University of Liege

Master Thesis Summary

Energy-based Multi-Modal Attention

Author:
Aurélien Werenne

Supervisor:
Dr. Raphaël Marée

June 3, 2019

Contents

1 Introduction 2
2 Problem Statement 3
3 EMMA 3
3.1 Framework Overview 3
3.2 Energy-step 5
3.3 Composition-step 8
3.4 Attention-step 9
4 Regularization 11
5 Experiments & Results 14
5.1 Experiment 1 14
5.2 Experiment 2 16
6 Appendix 17

1 Introduction

In our daily lives we often use multiple resources to gain a maximal amount of information about a particular situation. For example, when we are at a museum admiring a painting, both the image of the painting itself as the written description next to it gives us different but complementary information. The term *modality* is generally understood to mean a way in which something happens. Multimodal Deep Learning (MMDL) has received much attention in the last three years [1, 6, 8], it aims at building models that can process and combine information from multiple modalities. One of the success stories using MMDL is the work of Afouras et al. [1] using both audio and images of lip movements to obtain state-of-the art speech recognition results. Throughout this report we use the terms modality and mode interchangeably.

Previous work has mostly been focused on studying the best ways to combine different modalities. A major flaw of those models is their lack of robustness in real-world settings. Indeed the quality and usefulness of the modes can vary with time, and are dependent on the context. The desired behaviour of multimodal systems would be to adapt and distribute attention to the current most relevant modes. Humans have this ability, often called *cross-modal attention* [5]. A situation we all have encountered in which we use cross-modal attention is the cocktail-party effect [4], it refers to the difficulty we sometimes have understanding speech in noisy social settings. As a subconscious response, we tend to look at the mouth of our interlocutor i.e. we shift some attention from the auditory to the visual senses.

In this work I introduce a novel deep learning mechanism, *Energy-based Multi-Modal Attention* (EMMA), that aims at mimicking human's cross-modal attention. More specifically, the task of EMMA is to decide how much attention to devote to each mode taking several factors into account: noisiness, correlations between modes, unseen/missing values. The focus of this work is not to create a new MMDL architecture, but instead to design a module to be inserted in an already-trained MMDL model in order to increase its robustness.

The report is organized into five parts. First, section 2 formalizes the goal and setting of the problem. Next, the proposed architecture is described. Section 4 discusses a new regularizer to use during the training of the EMMA module. And to finish, section 5 examines the results of several experiments.

2 Problem Statement

A Deep Neural Network for multi-modal data $\mathbf{x}$ is typically constructed as a many-to-one encoder-decoder. Each modality goes through a different encoder extracting relevant features, those features are then fused together in the decoder. The decoder is trained to make predictions $\hat{y}$ as close as possible to the groundtruth y.

We define the i.i.d. dataset $(\mathbf{x}, y) \in \mathcal{D}^{(N)}$ of size N, all samples $\mathbf{x}$ are composed of M modalities. Each encoder i extracts a *feature vector* $\mathbf{m}_i$. Multi-modal machine learning is used to solve problems if it reduces the entropy about the output as follows

$$I(y\,;\mathbf{m}_i) \leq I(y\,;\mathbf{m}_i|\{\mathbf{m}_j\}_{1:M,i\neq j}) \tag{1}$$

A few examples of feature vectors are Fourier transforms of audio waves, flattened activation maps of Convolutional Neural Networks and context vectors produced by Recurrent Neural Networks.

The EMMA module is inserted in between the encoders and decoder (see Fig.1) to apply an attention operation to every feature vector as

$$\mathbf{m}'_i = \beta_i \mathbf{m}_i \quad \text{with} \quad \beta_i \in [0, 1] \tag{2}$$

where *attention factors* β_i are a function of all the feature vectors. Intuitively, if β_i is close to zero, the corresponding vector $\mathbf{m}'_i$ will also be close to zero and thus triggers less activations. To put it in other words, the influence of a mode on predictions is decreased for a low attention factor. Low values of β_i are desired for modes with low information but also for noisy/missing/unseen values. The latter three types of values are regrouped under the name *abnormal values*.

3 EMMA

3.1 Framework Overview

The module is composed of three sequential steps: the *Energy-*, the *Composition-* and the *Attention-step*. The main idea behind the three steps of the procedure are briefly outlined below.

