TNNLS DAAE Addressing Reviews

ac2211

October 2017

Introduction

We would like to thank the reviewers for the clear efforts put into reading, understanding and providing a critique of this draft. We apologise for the missing final page of references, which was an error that crept in on uploading the manuscript. The *original* final page is attached as an appendix to this document, as well as the revised manuscript.

In this response, we have included direct quotations from the three initial reviewers in blue; quotations from our original submission are written in *italic*, and quotations from other peer reviewed or widely cited/accepted work are in .

Reviewer: 1

Comments to the Author

The paper proposes DAAEs, which use adversarial training to match the posterior distribution to the prior for a denoising autoencoder. The proposed approach is like a combination of existing methods like DAEs, AAEs and DVAEs.

Whilst we do not question that there exists a relation of the proposed DAAE and iDAAE to other techniques, we believe that both represent a contribution. The iDAAE and DAAE are as distinct from each other, and from other flavours of autoencoder, as (for example) the AAE, DAE, VAE and DVAE are from each other. To ensure that the message of the paper clearer, we have now modified Figure 1, presenting the models we propose in a different colour. These differences of the DAAE and iDAAE to the DAEs, the VAE, DVAE and AAE are sufficient to warrant study, and in the revised paper we now provide a much more extensive set of experiments (to be discussed later).

Some benefits of the proposed method are highlighted when presenting the method but not well supported in the experiments.

We agree with this statement, and have now performed much more extensive experiments in the widely used celebA dataset. We have performed attribute classification experiments and achieve competitive or superior performance in synthesis and classification compared to the AAE and VAE. In addition, because the field is moving fast, we also provide performance comparisons to the β-VAE and DIP-VAE.

The writing of the paper should be further improved as there is a lot of unclear statements and confusing notations throughout the paper (I list some major issues in the detailed comments.).

We are very grateful to the reviewer for this comment, and the effort spent on identifying exactly were there are opportunities for improvement. We have tried to improve the writing generally, and to address the specific issues identified (see, for example, the next section).

The criteria of evaluation and datasets in the experiment part are not satisfactory. The results are not sufficient for a good publication. I suggest the authors proofread and rewrite the paper and provide strong experimental results to support the motivation of the method.

We have moved the MNIST results to an appendix. Extensive experiments have now been conducted on the CelebA face database. These experiments on image generation (synthesis), strengthen the comparison of both the proposed iDAAE and DAAE to comparable methods. In addition, we demonstrate classification of facial attributes from latent space representations. We provide direct comparisons of the DAAE and iDAAE performance for classification with the AAE as well as the VAE, beta-VAE and DIP-VAE, achieving competitive or superior performance on most attributes. The classification experiments are also very useful for comparing against previous work, and address another issue (see later) regarding the interpretation of performance results for synthesis.

Detailed comments

The paragraph "Two broad approaches ..." (Line 55-60, Column 1, Page 1) should be clarified. For 1), it is better to include some references.

We have added references to three papers that use the denoising criterion to train autoencoders. These are vincent2010stacked, bengio2013generalized, vincent2008extracting. The sentence now begins:

Two broad approaches to learning state-of-the-art generative models that do not require labelled training data include: 1) introduction of a denoising criterion {vincent2010stacked, bengio2013generalized, vincent2008extracting} – where the model learns to reconstruct clean samples from corrupted ones;

For 2), both VAEs [6] and GANs [13] use a latent space with a known prior distribution for generation. However, GANs [13] do not regularize the latent space while VAEs [6] do but directly for the training of the recognition model instead of the generative model.

We understand the concern. Our initial intention in writing the sentence was to point out that VAEs and GANs both have some constraints on latent space: GANs through the latent space arising from noise sources, and VAEs in the sense of having priors placed on latent space. We realise that this is potentially misleading, so we have removed the phrase "latent space with a known prior distribution".

The author should be careful about the correctness of "regularisation ... [1], [9], [13] or using ... [6], [9], [13]" and clarify the writing.

We agree that the sentence was confusing, particularly with regard to the use of the term "regularisation" and the choice of references. We have now altered the sentence to read:

"regularisation of the latent space to match a prior { kingma2013auto, makhzani2015adversarial }"

We would like to make an additional comment here: including GANs was not necessary for

us to convey the key ideas: that denoising and regularisation are core components of state of the art generative models.

The sentences "Autoencoders introduce" (Line 47-60, Column 1, Page 2) should be rewritten. On one hand, not all of the autoencoders are probabilistic. In contrary, for "autoencoders", people refers to the deterministic autoencoders by default.

We agree that not all autoencoders are probabilistic, but they are often referred to in a probabilistic framework, and we add the citation [9] Kingma et al to qualify this. In fact, this is exactly how VAE's in [9] were first introduced. We provide the quotation from [9] for context:

> In this paper we will therefore also refer to the recognition model $q_\phi(z|x)$ as a probabilistic encoder, since given a datapoint x it produces a distribution (e.g. a Gaussian) over the possible values of the code z from which the datapoint x could have been generated. In a similar vein we will refer to $p_\theta(x|z)$ as a probabilistic decoder, since given a code z it produces a distribution over the possible corresponding values of x.

