Energy-based Multi-Modal Attention

AURELIEN WERENNE

Master Thesis
2018-2019

Energy-based Multi-Modal Attention

Author:
Aurélien Werenne

Supervisor:
Dr. Raphaël Marée

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Computer Science and Engineering

Montefiore Institute
Faculty of Applied Sciences
Liège, Belgium

Academic Year 2018 - 2019

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed nec ultrices felis, eu suscipit orci. Nunc laoreet odio velit. Mauris tincidunt quam at purus vulputate vehicula. Etiam eu purus lorem. Sed hendrerit condimentum tellus, ut posuere libero laoreet nec. Sed maximus, tortor aliquam mollis blandit, lectus urna efficitur eros, at laoreet ante turpis suscipit quam. Phasellus eget sollicitudin felis. Sed enim nunc, rutrum vel velit ut, elementum tempor magna. Nullam ut tincidunt orci, et interdum lectus. Proin sed imperdiet tellus. Donec pharetra feugiat leo at fringilla.

Suspendisse consectetur maximus augue. Etiam eu tempor ipsum. Phasellus tempor at purus non cursus. Sed non quam vitae mi rutrum sodales vel posuere odio. Nunc elit arcu, finibus sit amet euismod et, aliquet vel mauris. Mauris eget enim lacus. Donec feugiat eget neque vitae dictum. Nullam turpis neque, mollis at dui quis, lobortis molestie quam. Sed faucibus arcu in odio venenatis, et eleifend lacus lobortis. Pellentesque tincidunt ante non mauris molestie efficitur. Nullam porta massa nulla, at lobortis ex pulvinar sed. Donec auctor consectetur ante, vitae vehicula odio gravida nec. Fusce arcu leo, imperdiet vel magna eu, ullamcorper lacinia orci.

Aliquam vulputate magna lectus, at consequat ante congue et. Sed congue ullamcorper erat, ut volutpat libero consectetur sit amet. Nulla gravida ullamcorper odio, in convallis quam congue sit amet. Nullam tempor ullamcorper pulvinar. Pellentesque rutrum massa eu massa mattis iaculis. Fusce rhoncus tortor id est cursus interdum. Morbi urna elit, commodo sed nisl eget, scelerisque porta dui. Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Sed nec egestas ligula.

Acknowledgements

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed nec ultrices felis, eu suscipit orci. Nunc laoreet odio velit. Mauris tincidunt quam at purus vulputate vehicula. Etiam eu purus lorem. Sed hendrerit condimentum tellus, ut posuere libero laoreet nec.

Contents

List of Figures

Notation and Abbreviations

Ω	Regularizer
$E^{(T)}$	True energy
$E^{(P)}$	Partial energy
$E^{(M)}$	Modal energy
$E^{(S)}$	System energy
Ψ	Potential energy

ML	**M**achine **L**earning
DL	**D**eep **L**earning
EMMA	**E**nergy-based **M**ulti-**M**odal **A**ttention module
AE	**A**uto**e**conder
DAE	**D**enoising **A**uto**e**conder

Chapter 1

Introduction

1.1 Motivation

- unimodal. AI – DL – lots of applications/advances/succes-stories in recent years perception (not reasoning yet). Cite important papers/reviews and exciting applications (autonomous driving, . . .)
- Cite/explain two multimodal examples: heart disease from multiple sensors (find better example) and lip/audio speech recognition. This is multimodal (more general than multisensorial, more on this in Section Background)
- Common good practice in engineering is if one sensor/process fails, other will take over (reliability engineering - what fails/what happens?). Humans cocktail-party effect, cross-modal attention. But also unseen, with lip beard example. Notion of robustness generalization (without saying technical word).
- Most AI systems are tested in ideal situations, not realistic. Imagine when one sensor fails, or noisy or unknown (outlying).. AI systems are not explicitly prepared for this.
- I introduce EMMA that solves previous paragraph problem, uncertainty, interpretability.
- What this thesis is about

crossmodal attention

1.2 Proposed solution

Introduce model + modes + prediction. We will only see DL models because better at perception.