Energy-based models (EBM, [9]) approximate the negative log-likelihood of the data. The approximated value is referred to as the energy, which will take low values on the data manifold and will be high everywhere else. Thus, we now have a way to quantify how normal/abnormal the modes are. However such

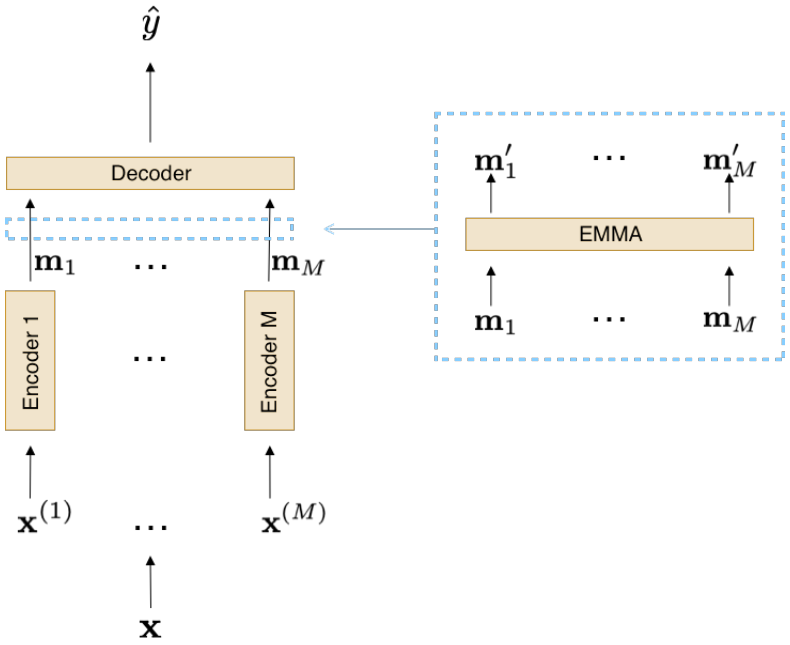


Figure 1: *A typical multi-modal model structure. The blue dashed rectangles show how an EMMA module can be inserted in between the encoders and the decoder.*

models are difficult to optimize in practice. An alternative approach is to use De-noising Autoencoders (DAE, Vincent et al. [10]), it was shown in [7] it is possible to derive the shape of the true energy function, the so-called *potential energy*.

Potential energies help us distinguish normal from abnormal values for a specific mode, but they do not express correlations with other modes or even how important the mode is for making good predictions. A parametric function composed of potentials, *modal energies*, are introduced to solve this issue. Finally, the Ising model [11] normalizes modal energies to the unit interval as *importance weights*.

Our attention mechanism is based on the work of the famous psychologist and economist Daniel Kahneman. He considers attention as a resource to be distributed among various stimuli [3], in contrast to the more classical viewpoint where attention is a selective process. Kahneman describes attention in terms of a capacity. Accordingly, during the training EMMA will learn how much the optimal capacity is for the task, it will also find the best filtering threshold.

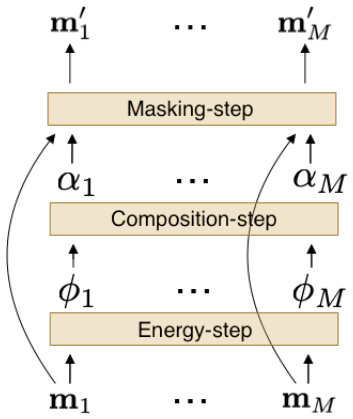


Figure 2: *High-level view of the architecture of EMMA.*

3.2 Energy-step

Autoencoders and potential energy

Autoencoders (AE) are simple models trained to attempt to copy their inputs to outputs. They are made out of two parts: an encoder $\mathbf{z}_i = f(\mathbf{m}_i)$ and a decoder $g(\mathbf{z}_i)$. The input-to-output mapping is called the reconstruction and is written as $r(\mathbf{m}_i) = g(f(\mathbf{m}_i))$. In this work we use an AE of the following specific architecture:

$$f(\mathbf{m}_i) = \sigma(W\mathbf{m}_i + \mathbf{b}_h)\,, \tag{3}$$

$$g(\mathbf{z}_i) = W^T\mathbf{z}_i + \mathbf{b}_r \tag{4}$$

Using tied weights is an important constraint necessary to derive the energy function.

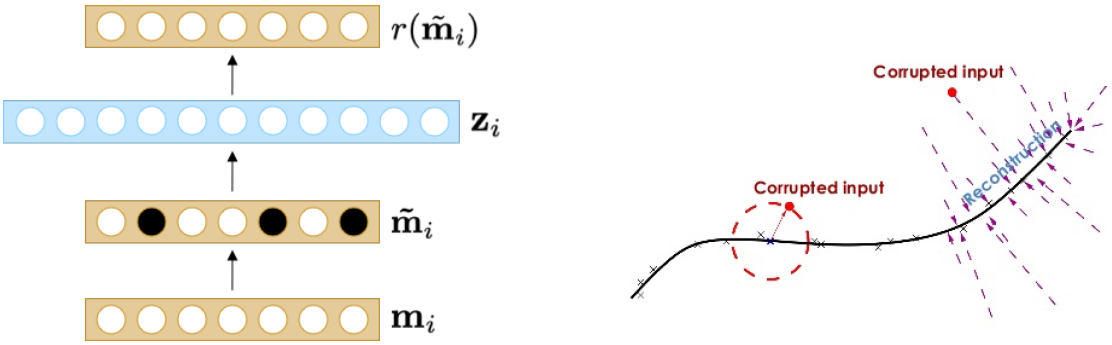
In this work we use a regularized variation of the AE: the de-noising autoencoder (Vincent et al. [10]). This variation is obtained by adding a corruption process to the inputs, $\tilde{\mathbf{m}}_i = \mathbf{m}_i + \epsilon$ (see Fig.3). De-noising autoencoders are then trained to minimize the reconstruction loss,

$$\mathcal{L}_{\text{DAE}} = \|r(\tilde{\mathbf{m}}_i) - \mathbf{m}_i\|_2^2 \tag{5}$$