We have purposefully used notation consistent with Kingma et al. [9]. Of course, in a sense, probabilistic autoencoders generalise (and include, as a clear subset) non-probabilistic ones, and our intention was to be *general* in this paper because it allows us to unify the discussion of autoencoders, and helps us to compare different approaches to creating autoencoders.

I also note that subsection C introduces VAEs again, which may be redundant. On the other hand, both the encoders and decoders in VAEs are trained jointly instead of "first learning ... form a training set ... trained in a supervised fashion." I seriously suggest the authors to present the literature carefully and precisely.

We are – of course – aware that the encoder and decoder are trained jointly in **most** generative models. Indeed, we did say this at the end of Section 1A (our original submission):

> In some situations the encoding distribution is chosen rather than learned, in other situations the encoder and decoder are learned simultaneously...

Our reason for explicitly pointing this out is to generalise to the case where the encoding process, was *not* learned. Doing this – which we agree may have been misleading – allows us to include a suggestion (by Bengio, perhaps not very widely referred to) in which encoding is treated as a *local corruption process* which is *not* learned. From our original submission, at the Top of 1B:

> Bengio et al. [4] treat the encoding process as a local corruption process, that does not need to be learned.

In [4], there is no latent space in the same sense as that for VAEs or GANs. To improve the clarity and visibility of this idea, we have added references:

> In some situations the encoding distribution is chosen rather than learned {bengio2013generalized}, in other situations the encoder and decoder are learned simultaneously {kingma2013auto, makhzani2015adversarial, im2017denoising}

Why did we do this? It allowed us to treat encoding in the most general way, so that we can bring the approach suggested in [4] under the same umbrella as other models. Taking this approach not only allows us to acknowledge Bengio's DAE, but also to distinguish between the "corruption" of that model and the noise process of our proposed DAAE more clearly. For example, the approach of [4] does not explicitly define a latent space. It is also worth pointing out

that taking Bengio's corruption process [4] to the extreme is equivalent to having a latent representation which is Gaussian noise {bachman2015variational}; so, this way of thinking seems to be self-consistent and flexible.

I'm confused by the reconstruction cost function (Line 50-54, Column 1, Page 4). The normal definition should be $E_{x\sim p(x)}E_{z\sim q_\phi(z|x)}logp(x|z)$ following VAEs. The author skip this equation but directly present one with Monte Carlo approximation by sampling x and z.

The cost function $E_{x\sim p(x)}E_{z\sim q_\phi(z|x)}\log p(x|z)$ is *not* used in denoising autoencoders: it is typically used in VAEs and AAEs: this expression lacks a corruption process. Because the writing in this section refers to denoising autoencoders, this is not the right expression for L_{rec}.

However, as x is sampled, there should not be a term of $p(x_i)$ in the equation (it should be 1/N, which already exists.). Another typo is "$logp(x_i|z_i)$" instead of "$logp(x|z_i)$". Did I miss something? It should be clarified.

Thanks for spotting this! There was indeed an extra i subscript in the Equation. However, to ensure clarity we have also rewritten the equation and amended the text immediately below it to avoid potential ambiguity:

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_{i=0}^{N-1}\log p_\theta(x|z_i)$$

where the z_i are obtained via the following sampling process $x_{i=0\ldots N-1} \sim p(x)$, $\tilde{x}_i \sim c(\tilde{x}|x_i)$, $z_i \sim q_\phi(z|\tilde{x}_i)$, and $p(x)$ is the distribution of the training data.

The experiments are restricted on simple datasets, which are out of date. In contrary, as one of the main baselines, AAEs are evaluated on faces and color images. Besides, the image quality in this paper is far from state-of-the-art on the simple datasets.

Our choice of datasets was explained in Section VI.B of the original submission. We acknowledge that in one sense the data might be regarded as simple, but please note that

- the sprites dataset (of the orignal paper) is a *colour* dataset
- the AAE paper {makhzani2015adversarial} presented results on MNIST, SVHN and heavily cropped faces from the Fray dataset which is in **not** in colour.
- Kingma's VAE paper presents results on Fray and MNIST, again **not** in colour.
- the DVAE paper contains examples of synthesised images.

. Notwithstanding this, in the revised paper we have now included synthesized examples of colour faces. For ease of reference, we show some examples of synthesized images below:

TODO – use figures finally used in paper

In addition, we have introduced new experiments in facial attribute classification based on colour faces; these experiments were not present in the original paper, and represent a significant addition. These experiments are to be found in Section VI-F-2.

The reconstruction results are not so interesting and some of them can be removed. In addition, the reconstruction results on MNIST and Omniglot are inconsistent, which should be analysed in depth. Current version does not provide insight for the phenomena.

Though we have removed some reconstruction results, we now give motivation for maintaining an examination of reconstruction. We have included the following paragraph to justify these experiments:

> We are interested in reconstruction for several reasons. The first is that if we wish to use encodings for down stream tasks, for example classification, a good indication of whether the encoding is modeling the sample well is to check the reconstructions. For example if the reconstructed image is missing certain features that were present in the original images, it may be that this information is not preserved in the encoding. The second reason is that checking sample reconstructions is also a method to evaluate whether the model has overfit to test samples. The ability to reconstruct samples not seen during training suggests that a model has not overfit. The final reason, is to further motivate AAE, DAAE and iDAAE models as alternatives to GAN based models that are augmented with encoders {li2017alice}, for down stream tasks that require good sample reconstruction.