Problem is when outlying – possibly high activations – drop in performances (more on this in next chapter). Principle of solution (loosily inspired by dropout) is quiet simple (oversimplified on purpose): outlying is multiplied by small number close/greater to zero, other by number close/less to 1 (unchanged). Explain diagram only test-time (not speaking about training).

all experiments and model are available code link. refer that more details in appendix.

1.3 Thesis Outline

High level overview of this work. See other thesis. Contributions?

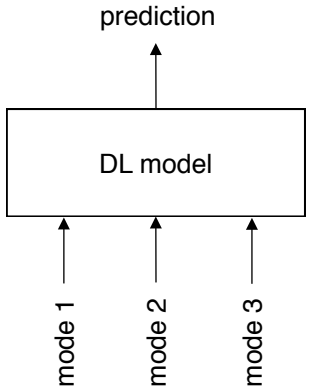


FIGURE 1.1: Main idea

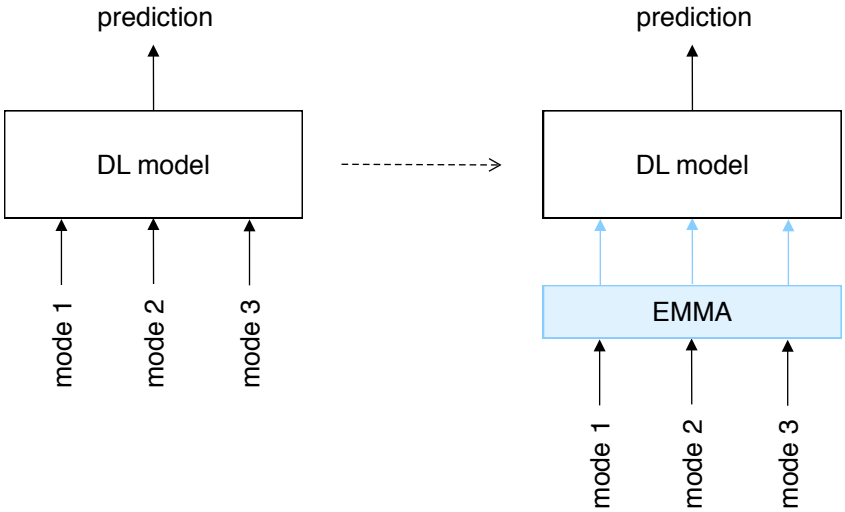


FIGURE 1.2: Main idea

- **chapter 1**: Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed nec ultrices felis, eu suscipit orci. Nunc laoreet odio velit. Mauris tincidunt quam at purus vulputate vehicula. Etiam eu purus lorem. Sed hendrerit condimentum tellus, ut posuere libero laoreet nec.

- **chapter 2**: Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed nec ultrices felis, eu suscipit orci. Nunc laoreet odio velit. Mauris tincidunt quam at purus vulputate vehicula. Etiam eu purus lorem. Sed hendrerit condimentum tellus, ut posuere libero laoreet nec.

Chapter 2

Background

Explain goal of chapter: introduce necessary background and concepts used in rest of work. Notation: vectors in bold.

2.1 Machine Learning

- supervised, unsupervised, RL
- supervised
  - task T, performance P, data D – D high, P high on T
  - types: generative , discriminative (classification, regression) are most common tasks T

2.1.1 Deep Learning

- subcategory of ML (difference with others, it extracts its own features)
- Data distribution, manifold (explain why DL works, intuition of manifold, definition mathematics)
- Cost, (surrogate) loss (crossentropy), metric, gradient, forward/backward (backprop, update optimization)
- Gigantic non-linear composed parametric functions
- MLP, CNN, LSTM

2.2 Modalities

- More general than sensorial. Give examples. Toy example of multimodal DL.
- Outlying. What happens on DL model. Show on our simple toy example, noise – high activations – wrong prediction – model seems certain, we cannot know model was uncertain

2.3 Energy-based models and Statistical Mechanics

- Negative log likelihood. Ising model (thermodynamics, stat mechanics). Gibbs distribution. Explain how relate to outlying (high E, low p). EBM = high everywhere except on data manifold.