Interestingly, Alain & Bengio [2] demonstrated that the reconstruction vector of a de-noising autoencoder is proportional to the score (log-likelihood),

$$r(\tilde{\mathbf{m}}_i) - \mathbf{m}_i \propto \frac{\partial \log p(\mathbf{m}_i)}{\partial \mathbf{m}_i} \tag{6}$$

To put it in another way, the reconstruction minus the input vector points towards the most likely corresponding de-noised input. We can visualize the vector field of reconstructions evaluated in the region around a toy data manifold in Fig.4. Notably, the data manifold acts as a sink in the vector field of the input space.



(a) *Architecture*

(b) *Denoising process. (Image from [10])*

Figure 3: *Denoising Autoencoder (DAE).*

Furthermore, Kamyshankska & Memisevic [7] observed that those vector fields are also gradient fields, due to the fact that the integrability criterion is satisfied when using tied weights[1]:

$$\frac{\partial(r_k(\tilde{\mathbf{m}}_i) - m_{i;k})}{\partial m_{i;l}} = \frac{\partial(r_l(\tilde{\mathbf{m}}_i) - m_{i;l})}{\partial m_{i;k}} \tag{7}$$

Therefore, the vector field can be written as the derivative of a scalar field. In analogy to physics, the scalar field may then be thought of as a potential energy and the vector field as a corresponding force. [7] derives the potential energy function as,

$$\begin{aligned} \mathrm{V}_i &= -\int (r(\tilde{\mathbf{m}}_i) - \mathbf{m}_i) d\mathbf{m}_i \\ &= -\int f(\mathbf{m}_i) d\mathbf{m}_i - \frac{1}{2}\|\mathbf{m}_i + \mathbf{b}_r\|_2^2 + \text{const} \\ &= -\sum_k \log(1 + \exp(W_{.k}^T \mathbf{m}_i + b_k^h)) + \frac{1}{2}\|\mathbf{m}_i - \mathbf{b}_r\|_2^2 + \text{const} \end{aligned} \tag{8}$$

Details of the computation can be found in Appendix B. By simply integrating

[1] See appendix A

Eq.6, we can corroborate the significance of the potential energy V_i:

$$\begin{aligned} r(\tilde{\mathbf{m}}_i) - \mathbf{m}_i &\propto \frac{\partial \log p(\mathbf{m}_i)}{\partial \mathbf{m}_i} \\ -\int (r(\tilde{\mathbf{m}}_i) - \mathbf{m}_i) d\mathbf{m}_i &\propto -\log p(\mathbf{m}_i) \\ \mathrm{V}_i &\propto -\log p(\mathbf{m}_i) + \text{constant} \end{aligned} \tag{9}$$

The potential energy V_i is proportional to the true energy, the negative log-likelihood (if we consider the constant as the partition function).

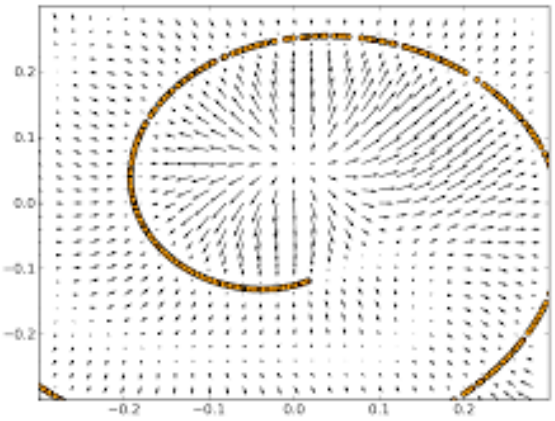


Figure 4: *Vector field of the reconstruction error $r(\tilde{\boldsymbol{m}}_i) - \boldsymbol{m}_i$, and the samples of the data manifold. (Image from [2])*

Abnormal values

Another interesting quantity is the Laplacian of the energy function, $\Delta \mathrm{V}_i$. It can be computed using the expression[2] below,

$$\Delta \mathrm{V}_i = \mathrm{Tr}[W^T \,\mathrm{diag}(\sigma'(W\mathbf{m}_i + \mathbf{b}_h))\, W] - M \quad \text{with} \quad \sigma'(x) = \sigma(x)(1-\sigma(x)) \tag{10}$$

The Laplacian is the divergence of the gradient and will be negative at a sink. Since we know that the data manifold behaves as a sink in its vector field, this helps us construct a simple normal/abnormal classifier,

$$\eta(\mathbf{m}_i) = \begin{cases} \eta_+ & \text{if } \Delta \mathrm{V}_i < 0 \\ \eta_- & \text{otherwise} \end{cases} \tag{11}$$

We will appreciate how relevant the classifier η is in Section 4.

