FORMATTING INSTRUCTIONS FOR ICLR 2019 / CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reproducibility of results obtained in research papers is an important part of the research process, as it allows to support the conclusions of the paper, or otherwise find potential inconsistencies in the results. The paper we have chosen to reproduce is called *MAE : Mutual Posterior-Divergence Regularization for Variational Autoencoders*, and aims to improve the performances of Variational Autoencoders (VAE) by constraining their latent variables as to ensure they do contain useful information. VAEs are one of two very powerful generative models and can be useful in many different fields, which is why we decided to study this paper.

1 INTRODUCTION

Our research history.

The paper *MAE : Mutual Posterior-Divergence Regularization for Variational Autoencoders* is based on research on Variational Autoencoders (VAEs) and tries to mitigate performance issues araising when the latent representation becomes uninformative, collapsing the model to an unconditional generative model. We had to read multiple papers about previous research on VAEs, as well as different Neural Network models used in the studied paper. We will here introduce those concepts and talk about the need for a better model than simple VAE as described in the first paper introducing the model (Kingma & Welling (2006)).

1.1 VARIATIONAL AUTOENCODERS

We learned about VAE, how it works

what's an autoencoder

An autoencoder is a generative model and can be decomposed into two parts : an *encoder* and a *decoder*, which both can be many kinds of networks. The encoder generates a so-called *latent representation* of the original data received as input, which the decoder will then use to try and recover the original information. The goal is then to train the encoder to generate the most informative latent variables, and the decoder to reconstruct the original data as closely as possible. Such a model can be very useful for many reasons : an efficient latent representation can be useful for analysis and understanding of some natural process or for data representation tasks, and when used along with the decoder, it can be used to generate realistic data from small representations.

For the rest of our discussion, let x be an observable random variable which is assumed to be generated from a hidden continuous random variable z, following a conditional distribution $p_\theta(x|z)$.

Kingma and Welling in their paper wish to design an autoencoder that works well even on latent variables with intractable posterior distributions $p_\theta(x|z)$, i.e. that cannot be evaluated or differenciated. To this end, they introduce what they call a *recognition model* $q_\phi(z|x)$, an approximation to the true posterior $p_\theta(z|x)$. They also call $q_\phi(z|x)$ the *encoder* and consequently the conditional distribution $p_\theta(x|z)$ the *decoder*.

In fact, they point out that the marginal likelihood $\log p_\theta(x)$ can be lower bound by the following :

$$\log p_\theta(x) \geq \mathcal{L}(\theta, \phi; x) = -D_{KL}\left(q_\phi(z|x)||p_\theta(z)\right) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \tag{1}$$

which they explain has too high a variance to be practical when estimated using a standard Monte Carlo gradient estimator, and where D_{KL} is the KL divergence of two distributions. To solve this problem, they introduce a new lower bound that takes advantage of a *parameterization trick*, which consists in rewriting the distribution of z as a function of a noise variable ε :

$$z \sim q_\phi(z|x) = g_\phi(\varepsilon|x) \tag{2}$$

where ε follows an appropriate distribution $p(\varepsilon)$. This makes it possible, after replacing the lower bound in 1, to apply a Monte Carlo estimator on the second term, yielding :

$$\tilde{\mathcal{L}}(\theta, \phi; x) = -D_{KL}\left(q_\phi(z|x)||p_\theta(z)\right) + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(x, g_\phi(\varepsilon_l, x)) \tag{3}$$

where L is some sample size and ε_l is a sample from the distribution of ε.

1.2 AUTOENCODING CAPABILITIES OF VAEs

Read about VLAE, which explains well why VAE is not enough. Talk about KL-varnishing at some point

As a followup reading, cited in our paper of interest about MAEs, we looked at *Variational Lossy Autoencoders* (Chen et al. (20017)), where we found good insight as to why VAEs on their own are not always sufficient for autoencoding.