- Boltzmann. MC training. Contrastive divergence. Explain why difficult (see paper of potential energy, good explanation!!). Thus we will explore approximators.
- To avoid confusion when we speak about neg likeli, we name it true energy and write it $E^{(T)}$. Helmoltz and free energy.

The potential energy article has some paragraphs. Also explain what the score is. Boltzmann distrib Energy-based models Alleen Mara!!!

2.4 Attention mechanisms

- Attention how it is used in DL (image captioning, speech recognition)
- Different types of attention (most important ones)
- Link to humans, different models and not exclusive, most interesting one in my opinion Kahnemann (the attention part of our model is inspired from this)

Kahnemann book tutorial on attention soft vs hard attention other tutorial attention

2.5 Physics

Integr criterion, Motivation why, otherwise it will be weird. With a fun example go through model differential equations – condition field – potential - force field - gradient, divergence. integrability criterion $\partial F/\partial x$... really important (see potential energy paper).

Chapter 3

Literature Review

two three pages review current crossmodal attention, multimodal, single mode, weakness, strength, citations. And how they differ from this work. Build up. speak about showing uncertainty in predictions. about attention between modes and its interpretation. biggest difference with mine is their are specific and implicit solutions. Ours is explicit (thus interpretable and does not burden the model) and general (easily inserted to any model without architecural modifications). interpreta deep lip reading comparisons other cocktail-party solution multi input cross attention transformer example cross attention1 example cross attention2 stacked cross attention latent attention Noise-tolerant Audio-visual Online Person Verification using an Attention-based Neural Network Fusion attention is all you need deep audio-visual speech recognition Deep Structured Energy Based Models for Anomaly Detection Learn to Combine Modalities in Multimodal Deep Learning Multimodal Machine Learning: A Survey and Taxonomy attention a review

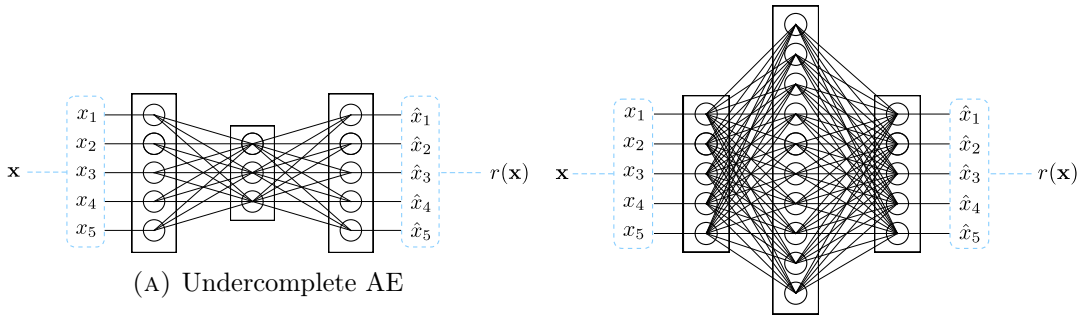
Chapter 4

Energy Estimation

As mentioned earlier, EMMA takes the outlyingness of each mode into account to determine which modes are important and which are not. Negative log-likelihood (NLL) gives us such a measure of outlyingness, which can be approximated with energy-based models. However, in section 2.3 we explained that classical energy-based models are difficult to train. This chapter discusses two alternative estimators obtained simply by training an autoencoder. First, we will explain autoencoders and how the two estimators are derived. Next, we will compare both estimators on simple generated datasets.