[2] Appendix C

3.3 Composition-step

From the previously computed potential energies we can construct composed energy functions, modal energies, of the general form

$$E_i = f(\mathrm{V}_i) + \sum_{k \neq i}^{M} g(f(\mathrm{V}_i), f(\mathrm{V}_k)) \tag{12}$$

The reason of this transformation is three-fold. First, the potential energies are approximations of the shape of the true negative-log likelihood (Eq.9, as a consequence potential energies are possibly be on very different scales with different biases. Secondly, as stated in [2] there can be "bad" attractors in the vector fields, leading to erroneous values of the energy function V_i. By leveraging the correlations between the energy functions (via g), the modal energies can be more robust to those errors. Lastly, the parameters of the modal energies can be trained to minimize the loss with respect to predictions. For example, the function g can be designed as a switch, if $g(\mathrm{V}_i, \mathrm{V}_j)$ and $g(\mathrm{V}_j, \mathrm{V}_i)$ are asymmetric.

As a side-note, the composition step is the most "engineered" part of the model. There are thus many other ways to design the system that could possibly obtain better results.

A problem we overlooked so far is that we neglected the integration constant (corresponding to the partition function) in the potential energy computation. As a consequence, V_i is biased. To correct the bias, an easy fix is to subtract the lowest value potential energy $\mathrm{V}^{(i)}_{min}$ of the training samples. At the same time we can multiply it by a weight to solve the scaling problem:

$$\phi_i = w_i V_i' + b_i \ (= f(\mathrm{V}_i)) \quad \text{with} \quad V_i' = \max(\mathrm{e}, \mathrm{V}_i - \mathrm{V}^{(i)}_{min} + \mathrm{e}) \tag{13}$$

Instead of mapping the lowest values to zero, we map it to e, Euler's number. This is because the computation of the gradient during the optimization involves a logarithm of ϕ_i[3]. Hence, to prevent exploding gradients we add e. The reason we use a max-operator is because lower energy values than $\mathrm{V}^{(i)}_{min}$ can occur during inference.

Now, let us introduce the building blocks of each modal energy i, called *partial energies*:

$$E_{ij} = \phi_i^{\gamma_{ij}} \phi_j^{1-\gamma_{ij}} \ (= g(f(\mathrm{V}_i), f(\mathrm{V}_j))) \tag{14}$$

[3]See appendix D

with $w_i \in \mathbb{R}^+$, $\gamma_{ij} \in [0, 1]$. Note that $E_{ii} = \phi_i$. Partial energies satisfy the switch property previously discussed if we add the constraint $\gamma_{ij} = \gamma_{ji}$. As a consequence, the two modes i and j will be strongly dependent when $gamma_{ij}$ equals zero. At the opposite spectrum, they will be completely decoupled when gamma tends to one.

Furthermore, modal energies are defined as sum over their respective partial energies:

$$E_i = \sum_{k=1}^{M} E_{ik} \tag{15}$$

Finally, the system energy $\mathcal{S}$ is the sum of all modal energies, which can be used as an indicator of how uncertain the model is about the data.

The importance weights are computed as likelihoods in a Gibbs distribution:

$$\alpha_i = \frac{1}{Z} e^{-\tau E_i} \quad \text{with the partition function} \quad Z = \sum_{i}^{M} e^{-\tau E_i} \tag{16}$$

where the *coldness* τ is a hyperparameter controlling the amount of entropy in the system. At high temperature/low coldness the distribution is uniform, and at low temperature/high coldness the importance weight corresponding to the lowest energy tends to 1, while the other importance weights approach zero.

3.4 Attention-step

Attention factors

The attention factors β_i are determined from the importance weights:

$$\beta_i = \text{ReLU}(\tanh(G\alpha_i - b)) \tag{17}$$

where the gain (G) and bias (b) are learnable in $\mathbb{R}^+$. As previously mentioned, the output feature vectors $\mathbf{m}'_i$ are obtained by multiplying the inputs $\mathbf{m}_i$ with the factors β_i (Eq.2). Thus,

$$\text{if } \alpha_i \leq \text{threshold then } \mathbf{m}'_i \approx 0\,, \text{ else } \mathbf{m}'_i \to \mathbf{m}_i \tag{18}$$

We can distinguish those phases if we plot the activation function (see Fig.5). The convergence rate of $\mathbf{m}'_i$ to $\mathbf{m}_i$ is characterized by the gain. Whereas, the output $\mathbf{m}'_i$ depends on both parameters G and b. In particular, $\mathbf{m}'_i$ is completely masked

out if

$$\begin{aligned} &\Leftrightarrow G\alpha_i - b \leq 0 \\ &\Leftrightarrow \log(\alpha_i) \leq \log(b/G) \\ &\Leftrightarrow E_i \geq \frac{\log(G/b) - \log(Z)}{\tau} \end{aligned} \tag{19}$$

From the inequality we deduce coldness also influences the masking threshold. When coldness equals zero, the inequality is never verified and thus their is no masking.