We have, however, provided significant improvements in the evaluation of the two proposed models by including experiments on facial attribute classification on colour faces.

The Parzen window estimator is not reliable [1]. It is better to use the methods proposed in [2] or [3] to compare the generative ability quantitatively.

The corresponding references are:

[1] Theis L, Oord A, Bethge M. A note on the evaluation of generative models[J]. arXiv preprint arXiv:1511.01844, 2015.

[2] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training gans[C]//Advances in Neural Information Processing Systems. 2016: 2234-2242.

[3] Wu Y, Burda Y, Salakhutdinov R, et al. On the quantitative analysis of decoder-based generative models[J]. arXiv preprint arXiv:1611.04273, 2016.

We were aware of the literature [1] and did cite Theis, in the last paragraph VII.E.- unfortunately in the uploading process, the last page of references was accidentally omitted. In the original submission paper, we had the following paragraph:

> Finally, evaluating generated samples is challenging: log-likelihood is not always reliable {theis2015note}, and qualitative analysis is subjective. For this reason, we provided both quantitative and qualitative results to communicate the benefits of introducing Markov chain sampling when a trained DAAE, and the advantages of iDAAEs over AAEs.

For this reason, we provided both quantitative and qualitative results to communicate the benefits of introducing Markov chain sampling with a DAAE, and the differences and advantages of iDAAEs compared to AAEs. We have also introduced new ways of communicating the results of image synthesis which are easier to interpret, which we hope makes the contributions clearer, even given the limitations of log-likelihood as a means of evaluation. We hope that the revised version of the paper – with improved face synthesis and experiments on facial attribute classification – further emphasises the improvements afforded by the proposed DAAE and iDAAE.

We would also like to emphasise the usefulness of the Omniglot dataset, which contains a testing dataset with examples of *entire alphabets* that are completely absent from the training data. We have since created a new visualisation of these results, to emphasise that they correspond to separate datasets in Omniglot. Notably (and this is something that enforces the reviewers point):

- if we evaluate log-likelihood on samples from alphabets that are different to those in the training data, the log-likelihood values get *worse* with an extended number of iterations
- if we evaluate log-likelihood on samples from alphabets similar to those in the training data, the likelihoods get *better* with an extended number iterations.

These observations were made in the original paper as follows:

> Conflicting log-likelihood values of generated samples between testing and evaluation datasets means that these measurements are not a clear indication of how the number of sampling iterations affects the visual quality of samples synthesized using a DAAE. In some cases it may be necessary to visually inspect samples in order to assess effects of greater or smaller numbers of sampling iterations.

The main advantage of the proposed method over DVAEs is providing the flexibility to choose complicated priors and posteriors. However, I do not observe any practical benefit of using mixture of Gaussian instead of high-dimensional Gaussian priors in terms of reconstruction error, log likelihood or sample quality. So, the motivation of the method is not well supported according to the current experiments.

To be clear, we do not suggest *replacing* high-dimensional Gaussians, but instead suggest using mixtures of high-dimensional Gaussians. The reason is that by supporting mixtures of Gaussians, one can encourage latent space to have an organisation that mirrors the semantic aspects that one wishes to represent in the data. To take a very simple case, if we have an observation space that is clearly bimodal (containing dogs and cats), then it would seem imminently sensible that latent space should have a distribution that reflects this need an example for which z_k "label" separability is a bad idea vs multimodal version.

Further, note that the distribution of encoded samples in the DAAE is a mixture of Gaussian. In the original paper we stated this:

> If $q_\phi(z|\tilde{x})$ is chosen to be Gaussian, then in many cases $\tilde{q}_\phi(z|x) = \int q_\phi(z|\tilde{x})c(\tilde{x}|x)d\tilde{x}$ will be a mixture of Gaussians.

an explanation that was also expressed by Im et al. {im2017denoising}.

As for the classification, the results should be compared with other unsupervised learning methods with adversarial losses like [4].
[4] Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning[J]. arXiv preprint arXiv:1605.09782, 2016.

- {4} Do NN1 classification for MNIST and semi-supervised learning for ImageNet
- We chose not to do semi-supervised learning because we wanted to evaluated the latent representation learned - which is best ahcieved using NN or training an SVM on top of the representation
- Our motivation was learning in the absence of larger amounts of labelled training data - for this the Omniglot dataset is ideal - since it has only 20 examples per class.
- This approach is used elsewhere in the literature to measure the linear seperability of learned encodings {kumar2017variational}.

In our revised version of the paper, we present results on the CelebA dataset. We perform facial attribute classification and compare the results to those obtained using a VAE and beta-VAE {kumar2017variational}. In each of these experiments presented by Kumar et al. {kumar2017variational} a linear classifier was applied to the encoding, our experiments are thus consistent with theirs.

Though the sampling method proposed do not rely on a given data but it is still an iterative one, which is less efficient than VAEs and GANs. Indeed, throughout this paper, I find no evidence that the proposed method can outperform VAEs and GANs, which are simple, effective and well-known.