4.1 Autoencoder

Autoencoders (AE) are models trained to reproduce the inputs to their outputs. Let $\mathbf{x} \in \mathbb{R}^L$ be the input, autoencoders consist of two parts: the encoder $\mathbf{u} = f(\mathbf{x}) = h(W_f\mathbf{x} + \mathbf{b}_f)$ with $h(\cdot)$ an activation function, and the decoder outputting the reconstruction $g(\mathbf{u}) = W_g\mathbf{x} + \mathbf{b}_g$. The reconstruction can also be written as $r(\mathbf{x}) = f(g(\mathbf{x}))$. Autoencoders are trained in an unsupervised manner, most commonly using the mean-squared error between input and output, $\|r(\mathbf{x}) - \mathbf{x}\|_2^2$. But why is it useful to train a model to copy its input? Well to answer this question we need to distinguish two types of autoencoders: the undercomplete and overcomplete autoencoders.



(A) Undercomplete AE

(B) Overcomplete AE

FIGURE 4.1: Two types of autoencoder architectures.

4.1.1 Undercomplete

In an undercomplete AE, the size of the hidden layer $\mathbf{u}$ is smaller than the size of the input/output layers (see Fig. 4.1a). Having a bottleneck forces the model to loose some information and keep only the most relevant features. The values of the hidden layer can be interpreted as a representation in latent space of the input.

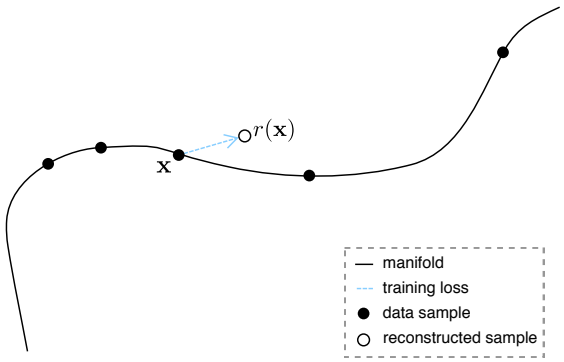


FIGURE 4.2: Vector representation of an undercomplete reconstruction process.

4.1.2 Overcomplete

On the contrary, an overcomplete AE has more hidden units than its input and output layer (see Fig. 4.1b). In this case, a trivial hack for the model to learn is to copy its input through the hidden layer to the output. However, by adding some noise to the input and forcing the AE to reconstruct the original input, the model can then be used to denoise signals. The latter is how a denoising autoencoder works (DAE). The input is corrupted with some small isotropic noise $\tilde{\mathbf{x}} = \mathbf{x} + \epsilon$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and a training loss

$$\|r(\tilde{\mathbf{x}}) - \mathbf{x}\|_2^2 \tag{4.1}$$

On Fig. 4.3 we verifiy that minimizing the loss, $r(\tilde{\mathbf{x}}) \to \mathbf{x}$, is equivalent to learning to invert the corruption, $r(\tilde{\mathbf{x}}) - \tilde{\mathbf{x}} \to -(\tilde{\mathbf{x}} - \mathbf{x})$.

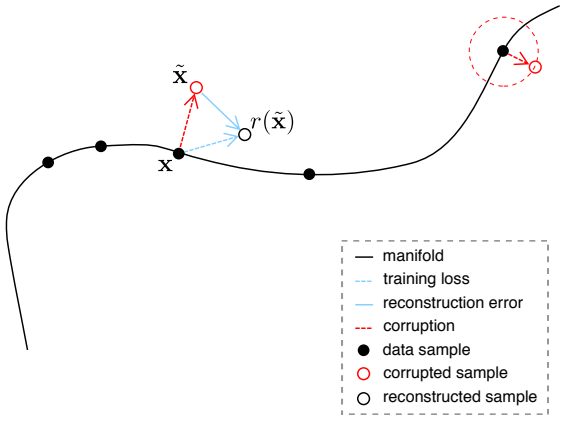


FIGURE 4.3: Vector representation of an overcomplete reconstruction process.

