

- To sample $r(x, v)$, first choose a group i, then choose two instances from that group $k, l \in K_i$. x will be the observation x_{ik} and v will be sampled from $q(v|x_{il})$.
- To sample $r(x)r(v)$, choose two groups i, j and two instances in each group $k \in K_i, l \in K_j$. Take x_{ik} for x and sample $q(v|x_{jl})$.

In our view, this method has the following limitation: Even when the instance variable does contain group information, the value of $I_r(x, v)$ might still be small, because it might be difficult to ascertain the group based on one single observation x. In our work, we propose a modification to this regularization term such that the network T takes as input all the observations in the group instead of only one.

3. Entangled Group and Instance Variables

We call the group and instance variables *entangled* when they are not independent conditioned on the data $p(u, v|x) \neq p(u|x)p(v|x)$. A useful heuristic for establishing whether the variables are entangled is to ask "Does knowing the group variable for an observation influence my belief about its instance variable?"

This property of the generative model is present in many machine learning tasks, such as collaborative filtering, 3D novel view synthesis. For example, in the context of the Netflix Challenge, where the task was to predict what score a user would give to a new film, one cannot infer what film is associated with a given score without also knowing the user.

Strictly speaking, most real-world models are entangled. However, in many cases, the mutual information between the group and instance variable, conditioned on the observation, is negligible. For example, in handwritten digit recognition, one can infer the digit value depicted in an image without knowing the author.

In this paper, we claim that the current methods in the GVAE family do not perform well in tasks where the group and instance variables are entangled.

3.1. Exam-Scores Problem

Suppose we wanted to model the exam scores of students from different schools. Our model must separate the school-level effect (the group factor) from the student-level effect (the instance factor). We define the following generative model:

$$x_{ik} = 2\ \mu_i + (\sigma_i^2 + 1)v_{ik} + \epsilon_{ik},\ i \in 1:N,\ k \in 1:K_i$$

where x_{ij} is the student score, $u_i = (\mu_i, \sigma_i)$ is the school-level effect, v_{ij} is the student-level effect, and ϵ_{ik} is a normally-distributed error term. We assume a $\mathcal{N}(0, 1)$ prior distribution for the latent variables.

We first sample the model to generate a dataset ($N = 32,768,\ K_i \sim \mathrm{Poisson}(16) + 8$) and then use the same model as the generative model in our Variational Autoencoder, instead of a neural network. The figure below shows what the data looks like.

Looking at the data, it is easy to see that this model is entangled, because the relative performance v of a student within

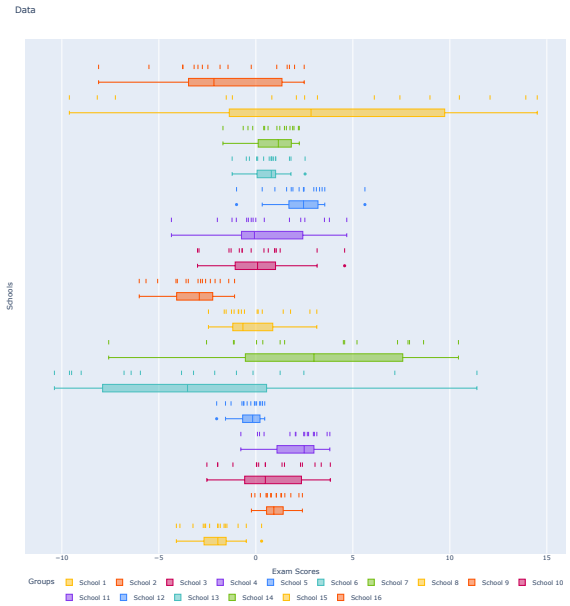

Figure 2. **Normalized exam scores of individual students grouped by school.**

their own school, given their absolute score x, depends on the distribution of scores within the school u.

4. Context-Aware Variational Autoencoder (CxVAE)

We propose a new model which can perform well on datasets generated from entangled group and instance variable. We call our model the Context-Aware Variational Autoencoder. Our model comprises the following changes with respect to the standard GVAE:

1. The group encoder is implemented as a Deep Sets network (Zaheer et al., 2017). The encoder has the following form:

$$\mu, \Sigma = E_u(\mathbf{x}) = E_u^B\left(\sum_{k=1}^{K} E_u^A(x_k)\right)$$

where E_u^A, E_u^B are two neural networks.

2. The variational instance posterior is dependent on the inferred group variable:

$$q(u, \mathbf{v}|\mathbf{x}) = q(u|\mathbf{x}) \prod_{k=1}^{K} q(v_k|x_k, u)$$

In practice, our instance encoder takes as input a vector concatenating x_k and u. This idea is not new, and has been used previously in sequence disentanglement (Li & Mandt, 2018). This allows the instance encoder to differentiate the between observations with similar values but which come from different groups.