- In the original paper, we identified an issue related to drawing samples from these models, which also applies to the DVAE.
- We propose methods to overcome these issues, and explain why it may be difficult to obtain visually plausible image samples from some denoinsing autoencoders.
- The purpose of the image samples is to demonstrate the issue and demonstrate empirically that our approach addresses the issue.

Minor things:

Some references are missing. For instance, [23] is not found in the reference list.

Unfortunately in the submission process the last page of references were cut off. They are now included.

The notations should be explained in detail, e.g. f_ϕ, p_θ.

We believe the reviewer is referring to p_ϕ and p_θ or possibly f_ψ? There is no f_ϕ in this paper. Notations p_ϕ and p_θ might be used in different ways in the literature, and so when we introduced them, we cite the work of Kingma et al.

From the original paper:

> *learning a probabilistic encoder cite{kingma2013auto}, $q_\phi(z|x)$, conditioned on observed samples, and a second probabilistic decoder cite{kingma2013auto}*

These sentences remain as-is in the current paper, as we believe them to be correct.

It is better to include some instances in Line 4-9, Column 2, Page2.

Have we addressed this?

I have no idea why the word “simple” in Line 19, Column 1, Page 3 is bold.

We agree that this was unnecessary. Removed.

It should be $\tilde{q}_\phi(z|x)$ instead of $\tilde{q}_\phi(x|z)$ in the last term of the equation in Line 34-35, Column 1 Page 3.

We have fixed this.

1 Reviewer: 2

Comments to the Author
This work introduces two new types of denoising autoencoders, DAAE and iDAAE. The authors propose an adversarial training methodology to address the analytical intractability associated with including a denoising criterion in the variational cost function. While sufficient theoretical development is provided, additional work is needed to validate the proposed solutions empirically. I suggest that the authors address the following comments in order to improve this manuscript:

(1) What process did you follow to select the parameters for your models? It would be helpful to outline a strategy for choosing at least some of the key parameters, such as the size of the encoding space.

- We used similar architectures to those in previous work for MNIST we used the same architecture as the AAE Makhzani.
- We modified this network for the Omniglot dataset, as discussed (Section VI.C.2):

  > Compared to the networks used for MNIST, deeper networks were needed in order for the loss functions to converge.

- We also modified the network of makhzani et al. for the sprite dataset; this was detailed in the original paper as follows:

  > For the decoder, we found that it was necessary to use a 3-layer fully connected neural network in order to capture the complexity of the sprite dataset. However, by comparing training and test data reconstruction error during training, we found that using 1000 neurons in each layer led to over-fitting. Rather, we used a 3-layer fully connected neural network with 1000 neurons in the first layer and 500 in each of the last layers.

In our revised version we have incorporated additional experiments on the CelebA dataset, which required us to extend the fully connected AAE to a convolutional version. We now include the following in our revised paper:

> For the CelebA dataset 3 types of model were trained: an AAE, a DAAE and an iDAAE. The models were constructed with convolutional layers, rather than fully connected layers since the CelebA dataset is more complex than the Toronto face dataset use by Makhzani et al. $\{$makhzani2015adversarial$\}$. The encoder and decoder consisted of 4 convolutional layers with a similar structure to that of the DCGAN proposed by Radford et al. $\{$radford2015unsupervised$\}$. We used a 3-layer fully connected network for the discriminator.

(2) In Section V.A., please indicate the specific nonlinear activation functions you used in the intermediate layers of the network.

We use ReLU's. We have included information on this as follows (revisions to the paper):

> Rectifying Linear Units (ReLU) were used between all intermediate layers to encourage the networks to learn representations that capture multi-modal distributions.
> ...
> Intermediate layers of the network have ReLU activation functions to encourage the network to capture highly non-linear relations

(3) How sensitive are your models to changes in the parameter configuration? Is extensive parameter tuning required in order to achieve your reported results?

We did very little hyper parameter tuning. To make our comparisons as "fair" as possible we used the same hyper parameters for each experiment, and the choice for hyper-parameters was based on previous work:

> In order to compare models trained on the same datasets, the same network architectures, batch size, learning rate, annealing rate, size of latent code and number of epochs is used for each

However, in the revised submission, in which we now include a significant number of CelebA experiments, we have had to train an AAE had not been trained on the CelebA dataset using our proposed architecture. For this, a more extensive hyper parameter search was performed. This is detailed in the **revised paper** as follows:

> Networks were trained for 100 epochs with a batch size of 64 using RMSprop with learning rate 1^{-4} and momentum of 0.1 for training the discriminator. We found that using smaller momentum values lead to more blurred images, however larger momentum values prevented the network from converging and made training unstable. When using Adam instead of RMSprop (on the CelebA dataset specifically) we found that the values in the encodings became very large, and not consistent with the prior. The encoding is made up of 200 units and the prior 200D Gaussian. The corruption process used to train the DAAE and iDAAE was additive Gaussian noise. We experimented with different noise level, σ between $[0.1, 1.0]$, we found several values in this range to be suitable. For our classification experiments we fixed $\sigma = 0.25$ and for synthesis from the DAAE, to demonstrate the effect of sampling, we used $\sigma = 1.0$. For the iDAAE we experimented with $M = 2, 5, 20, 50$. We found that $M < 5$ (when $\sigma = 1.0$), was not sufficient to train an iDAAE. By comparing histograms of encoded data samples to histograms of the prior, for an iDAAE trained with a particular M value, we are able to see whether M is sufficiently larger or not. We found $M = 5$ to be sufficiently large.