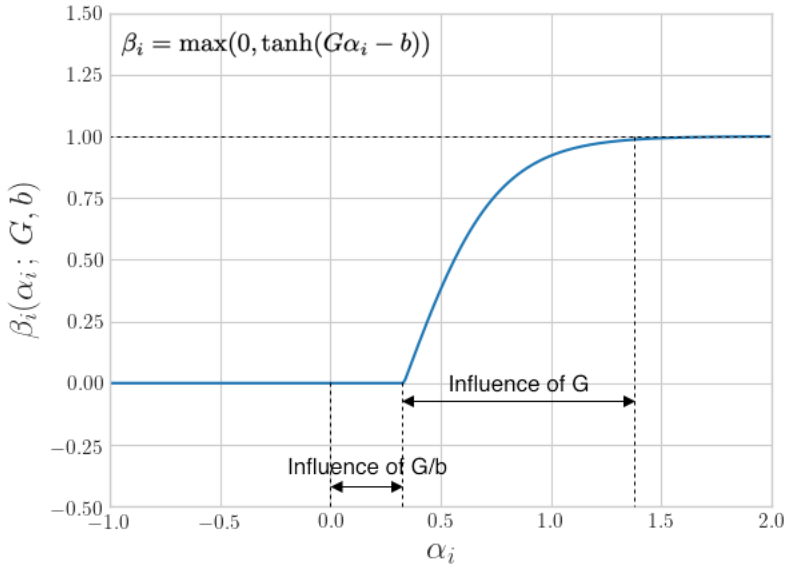


Figure 5: *Activation function representing the masking factor.*

Interpretability & Symmetry

The choice of having a common gain and bias is unconventional and makes the system less expressive. However, the cost of keeping them common is motivated by two properties:

- *Interpretability*: one nice advantage of EMMA is its interpretability, indeed importance weights quantify how much attention is applied to each mode. Using separate gains and biases makes importance weights less meaningful.
- *Symmetry*: using separate parameters G_i, b_i breaks the symmetry of the attention operation. To illustrate this, take the case where we have two

modalities, one contributing in general much more (strong mode) than the other one (weak mode). The learned masking factors would look approximately as draw in Fig.6. For most samples attention will work as desired. However a problem occurs if the sample has a noisy strong mode: even though $\alpha_{\text{weak}} \approx 1$, its corresponding factor β_{weak} is still small. As a consequence, both modes will be masked out.

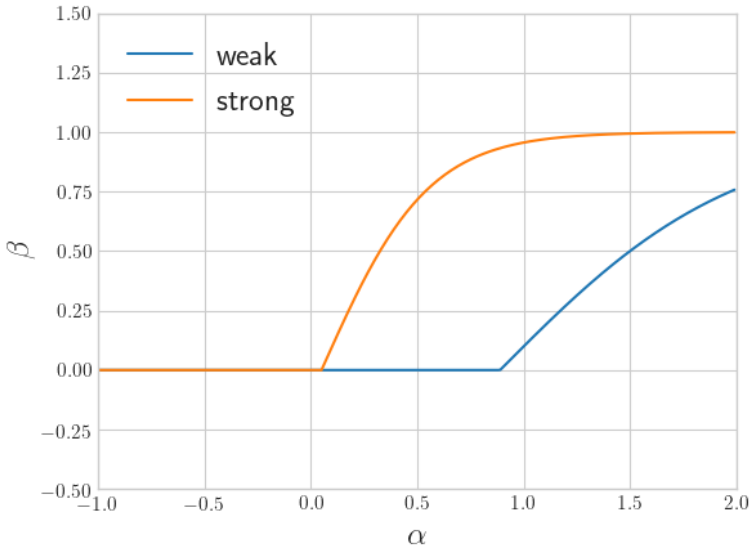


Figure 6: *Asymmetric effect.*

4 Regularization

To avoid modal energies to be too different from their original potential energies, I introduce a regularization term based on the classifier η. The loss per sample is written as,

$$\tilde{\mathcal{L}} = \mathcal{L}(y, \hat{y}) - \lambda\Omega \quad \text{with} \quad \Omega = \sum_{k=1}^{M} \eta(\mathbf{m}_k) \log(\alpha_k) \tag{20}$$

Intuitively, normal/abnormal modes have positive/negative η, thus as an effect of Ω the importance weights α_i will be maximized/minimized. Additionally, we show below that under some conditions Ω operates as a contrastive divergence optimization.