3. We propose a regularization objective similar to the one in Németh (2020), but whereby we minimize *the mutual information between one inferred instance variable and all the other observations in the group*. More precisely, our objective is to minimize $I_r(\mathbf{x}_{-k}, v_k)$ where $r(v_k|\mathbf{x}_{-k}) = q(v_k|x_k, u)$. Following the same approximation as in Németh (2020), the objective takes the following form:

$$\begin{aligned} I_r(\mathbf{x}_{-k}, v_k) \approx \max_T \; & \mathbb{E}_{r(\mathbf{x}_{-k}, v_k)}[T(\mathbf{x}_{-k}, v_k)] \\ & - \log \mathbb{E}_{r(\mathbf{x}_{-k})r(v_k)}[\exp T(\mathbf{x}_{-k}, v_k)] \end{aligned} \tag{3}$$

T is implemented as a Deep Sets neural network for the observations, with the instance code concatenated in the middle:

$$T(\mathbf{x}_{-k}, v_k) = T^\beta\left(v_k, \frac{1}{K}\sum_{l=1,\, l\neq k}^{K} T^\alpha(x_l)\right)$$

Again, the expectations are computed by sampling:

- To sample $r(x, v)$, first choose a group i, then choose one instance from that group $k \in K_i$. $\mathbf{x}_{i,-k}$ will be all the observations in the group apart from k, and v_k will be sampled from $q(v|x_{ik})$.
- To sample $r(\mathbf{x}_{-k})r(v_k)$, choose two groups i, j and two instances in each group $k \in K_i$, $l \in K_j$. Take $\mathbf{x}_{i,-k}$ for $\mathbf{x}_{-k}$ and sample $q(v|x_{jl})$.

This regularization objective also minimizes the mutual information between the inferred instance variables and the true data generating group variable, but uses as a proxy for the latter all but one of the observations in the group, instead of just one observation.

5. Measuring Disentanglement

In the context of the GVAE family, disentanglement is a property of the variational latent posterior. The inferred group and instance variables are disentangled when they are maximally informative about the group and instance variables of the true data generating distribution. We assume this true model has the same factorization of the joint distribution as the generative model, but the parameters are unknown.

Therefore, we can use the mutual information between the true and inferred latents as a measure of disentanglement.

$$I(U^t; U^q) = \mathrm{KL}[\Pr_{U^q, U^t} || \Pr_{U^q} \Pr_{U^t}]$$

$$= \mathbb{E}_{U^q}\mathbb{E}_{U^t|U^q}\left[\log \frac{\Pr_{U^t|U^q}}{\Pr_{U^t}}\right]$$

and since $\Pr_{U^t}$ does not depend on q

$$= \mathbb{E}_{U^q}\mathbb{E}_{U^t|U^q}[\log \Pr_{U^t|U^q}] + \text{const}$$

We estimate the density $\Pr_{U^t|U^q}$ by training a regression network to predict the value of U^t given U^q. The mean-squared error of the prediction on the holdout set is a monotonic function of the mutual information with respect to changes in q.

5.1. Unsupervised Translation

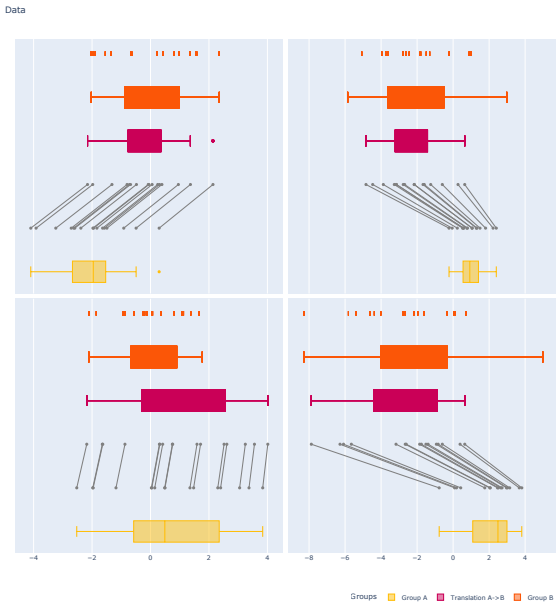

Figure 3. **Example of translation on the Exam-Scores dataset.** The yellow and orange show the distributions of the source and target group, respectively. The red shows the distribution of the translated group. The grey lines show the translation of individual observations from the source to the target. We expect a good translation to have the translation lines uncrossed (the instance variable is preserved) and to have the same distribution as the target group (the group variable is changed).