(4) You use a standard AAE as a benchmark for evaluating DAAE and iDAAE. It would be helpful to also include a regular autoencoder in your results for reference.

Regular autoencoders may not be used for sample synthesis because the latent distribution is not known - only a denoising variant may be used to synthesise samples. Thus, in order to perform a comparison, we need to do something else.

So, in our revised paper, we test classification performance in a standard dataset that is used for generative models: we use the CelebA dataset, and perform facial attribute classification. This allows us to make comparisons to well-known models including the VAE and β-VAE (both are considered state-of-the-art). The classification resutls are described in Part F of Section IV. Robustness of parameters is also included in these studies (see for example, Figure 10 of the revised paper).

(5) All of your experimental results use image datasets. Can you comment on the applicability & expected performance of your methods on non-image data?

We deliberately describe the method in a very general framework. The models $q_\phi(z|x)$ and $p_\theta(x|z)$ may be instantiated using any differentiable parameterized model. What we have done

is test the models in vastly different image sets that are not described in the paper. For example, we have looked at skin lesion classification, a very narrow sub-domain of image data space. Results hold up extremely well.

(6) In Section VI.D., you state "The reconstruction is evaluated by computing the mean squared error between the reconstruction and the original sample." Please clarify if "original sample" means the uncorrupted withheld test data. The uncorrupted data should be used to evaluate the quality of the reconstruction across all methods in order to provide an accurate comparison, regardless of whether the samples were corrupted going into the model.

We agree. All experiments should have been performed on uncorrupted data. We have now done this and updated the reconstruction values which are shown in Table IV.

We have also clarified the writing in the revised version of the paper:

> The reconstruction task involves passing samples from the test dataset (i.e. a samples not seen during training) through the trained encoder and decoder to recover a sample similar to the original (uncorrupted) sample. The reconstruction is evaluated by computing the mean squared error between the reconstruction and the original sample.

(7) There is no comparison in training times between the different methods. A comparison of both performance and training complexity is needed to evaluate the practical merits of this work. Ideally this should be provided for all of the methods you considered as well as a standard autoencoder without regularisation or adversarial training. This is particularly important to consider given how close your proposed methods are in performance to much simpler methods like PCA.

Standard autoencoders may not be used for sample synthesis

It is becoming rather rare in this field to include information on training time, which will vary based on resource available, choice of framework and machine. The additional computation cost from corruption is minimal (compared to, say, a vanilla AAE), and the computational cost associated with the adversarial loss is dependant on the size of the discriminator. Comparison of training times are likely to be less meaningful. Instead, we had added additional results to demonstrate the benefits of using DAAE or iDAAE compared to other state of art models.

In the revised version, we include facial attribute classification experiments which compare our model to state-of-the-art models rather than PCA, which we hope motivates the models further. Experiments are now comparable with very recent work in this field Citations needed here.

2 Reviewer: 3

Comments to the Author The paper introduces two denoising adversarial autoencoders combining concepts from denoising and adversarial autoencoders known in the literature.

The paper presents an interesting approach for unsupervised representation learning, combining advantages of previously proposed models. The authors also address the issue of generating samples when a posterior of a latent representation based on corrupted input is matched to prior rather than posterior based on uncorrupted input. The authors propose a Markov Chain to generate samples in this case and provide theoretical guarantees for this although here they highly rely on the previous work by Bengio et al 2013, 2014.

The main concerns regarding the paper are about the experimental section. Although the authors provide a very thorough evaluation of their methods on three different datasets for three

different tasks there are still some moments that could potentially improve the value of the paper. First of all, the comparison is provided only with respect to the adversarial autoencoder. It is a very interesting comparison as it allows to empirically evaluate the influence of denoising training of the adversarial autoencoder. On the other hand the proposed methods can be considered as adversarial extension of denoising autoencoders, and it would be interesting to see how adversarial learning to match the prior works in comparison to Kullback-Leibler minimisation used in DVAE, for example.

Makhzani citation needed here has already convincingly shown the benefits of using adversarial training rather than the KL-divergence to match a prior in the case of AAE and VAE. In short, it is because the KL divergence admits an analytic modification to the loss function under only a limited number of distributions.

For this reason we did not make experiments comparing DVAE with DAAE a priority, rather we focused on comparing DAAE and AAE. This second point appears completely out of sync with the reviewers comment.

In our revised version we have performed additional experiments with the CelebA dataset of colour faces, performing facial attribute classification. We now compare results to those obtained for well known models including VAE and β-VAE. This is in alignment with methods of evaluation in the field.

The second concern regarding the experimental setup is about choice of the prior used in the experiments. Except DAAE with 2D 10-GMM prior that shows poor results and is outperformed by all the competitors the experiments are conducted with Gaussian prior that could have been trained with DVAE and VAE. The main motivation of the proposed methods in compariosn to DVAE claimed by the authors was the possibility of using different more complex priors. It would be good to see this in the experiments.