They note that when the decoder is made too expressive, it is easy for the model to set the approximate posterior $q_\phi(z|x)$ to equal the prior $p_\theta(z)$ and hence avoid any cost from the regularizing term $D_{KL}(p_\phi(z|x)||p_\theta(z))$ in 3. Although this is explained by "optimization challenges" of VAE in most research , they present another observation that, even on a perfectly optimized VAE, ignoring the latent variables should still result in optimum results in most instances of VAE with intractable true posterior and powerful enough decoders.

1.3 CONVNET

Finally, as most experiments in our paper of interest make use of Neural networks for the VAE encoder and decoder, we describe here in essence what they are. Note that we focus on Convolutional Neural Networks (ConvNet) rather than Residual Neural Networks (ResNet), even though the latter is also used in the paper. We unfortunately have had technical and timing constraints, and Resnet being a fairly big model, we decided to focus on having good results on ConvNet alone.

Regular Neural Nets A Neural Network is composed of a sequence of hidden layers, each consisting of neurons which connect to all neurons of the previous layer. The first layer takes a single vector as input, and the last layer outputs the result. For example, in a classification problem, it outputs a score for each class.

Convolutional Neural Nets An issue with Regular Neural Networks is the fact that each neuron must be connect to all neurons of the previous layer, which makes the model impractically big for large input data. Convolution Neural Networks (ConvNet) take advantage of the fact that, in inputs that represent images, localized information in the image can be enough for learning. Each layer organizes its neurons in a three dimensions (width, height, depth) and each neuron is not constrained to be fully connected to the previous layer. The first layer then usually has the same dimensions as the input images, and for the example of a classification problem with c classes, the output data will have dimension $1 \times 1 \times c$.

2 SUBMISSION OF CONFERENCE PAPERS TO ICLR 2019

ICLR requires electronic submissions, processed by `https://openreview.net/`. See ICLR's website for more instructions.

If your paper is ultimately accepted, the statement `\iclrfinalcopy` should be inserted to adjust the format to the camera ready requirements.

The format for the submissions is a variant of the NIPS format. Please read carefully the instructions below, and follow them faithfully.

2.1 STYLE

Papers to be submitted to ICLR 2019 must be prepared according to the instructions presented here.

Authors are required to use the ICLR LaTeX style files obtainable at the ICLR website. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

2.2 RETRIEVAL OF STYLE FILES

The style files for ICLR and other conference information are available on the World Wide Web at

`http://www.iclr.cc/`

The file `iclr2019_conference.pdf` contains these instructions and illustrates the various formatting requirements your ICLR paper must satisfy. Submissions must be made using LaTeX and the style files `iclr2019_conference.sty` and `iclr2019_conference.bst` (to be used with LaTeX2e). The file `iclr2019_conference.tex` may be used as a "shell" for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

The formatting instructions contained in these style files are summarized in sections 3, 4, and 5 below.

3 GENERAL FORMATTING INSTRUCTIONS

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing of 11 points. Times New Roman is the preferred typeface throughout. Paragraphs are separated by 1/2 line space, with no indentation.

Paper title is 17 point, in small caps and left-aligned. All pages should start at 1 inch (6 picas) from the top of the page.

Authors' names are set in boldface, and each name is placed above its corresponding address. The lead author's name is to be listed first, and the co-authors' names are set to follow. Authors sharing the same address can be on the same line.

Please pay special attention to the instructions in section 5 regarding figures, tables, acknowledgments, and references.

4 HEADINGS: FIRST LEVEL

First level headings are in small caps, flush left and in point size 12. One line space before the first level heading and 1/2 line space after the first level heading.

4.1 HEADINGS: SECOND LEVEL

Second level headings are in small caps, flush left and in point size 10. One line space before the second level heading and 1/2 line space after the second level heading.

4.1.1 HEADINGS: THIRD LEVEL

Third level headings are in small caps, flush left and in point size 10. One line space before the third level heading and 1/2 line space after the third level heading.

5 CITATIONS, FIGURES, TABLES, REFERENCES

These instructions apply to everyone, regardless of the formatter being used.