4.2 Estimators

The norm of the reconstruction error, $\|r(\tilde{\mathbf{x}}) - \tilde{\mathbf{x}}\|_2^2$, is commonly used as an approximate measure of outlyingness. The intuition behind it, is that inputs far away of the manifold in the dataspace, will have larger norms.

The authors in (Alain and Bengio, 2012) proved that the reconstruction error of a trained DAE is proportional to the score (derivative of log-likelihood)

$$r(\tilde{\mathbf{x}}) - \tilde{\mathbf{x}} \propto \frac{\partial \log p(\tilde{\mathbf{x}})}{\partial \tilde{\mathbf{x}}} \tag{4.2}$$

To put it in another way, the reconstruction error points in the same direction as the most likely datapoint. To verify this visually, we can train a DAE on a generated circle manifold (see experiment in 4.3). As we can see below, the vector field of the reconstruction error does indeed point towards the data manifold.

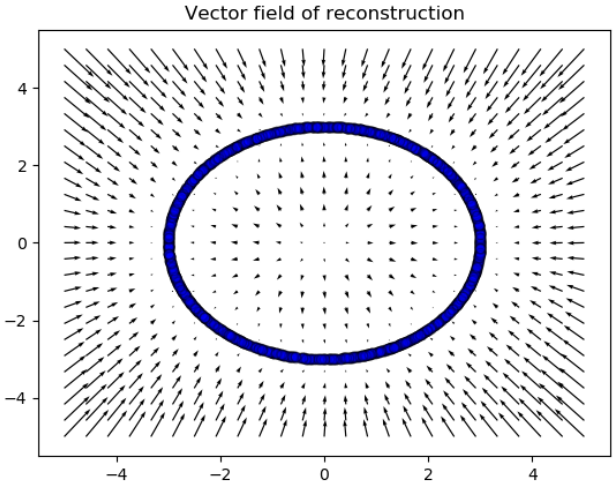


FIGURE 4.4: Vector field of reconstruction error.

In (Kamyshanska and Memisevic, 2014), authors observed that by using tied weights ($W_f = W_g^T$), the integrability criterion[1] became satisfied:

$$\begin{aligned} \frac{\partial (r(\tilde{\mathbf{x}})_i - x_i)}{\partial x_j} &= \sum_k W_{ik} \frac{\partial h(W\tilde{\mathbf{x}} + \mathbf{b}_f)}{\partial (W\tilde{\mathbf{x}} + \mathbf{b}_f)} W_{jk} - \delta_{ij} \\ &= \frac{\partial (r(\tilde{\mathbf{x}})_j - x_j)}{\partial x_i} \end{aligned} \tag{4.3}$$

where δ_{ij} denotes the Kronecker delta. As the integrability criterion is satisfied, the vector field can be expressed as a gradient of a scalar field $-\Psi$, such that $r(\tilde{\mathbf{x}}) - \tilde{\mathbf{x}} = -\partial\Psi(\tilde{\mathbf{x}})/\partial\tilde{\mathbf{x}}$. In analogy to physics, the vector field can be tought as a force applied on the input and the scalar field as a potential energy. Furthermore, the reconstruction process can be seen as a gradient descent in the potential energy landscape. For our purpose, an important observation is to note that the potential energy is proportional to the NLL,

$$-\frac{\partial \Psi(\tilde{\mathbf{x}})}{\partial \tilde{\mathbf{x}}} \propto -\frac{\partial \log p(\tilde{\mathbf{x}})}{\partial \tilde{\mathbf{x}}} \Rightarrow \Psi = -\log p \tag{4.4}$$

The potential energy being the gradient of the reconstruction error, we can compute Ψ as

$$\Psi(\tilde{\mathbf{x}}) = -\int (r(\tilde{\mathbf{x}}) - \tilde{\mathbf{x}}) d\tilde{\mathbf{x}} \tag{4.5}$$