CHECK THAT DVAE is more similar to DAAE than iDAAE: We have now moved the MNIST results to the appendix as we agree that the MNIST results are not meaningful. However, the DVAE is similar to the DAAE, but not the iDAAE. In the iDAAE the KL divergence between $\tilde{q}_\phi(z|x)$ and $p(z)$ is minimized, since $\tilde{q}_\phi(z|x)$ is often a mixture of Gaussian (as stated in the paper), there is no analytic solutional for this KL divergence and so it is still neccesary to implement the model using an adversarial approach.

(The other drawback mentioned for DVAE that is non-trivial synthesis is also valid for the proposed DAAE but successfully overcome by the authors and could have been overcome for DVAE).

Yes, exactly. Although we did not explore this further, since we were focusing on AAE, but this could be an avenue for future work. we will add a sentence on this in the Conclusions ? The quote below is unclear

> 6) applying our proposed sampling process to the DVAE {im2017denoising} - for which sampling is non-trivial for the same reasons discussed for the DAAE

Other comments:

1. The first two paragraphs of the introduction are general and do not mention anything about neural networks although all the references are for papers related to NN. It would be better to either mention in the text that the authors are talking about NNs or to add some not NN references.

We chose to present the model and theory in a more general framework and then show that neural networks may be used for implementation.

In the last paragraph of section IV.B. in the original paper, we had explained:

> *The analyses of Sections* **??** *and* **??** *are deliberately general: they do not rely on any specific implementation choice (e.g. particles, artificial neural networks, or parametrised forms of distribution) to capture the model distributions.*

In addition, at the start of the next section, section V.A. we talk about the implementation using neural networks.

> *Under the autoencoder framework,* $E_\phi(x)$ *is the encoder and* $R_\theta(z)$ *is the decoder. We used fully connected neural networks for both the encoder and decoder.*

Further, in Figure 1, on page 1 where we present our model in the context of other relevant recent work, the caption in the original paper says:

> *Arrows in this diagram represent mappings implemented using trained neural networks.*

and we refer to Figure 1, multiple times in the Introduction. We have now modified Figure 1 to hilight its importance and we now include the word *deep* in a modification of the caption of Figure 1.

2. In the list of references there are only 21 papers whereas in the text works are cited upto 24

In the submission process the last page was unintentionally left out. The rest of the references have now been included.

3. Section I.A First paragraph. "... there is a ground truth label for every training image." - It has not been specified before that classification of images only is considered.

Thank you for catching this, his has been replaced by:

> because there is a ground truth label for every training data sample.

We deliberately keep the method general.

4. Section I.A Autoencoders do not have to be probabilistic, generally speaking. It would be better to mention it and probabilistic interpretation is considered here.

It is advantageous for us to talk about autoencoders as probabilistic models – since our notation follows that of a probabilistic framework. Further, VAE's introduced by Kingma et al. were introduced in a probabilistic framework, this is why we cite this work, when introducing autoencoders in a probabilistic framework:

> *probabilistic encoder* cite{kingma2013auto}, $q_\phi(z|x)$, conditioned on observed samples, and a second *probabilistic decoder* cite{kingma2013auto}

Indeed, Kingma et al. introduce autoencoders as follows:

> learning a *probabilistic encoder* cite{kingma2013auto}, $q_\phi(z|x)$, conditioned on observed samples, and a second *probabilistic decoder* cite{kingma2013auto},

Please see our response to Reviewer 1 regarding the generality of a probabilistic treatment, under which deterministic autoencoders are merely a special case.

5. Section I.A. Last paragraph. It may be worth to add some references here.

We have added references as follows:

> In some situations the encoding distribution is chosen rather than learned {bengio2013generalized}, in other situations the encoder and decoder are learned simultaneously {kingma2013auto, makhzani2015adversarial, im2017denoising}

6. Please use the same tense citing other works: "Bengio at al. [4] treat...", "Im at al. [8] showed...","Im et al. [8] do not address...", "Makhzani et al. [13] introduced..."

We have updated the writing to be more consistent in the use of tenses:

- Im at al. [8] showed → Im at al. [8] show
- Makhzani et al. [13] introduced → Makhzani et al. [13] introduce (the rest of the paragraph was update to be in the present tense.)
- Bengio et al. {bengio2013generalized} proposed → Bengio et al. {bengio2013generalized} propose

7. Organisation of the paper. It is not clear why denoising autoencoders are reviewed in "Background" section whereas adversarial autoencoders are reviewed in "Related work" section.

- In the "Background" we talk about Autoencoders in a more general framework, this allows us to introduce many concpets in this framework.
- Since our work specifically builds on Adversarial autoencoders, we have a specific section for discussing them, which we call "Related work".

8. Section I.C. Last paragraph looks a bit odd especially given the first paragraph of Section I.D

We thank the reviewer for the comment. Perhaps this will clarify: At the end of section I.C. we are referring to the original VAE paper, hence the citation to Kingma et al. – we are pointing out that the Kinga paper does not use a denoising principle.

In contrast, at the start of section I.D. we are beginning to introduce the work of Im et al. Cited in paragraph two if Section I.D. who introduce DVAEs, which are VAE's with denoising.

In our revised version, end Section I.C. with the following:

> no corruption process was introduced by Kingma et al. during VAE training.