5.1 CITATIONS WITHIN THE TEXT

Citations within the text should be based on the `natbib` package and include the authors' last names and year (with the "et al." construct for more than two authors). When the authors or the publication are included in the sentence, the citation should not be in parenthesis (as in "See Hinton et al. (2006) for more information."). Otherwise, the citation should be in parenthesis (as in "Deep learning shows promise to make progress towards AI (Bengio & LeCun, 2007).").

The corresponding references are to be listed in alphabetical order of authors, in the REFERENCES section. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

5.2 FOOTNOTES

Indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).[2]

5.3 FIGURES

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction; art work should not be hand-drawn. The figure number and caption always appear after the figure. Place one line space before the figure caption, and one line space after the figure. The figure caption is lower case (except for first word and proper nouns); figures are numbered consecutively.

Make sure the figure caption does not get separated from the figure. Leave sufficient space to avoid splitting the figure and figure caption.

You may use color figures. However, it is best for the figure captions and the paper body to make sense if the paper is printed either in black/white or in color.

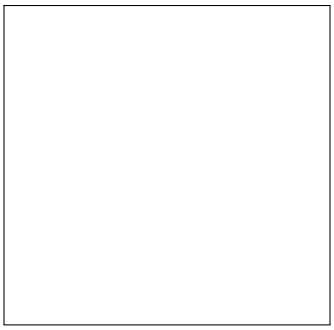

Figure 1: Sample figure caption.

5.4 TABLES

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

[1] Sample of the first footnote

[2] Sample of the second footnote

Table 1: Sample table title

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

6 DEFAULT NOTATION

In an attempt to encourage standardized notation, we have included the notation file from the textbook, *Deep Learning* Goodfellow et al. (2016) available at `https://github.com/goodfeli/dlbook_notation/`. Use of this style is not required and can be disabled by commenting out `math_commands.tex`.

Numbers and Arrays

a	A scalar (integer or real)
$\boldsymbol{a}$	A vector
$\boldsymbol{A}$	A matrix
$\mathbf{A}$	A tensor
$\boldsymbol{I}_n$	Identity matrix with n rows and n columns
$\boldsymbol{I}$	Identity matrix with dimensionality implied by context
$\boldsymbol{e}^{(i)}$	Standard basis vector $[0, \dots, 0, 1, 0, \dots, 0]$ with a 1 at position i
$\text{diag}(\boldsymbol{a})$	A square, diagonal matrix with diagonal entries given by $\boldsymbol{a}$
a	A scalar random variable
$\mathbf{a}$	A vector-valued random variable
$\mathbf{A}$	A matrix-valued random variable

Sets and Graphs

$\mathbb{A}$	A set
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	The set containing 0 and 1
$\{0, 1, \dots, n\}$	The set of all integers between 0 and n
$[a, b]$	The real interval including a and b
$(a, b]$	The real interval excluding a but including b
$\mathbb{A} \backslash \mathbb{B}$	Set subtraction, i.e., the set containing the elements of $\mathbb{A}$ that are not in $\mathbb{B}$
$\mathcal{G}$	A graph
$Pa_{\mathcal{G}}(\mathrm{x}_i)$	The parents of x_i in $\mathcal{G}$

Indexing

a_i	Element i of vector $\boldsymbol{a}$, with indexing starting at 1
a_{-i}	All elements of vector $\boldsymbol{a}$ except for element i
$A_{i,j}$	Element i, j of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{i,:}$	Row i of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{:,i}$	Column i of matrix $\boldsymbol{A}$
$\mathsf{A}_{i,j,k}$	Element (i, j, k) of a 3-D tensor A
$\mathsf{A}_{:,:,i}$	2-D slice of a 3-D tensor
a_i	Element i of the random vector $\mathbf{a}$