Unsupervised translation is the process of transforming an observation by changing its group code while keeping its instance code fixed. This is a common downstream task for disentangled representations because it requires a clean separation between the group and instance representations (Tenenbaum & Freeman, 2000). Therefore, it can be used to quantify the quality of disentanglement.

Formally, let i, j be the indices of the source and target group, respectively. Translation involves sampling a group of instance codes from the source group $q(\mathbf{v}_i|\mathbf{x}_i)$ and a group code from the target group $q(u_j|\mathbf{x}_j)$. Then, we generate the translated observations by combining each instance code with the group code $p(x'_k|u_j, \mathbf{v}_j)$.

We measure translation quality by taking the mean-squared error between the translation performed with each model and the ground-truth translation computed using the ground-truth factors and the true data-generating process.

6. Evaluation

We test CxVAE against the other models from the GVAE family.

6.1. Quantitative Evaluation

6.2. Ablation Study

In order to quantify the effect of each proposed improvement, we perform an ablation study whereby we measure the decrease in performance resulting from replacing a proposed element of our model with a current alternative.

The proposed improvements have the best individual performance with respect to both reconstruction and translation error in each of the three categories. The choice of a group-aware instance encoder leads to the most significant increase in performance.

References

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Hjelm, R. D., and Courville, A. C. Mutual information neural estimation. In *ICML*, 2018.

Bouchacourt, D., Tomioka, R., and Nowozin, S. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *AAAI Conference on Artificial Intelligence*, 2018.

Chen, J. and Batmanghelich, K. Weakly supervised disentanglement by pairwise similarities. *AAAI Conference on Artificial Intelligence*, 2020.

Hosoya, H. Group-based learning of disentangled representations with generalizability for novel contents. In *IJCAI*, 2019.

Jordan, M. I., Ghahramani, Z., Jaakkola, T., and Saul, L. K. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 2004.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

Kotovenko, D., Sanakoyeu, A., Lang, S., and Ommer, B. Content and style disentanglement for artistic style trans-

Table 1. **Comparison between errors on the holdout set of our model (CxVAE) and the rest of the GVAE family.** CxVAE has the best performance across the 4 criteria. All errors are MSE, so lower is better. A mean error and standard deviation over the 7 training runs is taken at every epoch, and then the last 20 training epochs (44 - 64) are averaged in order to be displayed.

MODEL	REC ERR	TRANS ERR	u-PRED ERR	v-PRED ERR
CXVAE (OURS)	**52.1** $\pm$ 0.6	**219.5** $\pm$ 4.7	**53.3** $\pm$ 1.2	**30.2** $\pm$ 0.3
ML-VAE (BOUCHACOURT ET AL., 2018)	69.8 $\pm$ 1.6	738.8 $\pm$ 13.6	63.4 $\pm$ 1.0	63.0 $\pm$ 0.3
GVAE (HOSOYA, 2019)	71.9 $\pm$ 4.3	742.9 $\pm$ 14.4	63.1 $\pm$ 0.9	63.0 $\pm$ 0.4
GVAE + REG (NÉMETH, 2020)	70.0 $\pm$ 1.4	735.3 $\pm$ 13.0	63.8 $\pm$ 1.1	63.0 $\pm$ 0.5

Table 2. **Ablation study comparing CxVAE (our model) with alternative models obtained by replacing each of the novel components (group encoder, instance encoder and regularization).** Each replacement leads to a worse performance across all criteria.

MODEL	REC ERR	TRANS ERR	u-PRED ERR	v-PRED ERR
CXVAE (OURS)	**52.1** $\pm$ 0.6	**219.5** $\pm$ 4.7	**53.3** $\pm$ 1.2	**30.2** $\pm$ 0.3
u ENCODER				
AVERAGE	53.6 $\pm$ 2.2	224.7 $\pm$ 10.0	53.4 $\pm$ 1.1	30.9 $\pm$ 1.1
MULTIPLICATION	54.9 $\pm$ 2.8	229.2 $\pm$ 7.8	53.4 $\pm$ 0.9	31.4 $\pm$ 1.0
v ENCODER				
UNCONDITIONAL	72.3 $\pm$ 3.6	729.0 $\pm$ 15.1	63.6 $\pm$ 1.2	63.0 $\pm$ 0.3
REGULARIZATION				
NONE	54.1 $\pm$ 1.9	235.5 $\pm$ 16.5	53.6 $\pm$ 1.4	31.6 $\pm$ 1.5
UNCONDITIONAL IB	53.7 $\pm$ 1.2	225.7 $\pm$ 5.7	53.5 $\pm$ 1.1	30.8 $\pm$ 0.6

fer. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4421–4430, 2019.

Li, Y. and Mandt, S. Disentangled sequential autoencoder. In *ICML*, 2018.