Let $\theta = \{\boldsymbol{\gamma}, \mathbf{w}, \mathbf{b}\}$ be the set of all parameters of the composition step. Now we can compute the gradient of the regularizer with respect to those parameters,

$$\nabla_\theta \Omega = \sum_{k=1}^{M} \eta(\mathbf{m}_k) \nabla_\theta \log(\alpha_k) \tag{21}$$

The gradient of the logarithm becomes

$$\begin{aligned}
\nabla_\theta \log(\alpha_k) &= \nabla_\theta \log \left[\frac{e^{-\tau E_k}}{Z} \right] \\
&= \nabla_\theta(-\tau E_k) - \nabla_\theta \log \sum_{l=1}^{M} e^{-\tau E_l} \\
&= -\tau \nabla_\theta E_k - \frac{\sum_{l=1}^{M} \nabla_\theta e^{-\tau E_l}}{\sum_{l=1}^{M} e^{-\tau E_l}} \\
&= -\tau \nabla_\theta E_k + \tau \frac{\sum_{l=1}^{M} \nabla_\theta (E_l) e^{-\tau E_l}}{\sum_{l=1}^{M} e^{-\tau E_l}} \\
&= \tau \left[-\big(1 - \frac{e^{-\tau E_k}}{Z}\big) \nabla_\theta E_k + \sum_{l \neq k}^{M} \frac{e^{-\tau E_l}}{Z} \nabla_\theta E_l \right] \\
&= \tau \left[-\big(1 - \alpha_k\big) \nabla_\theta E_k + \sum_{l \neq k}^{M} \alpha_l \nabla_\theta E_l \right]
\end{aligned} \tag{22}$$

We take the part of the gradient relative to the specific subset of parameters $\theta_i = \{[\gamma_{ik}]_{k=1}^{M}, w_i, b_i\}$. Equation 22 can be split as follows

$$\nabla_{\theta_i} \log(\alpha_k) = \begin{cases} -\tau(1 - \alpha_i) \nabla_{\theta_i} E_i, & \text{if } i = k \\ \tau \alpha_i \nabla_{\theta_i} E_i, & \text{if } i \neq k \end{cases} \tag{23}$$

Let us define $M' = \lfloor \frac{M+1}{2} \rfloor$, the number of normal modes in the sample during the training phase. Next, by substituting (23) in (21) and extrapolating the sums we obtain the expression of $\nabla_{\theta_i} \Omega$ if mode i is normal

$$\eta_+ \big[-\tau(1 - \alpha_i) \nabla_{\theta_i} E_i \big] + \big[(M' - 1)\eta_+ + (M - M')\eta_- \big] \alpha_i \tau \nabla_{\theta_i} E_i \tag{24}$$

and $\nabla_{\theta_i} \Omega$ if i is abnormal

$$\eta_- \big[-\tau(1 - \alpha_i) \nabla_{\theta_i} E_i \big] + \big[M'\eta_+ + (M - M' - 1)\eta_- \big] \alpha_i \tau \nabla_{\theta_i} E_i \tag{25}$$

Which can be summarized into the following equation if we pose $\eta = +1$ and $\eta = -1$:

$$\boxed{\nabla_{\theta_i} \Omega = -\big[(M - 2M')\alpha_i + \text{sign}(\eta_i) \big] \tau \nabla_{\theta_i} E_i} \tag{26}$$

Two cases can be distinguished:

- If the total number of modes M, is even then

$$\nabla_{\theta_i}\Omega = -\text{sign}(\eta_i)\tau\nabla_{\theta_i}E_i \tag{27}$$

  Minimizing the loss function $\tilde{\mathcal{L}}$, will maximize the regularizer Ω. Ignoring the second-order terms in the Taylor expansion of modal energies, we can conclude from equation (27) that gradient descent will update values of the energy function E_i downward/upward for normal/abnormal inputs.

- If M is uneven,

$$\nabla_{\theta_i}\Omega = \begin{cases} -(1-\alpha_i)\tau\nabla_{\theta_i}E_i & \text{if input } i \text{ is normal} \\ (1+\alpha_i)\tau\nabla_{\theta_i}E_i & \text{otherwise} \end{cases} \tag{28}$$

  In analogy, modal energies are forced to stay close to their original potential energies. Additionally, an interesting phenomenon occurs: high energies that have to be low and low energies that have to be high will have stronger gradients than their counterparts. This corresponds to the positive and negative phase in contrastive divergence optimization.

5 Experiments & Results

5.1 Experiment 1

In this experiment we train a DAE as described by equations 3 and 4 on manifolds in a two-dimensional space. The autoencoder has 8 hidden units and is trained on 4 types of manifolds:

- Linear manifold (first row of Fig.1)
- Sinus manifold (second row of Fig.1)
- Circular manifold (third row of Fig.1)
- Spiral manifold (fourth row of Fig.1)

We can observe from the results that the data distribution is more or less captured in the vector field of the autoencoder. Moreover, we can clearly see a false positive for the reconstruction of the linear manifold. This is because the erroneous region is a source in the vector field, which will thus be reflected in the reconstruction (small norm) but not in the potential heatmap. Néanmoins on n'obtient pas des résultats aussi précis que dans la FIg.4 par exemple. Il y a deux raisons à ce problème. Premièrement, je me suis rendu compte d'un léger bug dans le design du corruption process, ce qui cause le vecteurs à être plus dirigés vers le centre du manifold que vers les extrèmités. Et deuxièment, je n'ai pas beaucoup joué avec la capacité du modèle et tester d'autres fonctions d'activations.