Calculus

$\frac{dy}{dx}$	Derivative of y with respect to x
$\frac{\partial y}{\partial x}$	Partial derivative of y with respect to x
$\nabla_{\boldsymbol{x}} y$	Gradient of y with respect to $\boldsymbol{x}$
$\nabla_{\boldsymbol{X}} y$	Matrix derivatives of y with respect to $\boldsymbol{X}$
$\nabla_{\mathsf{X}} y$	Tensor containing derivatives of y with respect to X
$\frac{\partial f}{\partial \boldsymbol{x}}$	Jacobian matrix $\boldsymbol{J} \in \mathbb{R}^{m \times n}$ of $f : \mathbb{R}^n \to \mathbb{R}^m$
$\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x})$ or $\boldsymbol{H}(f)(\boldsymbol{x})$	The Hessian matrix of f at input point $\boldsymbol{x}$
$\int f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral over the entire domain of $\boldsymbol{x}$
$\int_{\mathbb{S}} f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral with respect to $\boldsymbol{x}$ over the set $\mathbb{S}$

Probability and Information Theory

$P(\mathrm{a})$	A probability distribution over a discrete variable
$p(\mathrm{a})$	A probability distribution over a continuous variable, or over a variable whose type has not been specified
$\mathrm{a} \sim P$	Random variable a has distribution P
$\mathbb{E}_{\mathrm{x} \sim P}[f(x)]$ or $\mathbb{E} f(x)$	Expectation of $f(x)$ with respect to $P(\mathrm{x})$
$\mathrm{Var}(f(x))$	Variance of $f(x)$ under $P(\mathrm{x})$
$\mathrm{Cov}(f(x), g(x))$	Covariance of $f(x)$ and $g(x)$ under $P(\mathrm{x})$
$H(\mathrm{x})$	Shannon entropy of the random variable x
$D_{\mathrm{KL}}(P \Vert Q)$	Kullback-Leibler divergence of P and Q
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution over $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

Functions

$f : \mathbb{A} \to \mathbb{B}$	The function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f \circ g$	Composition of the functions f and g
$f(\boldsymbol{x}; \boldsymbol{\theta})$	A function of $\boldsymbol{x}$ parametrized by $\boldsymbol{\theta}$. (Sometimes we write $f(\boldsymbol{x})$ and omit the argument $\boldsymbol{\theta}$ to lighten notation)
$\log x$	Natural logarithm of x
$\sigma(x)$	Logistic sigmoid, $\dfrac{1}{1 + \exp(-x)}$
$\zeta(x)$	Softplus, $\log(1 + \exp(x))$
$\|\|\boldsymbol{x}\|\|_p$	L^p norm of $\boldsymbol{x}$
$\|\|\boldsymbol{x}\|\|$	L^2 norm of $\boldsymbol{x}$
x^+	Positive part of x, i.e., $\max(0, x)$
$\mathbf{1}_{\text{condition}}$	is 1 if the condition is true, 0 otherwise

7 FINAL INSTRUCTIONS

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the REFERENCES section; see below). Please note that pages should be numbered.

8 PREPARING POSTSCRIPT OR PDF FILES

Please prepare PostScript or PDF files with paper size "US Letter", and not, for example, "A4". The -t letter option on dvips will produce US Letter files.

Consider directly generating PDF files using `pdflatex` (especially if you are a MiKTeX user). PDF figures must be substituted for EPS figures, however.

Otherwise, please generate your PostScript and PDF files with the following commands:

```
dvips mypaper.dvi -t letter -Ppdf -G0 -o mypaper.ps
ps2pdf mypaper.ps mypaper.pdf
```

8.1 MARGINS IN LATEX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the graphicx package. Always specify the figure width as a multiple of the line width as in the example below using .eps graphics

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.eps}
```

or

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

for .pdf graphics. See section 4.4 in the graphics bundle documentation (`http://www.ctan.org/tex-archive/macros/latex/required/graphics/grfguide.ps`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

Xi Chen, Diederik Kingma, Time Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder, mar 20017. arXiv:1611.02731v2.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes, may 2006. arXiv:1312.6114v10.