Locatello, F., Poole, B., Rätsch, G., Scholkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Németh, J. Adversarial disentanglement with grouped observations. In *AAAI*, 2020.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

Shu, R., Chen, Y., Kumar, A., Ermon, S., and Poole, B. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgSwyBKvr.

Tenenbaum, J. B. and Freeman, W. T. Separating style and content with bilinear models. *Neural Computation*, 12: 1247–1283, 2000.

Tschannen, M., Bachem, O., and Lucic, M. Recent advances in autoencoder-based representation learning. *Third workshop on Bayesian Deep Learning (NeurIPS)*, 2018.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Póczos, B., Salakhutdinov, R., and Smola, A. Deep sets. In *NIPS*, 2017.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

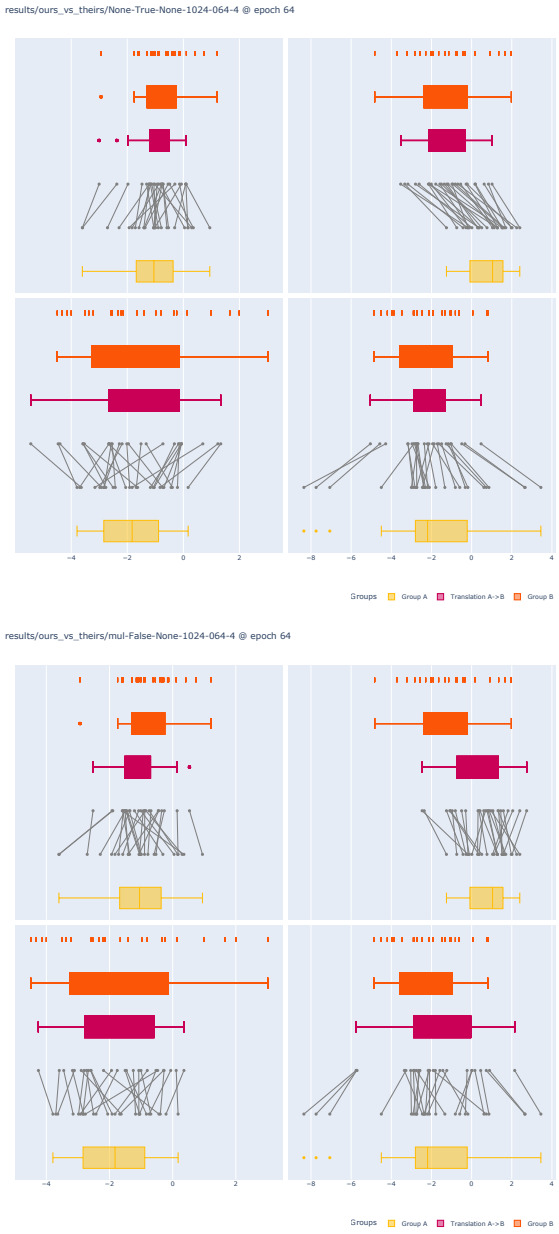

Figure 4. Qualitative comparison of unsupervised translation between CxVAE (top) and ML-VAE (Bouchacourt et al., 2018) (bottom). Our model captures the distribution of values in the target group better than the existing methods.

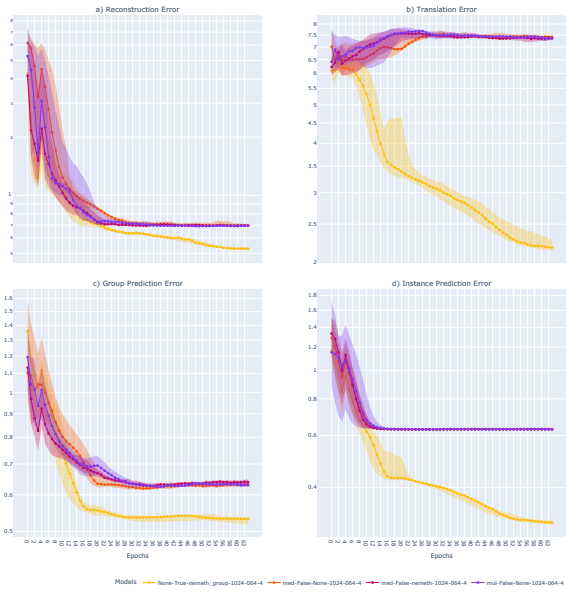

Figure 5. **Errors on the holdout set by training epoch.** We compare CxVAE (ours, yellow) with Bouchacourt et al. (2018) (purple), Hosoya (2019) (orange), and Németh (2020) (red). For each model we perform 7 training runs, displaying the lowest, highest, and median error at each epoch.

A. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one, even using the one-column format.