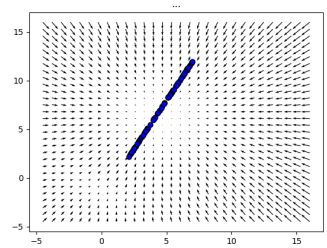


Figure 7: *Vector field*

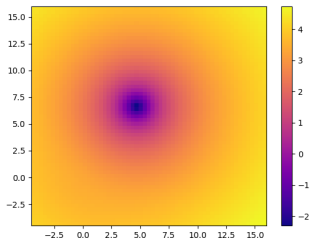


Figure 8: *Potential (log-scale)*

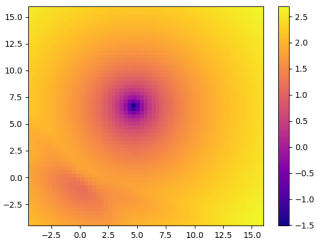


Figure 9: *Reconstruction norm (log-scale)*

Table 1: *Results of experiment 1*

5.2 Experiment 2

In this second experiment we augment the capacity of our autoencoder and we train it on the zero digits of the MNIST dataset. We then perform two tests:

- How does the potential energy evolve on a sample (100 zero images) with more and more noise? (Fig.10)
- How does the potential energy change on a sample of seen data (0 digit) versus a sample of unseen data (1 digit)? (Fig.11)

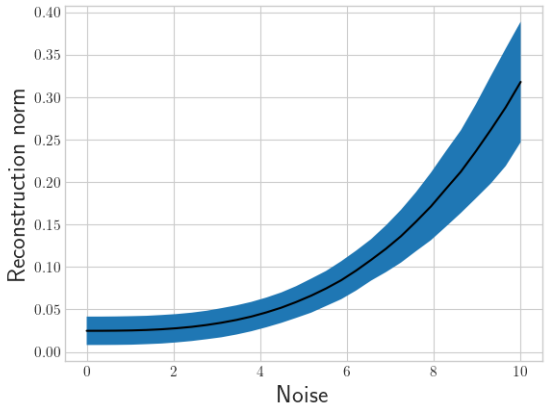


Figure 10: *Noise versus potential energy. Error bands represent two times the standard deviation.*

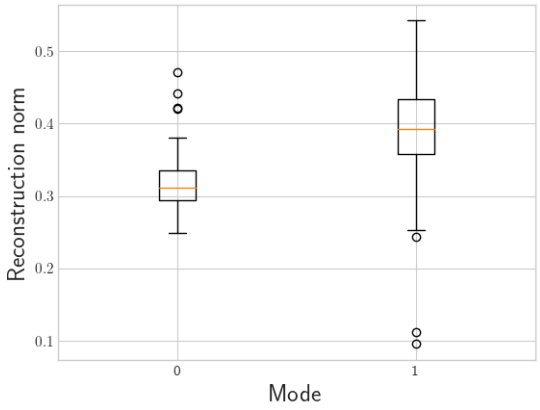


Figure 11: *Unseen/seen data and potential energy.*

6 Appendix

Appendix A, B and C are developments made by [7] and adapted to our formalism.

Appendix A

The integrability criterion holds since

$$\begin{aligned}\frac{\partial(r_k(\tilde{\mathbf{m}}_i) - m_{i;k})}{\partial m_{i;l}} &= \sum_j W_{kj} \frac{\partial \sigma(W\tilde{\mathbf{m}}_i + \mathbf{b}_h)}{\partial(W\tilde{\mathbf{m}}_i + \mathbf{b}_h)} W_{lj} - \delta_{kl} \\ &= \frac{\partial(r_l(\tilde{\mathbf{m}}_i) - m_{i;l})}{\partial m_{i;k}}\end{aligned} \tag{29}$$

where δ_{kl} denotes the Kronecker delta.

Appendix B

$$\begin{aligned}\mathrm{V}_i &= -\int (r(\tilde{\mathbf{m}}_i) - \mathbf{m}_i) d\mathbf{m}_i \\ &= -\int (W^T \sigma(W\mathbf{m}_i + \mathbf{b}_h) + \mathbf{b}_r - \mathbf{m}_i) d\mathbf{m}_i \\ &= -W^T \int \sigma(W\mathbf{m}_i + \mathbf{b}_h) d\mathbf{m}_i - \int (\mathbf{b}_r - \mathbf{m}_i) d\mathbf{m}_i\end{aligned} \tag{30}$$