9. Section II.A. It is unclear why it is the only section uses notation w and v instead of x and z, moreover w and v are undefined

The use of v and w instead of x and z was done deliberately. Adversarial training is usually performed on data samples, x rather than on latent samples, z. If we introduced adversarial training on the data, but then applied it to the latent space, this may have cause confusion in notation. If we had introduced directly for the latent space, readers familiar with adversarial training may have been confused.

For these reasons we deliberately introduce adversarial training with a general notation. We wanted to avoid calling samples, w data samples and sample, v latent samples. Instead we refer to samples w as being from a "target distribution".

To help avoid confusion, in our revised version we refer to v samples as "input samples" coming from our chosen prior distribution $p(v)$ and generated w samples as "output samples". In the revised version we say:

> produce output samples, w that match a target probability distribution $t(w)$.
> ...
> The generative model - fed with input samples v, drawn from a chosen prior distribution, $p(v)$ - is trained to generate output samples w that are indistinguishable from target w samples

10. Section IV.B. Second paragraph. "... we used PREVIOUSLY in Algorithm 2 of the Appendix..." - this is the first time this algorithm is mentioined and "previously" implies that a reader should have already been aware of it

- This was an error. We were referring to the wrong algorithm. To make this more clear, we describe the sampling process that we are referring to in the text. We now say:

> To ensure that we draw novel data samples, we do not want to draw samples from the training data at any point during sample synthesis. This means that we cannot use data samples from our training or test data to approximately draw latent samples, z from $\tilde{q}_\phi(z|x)$.

11. Theorem IV.3 formulation. "sufficient approximation" is ambiguous

- We have changed this to:

> Assuming that $p_\theta(x|z)$ is approximately equal to $\mathcal{P}(x|z)$,

12. Section V. "We represent an image in the form of a 3D array" - Does it mean the RGB representation?

- It means, $C \times W \times H$, C - colour channel (would be 1 for grey scale), W - width of the image, H - height of the image.
- Due to limited space, we have removed this comment since it does not add significatly to the understanding of the paper.

13. Section V.A. Last paragraph. It would be better to add some discussion about fundamental difference between encoder and decoder that lead to this.

We had updated this with the following: The vectors output by the encoder may take any real values, therefore minimising reconstruction error is not sufficient to match either $q_\phi(z|\tilde{x})$ or $\tilde{q}_\phi(z|x)$ to the prior, $p(z)$.

14. Algorithm 1. For clearer representation it may be better to include both iDAAE and DAAE in the Algorithm with the difference in Line 11

- We have considered this - however we do not see a tidy way to combine both algorithms in one and we would not like readers to be confused by writing the algorithms together.
- We think it is better to present one algorithm for the iDAAE and explain that by changing line 11 we obtain the algorithm for the DAAE, as we do already:

> Algorithm 1 shows the steps taken to train an iDAAE. To train a DAAE instead, all lines in Algorithm 1 are the same except Line 11, which may be replaced by $z_{fake} = E_\phi(\tilde{x})$.

15. Section V. Last paragraph. "... non-denoising adversarial autoencoders" → "... non-denoising adversarial autoencoders AAE [13]"

We have done this.

16. Section VI.B.2. The only datasets where the colour scheme of images is not mentioned

We have added the following:

> Each example in the dataset is 105-by-105 pixels, taking values $\{0,1\}$.

17. Section VI.C. It is unclear why 10-GMM prior is chosen only for DAAE and not for iDAAE

We have removed the MNIST results from the main body of the text.

18. Section VI.C. Which optimisation algorithm is used for training?

- To ensure that our experiments are repeatable we provided code in the Theano framework: `https://github.com/ToniCreswell/DAAE_/blob/master/DAAE_sprite_v2.ipynb`
- Since the submission of this paper, it was announced that the Theano framework would no longer be supported, for this reason, code used to run the additional experiments has been written in pyTorch as is available here: `https://github.com/ToniCreswell/pyTorch_DAAE`.
- We use adam
- We have added the following:

> In order to compare models trained on the same datasets, the same network architectures, batch size, learning rate, annealing rate, size of latent code and number of epochs is used for each. All modes were trained using the Adam cite$\{$kingma2014adam$\}$ optimisation algorithm.

19. Section VI.D. The first sentence makes it unclear that for AAE images are not corrupted and the following statements that for AAE uncorrupted images are used are confusing

The first secntence is making a distinction between the reconstruction task (for which, in the revised version, we no longer corrupt images) and the synthesis task. In the revised version of the paper we perform reconstruction on non-corrupted image, we have updated this sentence as follows:

> In reconstruction, we start from a test data sample, encode it to get a latent sample and decode to reconstruct the original.

20. Tables III and IV. "200D Gaussian" is placed such that it was used only for DAAE, moreover in the following tables "200D Gaussian" is repeated for each row

NEED TO UPDATE PAPER: For these experiments we use the same prior and so in the revised version of our paper we put this information in the caption as well as in the model description and remove it from the table.

21. Section VI.D.1. It would be better to have more discussion about why the performance on MNIST of the proposed methods is worse than AAE's.

We have now removed the MNIST results from the main body of the paper. The dataset was perhaps too tivial to see the benefits of denoinsing.