[1] See 2.5

By substituting $f = h(W\mathbf{x} + \mathbf{b}_f)$ and $g = W^T\mathbf{x} + \mathbf{b}_g$ with $W = W_f$, we obtain[2]

$$Psi(\tilde{\mathbf{x}}) = -\int f(\tilde{\mathbf{x}})d\tilde{\mathbf{x}} - \frac{1}{2}\|\tilde{\mathbf{x}} + \mathbf{b}_r\|_2^2 + \text{const} \tag{4.6}$$

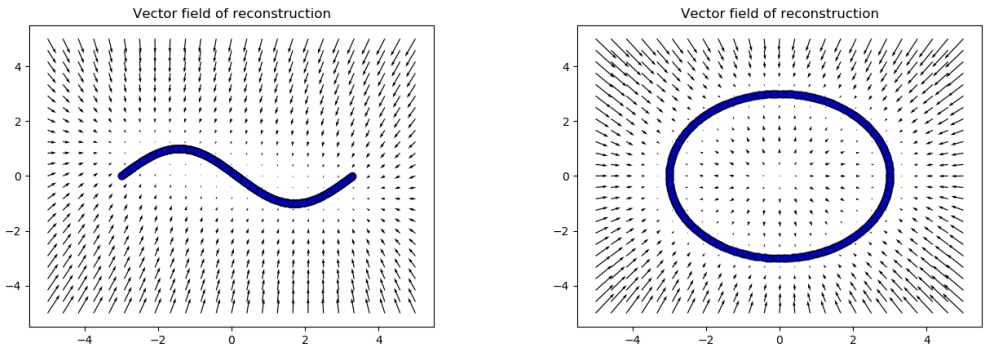
In this work we will only use the sigmoid activation function,

$$\boxed{\Psi(\tilde{\mathbf{x}}) = -\sum_k \log(1 + \exp(W_{.k}^T\tilde{\mathbf{x}} + b_k^f)) + \frac{1}{2}\|\tilde{\mathbf{x}} - \mathbf{b}_g\|_2^2 + \cancel{\text{const}} \propto -\log p(\tilde{\mathbf{x}})} \tag{4.7}$$

where $W_{.k}^T$ is the k^{th} column of W^T, and b_k^f the k^{th} element of $\mathbf{b}_f$.

4.3 Experiment I

As expected the potential energy obtains results being more theoretically grounded. For this experiment two data manifold were generated (wave and circle), for each one an DAE was trained as described previously. Experimental setup in appendix. The results of this experiment are shown in Fig. 4.7 and Fig. **??**.



(A) Wave manifold (B) Circle manifold

FIGURE 4.5: Two test manifolds, generation of 200 samples in a certain range.

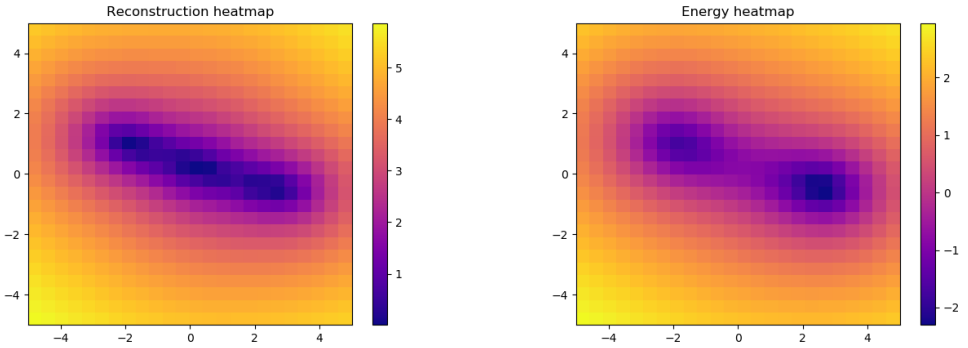
Results

Show results+ observatiions (explain why there are sources, explain why less convergence in center of wave (because more accumulation)) + conclusion we will use potential (better results, more robuts, more theoretically grounded). Table of six figures.