Let $\mathbf{u} = W\mathbf{m}_i + \mathbf{b}_h$ be the auxiliary variable such that

$$\frac{d\mathbf{u}}{d\mathbf{m}_i} = W^T \Leftrightarrow d\mathbf{m}_i = W^{-T} d\mathbf{u} \tag{31}$$

we can write,

$$\begin{aligned}\mathrm{V}_i &= -\int W^T W^{-T} \sigma(\mathbf{u}) d\mathbf{u} - \mathbf{b}_r^T \mathbf{m}_i + \frac{1}{2}\|\mathbf{m}_i\|_2^2 + \text{const} \\ &= -\int \sigma(\mathbf{u}) d\mathbf{u} + \frac{1}{2}\|\mathbf{m}_i - \mathbf{b}_r\|_2^2 + \text{const} \\ &= -\int (1 + e^{-\mathbf{u}})^{-1} d\mathbf{u} + \frac{1}{2}\|\mathbf{m}_i - \mathbf{b}_r\|_2^2 + \text{const} \\ &= -\sum_k \log(1 + e^{W_{.k}^T \mathbf{m}_i + b_k^h}) + \frac{1}{2}\|\mathbf{m}_i - \mathbf{b}_r\|_2^2 + \text{const}\end{aligned} \tag{32}$$

Appendix C

The Laplacian of the potential energy is defined as

$$\mathrm{V}(\mathbf{m}_i) = \sum_k \frac{\partial^2 \mathrm{V}_i}{\partial m_{i;k}^2} \tag{33}$$

If we know that,

$$\frac{\partial \mathrm{V}_i}{\partial \mathbf{m}_i} = r(\mathbf{m}_i) - \mathbf{m}_i = W^T \sigma(W\mathbf{m}_i) - \mathbf{m}_i \tag{34}$$

The Laplacian can be written as

$$\begin{aligned} \frac{\partial^2 \mathrm{V}_i}{\partial \mathbf{m}_i^2} &= W^T \mathrm{diag}(\sigma'(\mathbf{m}_i))W - I_M \\ &= \mathrm{Tr}\big[W^T \mathrm{diag}(\sigma'(\mathbf{m}_i))W\big] - M \end{aligned} \tag{35}$$

Appendix D

The gradient of the loss with respect to γ_{ij} is computed with the chain rule:

$$\frac{\partial \mathcal{L}}{\partial \gamma_{ij}} = \frac{\partial \mathcal{L}}{\partial E_i} \cdot \frac{\partial E_i}{\partial \gamma_{ij}} + \frac{\partial \mathcal{L}}{\partial E_j} \cdot \frac{\partial E_j}{\partial \gamma_{ij}} \tag{36}$$

Furthermore,

$$\begin{aligned} \frac{\partial E_i}{\partial \gamma_{ij}} &= \frac{\partial}{\partial \gamma_{ij}} \sum_{k=1}^{M} E_{ik} \\ &= \frac{\partial E_{ij}}{\partial \gamma_{ij}} \\ &= \frac{\partial}{\partial \gamma_{ij}} (\phi_i^{\gamma_{ij}} \phi_j^{1-\gamma_{ij}}) \\ &= \phi_i^{\gamma_{ij}} \frac{\partial}{\partial \gamma_{ij}} \phi_j^{1-\gamma_{ij}} + \phi_j^{1-\gamma_{ij}} \frac{\partial}{\partial \gamma_{ij}} \phi_i^{\gamma_{ij}} \\ &= \phi_i^{\gamma_{ij}} \phi_j^{1-\gamma_{ij}} (\log \phi_i + \log \phi_j) \end{aligned} \tag{37}$$

Bibliography

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep Audio-Visual Speech Recognition. *arXiv e-prints*, page arXiv:1809.02108, Sep 2018.

[2] Guillaume Alain and Yoshua Bengio. What Regularized Auto-Encoders Learn from the Data Generating Distribution. *arXiv e-prints*, page arXiv:1211.4246, Nov 2012.

[3] Brian Bruya and Yi-Yuan Tang. Is attention really effort? revisiting daniel kahneman's influential 1973 book attention and effort. *Frontiers in Psychology*, 9:1133, 2018.

[4] Cocktail party effect. Cocktail party effect — Wikipedia, the free encyclopedia, 2010. [Online; accessed 29-April-2019].

[5] Jon Driver and Charles Spence. Crossmodal attention. *Current Opinion in Neurobiology*, 8(2):245 – 253, 1998.

[6] Ayush Jaiswal, Ekraam Sabir, Wael AbdAlmageed, and Premkumar Natarajan. Multimedia Semantic Integrity Assessment Using Joint Embedding Of Images And Text. *arXiv e-prints*, page arXiv:1707.01606, Jul 2017.

[7] Hanna Kamyshanska and Roland Memisevic. The Potential Energy of an Autoencoder. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI)*, 2014.

[8] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to Combine Modalities in Multimodal Deep Learning. *arXiv e-prints*, page arXiv:1805.11730, May 2018.

[9] Takayuki Osogami. Boltzmann machines and energy-based models. *arXiv e-prints*, page arXiv:1708.06008, Aug 2017.

[10] Pascal Vincentl, Bengio Yoshua Larochelle, Hugo, , and Pierre-Antoine Manzagol. Extracting and Composing Robust Features with Denoising Autoencoders. *ICML 2008*, 2008.

[11] C. N. Yang. The spontaneous magnetization of a two-dimensional ising model. *Phys. Rev.*, 85:808–816, Mar 1952.