22. Section VI.E. It is claimed that "if we pass samples from the prior through the decoder of a trained DAAE, the samples are likely to be inconsistent with the training data". It would be interesting to see this in practise. There are results with $z^{(0)}$, but they are taken from some other distribution rather than from the prior.

There are visual examples of this in Figures 2,4,5 of the original paper and numerical results in Tables 5,6,7 and 8. The samples $x^{(0)}$ are examples of these. It is only for Figure 3 that we take $z(0)$ from a disitribution that is different to the prior.

This was indicated by saying this in the Omniglot figure captions:

> The chain was initialized with $z^{(0)} \sim \mathcal{N}(0, I)$.

which was the same as the prior for all experiments except the case where we used a 2D mixture of Gaussian.

We also say in the main body of the text, where we discuss our proposed sampling procedure:

> we draw an initial $z^{(0)}$ from any random distribution – we use a normal distribution for simplicity ... $x^{(0)}$ is the sample generated when $z^{(0)}$ is passed through the decoder.

We did not want to say explicitly that the inital samples was drawn from the prior, as this may confuse some readers.

In our revised version we also include celebA faces where $z^{(0)}$ is sampled from the prior, and there is a vivid difference between inital and final samples.

23. Section VI.E. It is better to stick with the same order of the datasets for all the tasks

We have updated the order to be Omniglot, Sprites, CelebA for all experiment sections.

24. Section VI.E.3. Phrases "less negative log-likelihood" and "more negative" are very confusing, as it seems like log-likelihood is still used whereas "negative log-likelihood" is another popular measure. It is better to just say "larger" and "smaller"

Anil - I dont know if I agree with this?

25. Section VI.F.1. Second paragraph, first sentence implies that there will be "second" with discussion DAAE with 2D 10-GMM

We have now removed experiments on the MNIST dataset and replaced them with experiments on te CelebA dataset. So this is no longer in the main body of the text.

26. Section VI.G. It would be good to provide time estimates here to understand how much longer iDAAE takes to train

This will depend heavily on the type of framework and GPU that you use. Any values we provide will either be trivial, for example, if the iDAAE has M interration we have to compute M times the number of encodings, which may take M times as long, however if you have a very large GPU you may be able to compute M encodings in one go. I did not have access to large GPUs (Ge Force 680 for experiments in the original paper) and so could not do this.

To make this clear we have added the following to the paper:

> On the other hand, it is possible to perfrom the intergration process in parallel provieded that sufficient computational resource is available.

27. Section VII. First two paragraphs can be safely skipped.

In our revised paper we have removed these two paragrpahs.

28. References. It is better to cite published version of the works rather than their preprints from arxiv, like in [8] - Proceeding of the 31 AAAI Conference on Artificial Intelligence, or in [9] arXiv reference is redundant

In the revised version we have updated reference [8] and removed the redunant arXiv reference in [9] and all other redunant arXiv references.

29. Appendix A. It is odd to use different names for the same lemmas as in the main body of the paper.

In the revised version we have re-numbered the Lemmas and Proofs so that they are consistent between the main body of text and the appendix.

30. Appendix A. Proof of Lemma A.1. "Similar to proof in Bengio [3]" looks like it was not supposed to be there.

This is supposed to be there. The proof is similar.

31. Appendix A. There is no point to state the theorem here as its proof has been provided in the main body of the paper.

In the reivsed version we have removed the theorem from the appendix.

Minor notes: 1. Page 1, left column, row 24. "...encoded data in THE latent space" - missing article

We have addressed this in our revised version.

2. Caption for Figure 1. All the models are mentioned both with their names and acronyms except Denoising Variational Autoencoders

We have addressed this in our revised version.

3. Page 2, left column, rows 30-32. "Methods to sample denoising adversarial autoencoders through Markov-Chain sampling". Should they sample synthetic data points rather than autoencoders?

In our revised version we now say:

> Methods to draw synthetic data samples from denoising adversarial autoencoders through Markov-Chain sampling;

4. Page 3, left column, rows 9-12. "In the second step, a random variable ... the Hadamard product" - missing verb

In our revised version we now say:

> In the second step, a random variable $\epsilon \sim \mathcal{N}(0, 1)$ is drawn and the encoding is calculated as $z = \mu_\phi(x) + \sigma_\phi(x) \circ \epsilon$ where $\circ$ is the Hadamard product, see the VAE in Figure **??**.

5. Page 3, left column, row 33. "... may be formed in the following two ways" - "two ways" are not very clear

6. Page 5, left column, row 37. "This will be come ..." -¿ "This will become ..."

We have addressed this in our revised version.

7. Page 5, right column, row 48 (and the same for Appendix). “... the sampling process in 2” -¿ “... the sampling process in (2)”

We have addressed this in our revised version.

8. Page 5, right column, row 52 (and the same for Appendix). “... by the transition operator, $T_{\theta,\phi}(z_{t+1}|z_t)$” -¿ “... by the transition operator, $T_{\theta,\phi}(z_{t+1}|z_t)(3)$”

We have addressed this in our revised version.

9. Page 8, left column, row 7. “Prior indicated...” -¿ “Prior indicates...”

10. Page 13, left column, row 23. Should it be “testing dataset”?

3 Additional Changes

The addition of Figure 2, which compares how the iDAAE and DAAE match encodings to the prior when trained on the CelebA dataset.