4.4 Limitations

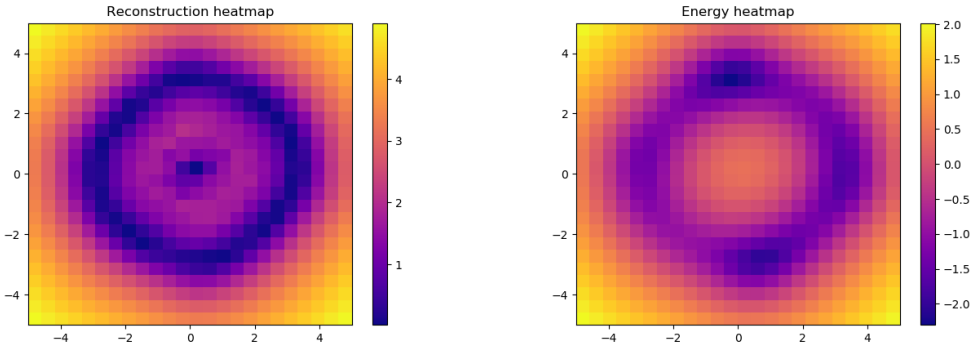
Explain easy but not easily extensible to images/sounds. Mention alternatives. Our goal is not to analyze best energy approximator bt to show that roughly any approx can be used to solve our problem. Hypothesis + relaxation + interstingly can we generalize to convolutions... - research direction alternative

[2] Details about computation are in (Kamyshanska and Memisevic, 2014).



(A) Wave manifold (B) Circle manifold

FIGURE 4.6: Two test manifolds, generation of 200 samples in a certain range.



(A) Wave manifold (B) Circle manifold

FIGURE 4.7: Two test manifolds, generation of 200 samples in a certain range.

Chapter 5

EMMA

In chapter 3, a review of the important works in multi-modal deep learning was done. We concluded that no MMDL architecture were enabled with an explicit module leveraging the multi-modality to improve the robustness. To adress this issue, we designed a novel generic module that can easily be inserted into every trained MMDL architecture. In this chapter we start by presenting the main ideas and justifications of the design (5.2) which we argue are general. Then we give more details about the architectural choices to achieve those ideas. Keep in mind for the latter that the specific choices are not unique, multiple other ways could further be explored.

5.1 Problem Statement

We define the i.i.d. dataset $(\mathbf{X}, y) \in \mathcal{D}^{(N)}$ of size N, with $\mathbf{X}$ the input composed of M modes $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ and y the target variable. The goal of the MLP (see Fig. **??**) is to make predictions $\hat{y}$ as close as possible to the groundtruth y.

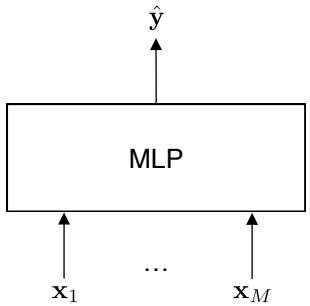


FIGURE 5.1: Mult-modal multi-layer perceptron.

1. define concepts: dynamic/static quantities. outlying. generalized robustness.
2. problem: impact of outlying on predictions. drop accuracy.
3. classical way against noise in unimodal DL - fix noise corruption on some part of training-set – problem is generalized robustness is not satisfied. 1) noise could vary on test environment 2) no solution for missing 3) even worse for unseen
4. MMDL has different representations/modes thus when one is corrupted other maybe is not (hypothesis)
5. human leverage this to improve robustness. Crossmodal attention
6. this is our inspiration. mimick. focus our attention dynamically (change for each sample) on most important ones. what is importance.

7. importance lead to importance lead to attention scores. why not directly importance scores. see figure.

8. two ways of viewing our solution 1) masking perturbation, others are unchanged. The MLP has to learn how to adapt. 2) given by multiplication, both mode and value of importance are passed. The model has to learn an algo taking also second in account.

9. advantages 1) more biologically plausible (splitting processes, reasoning, preprocess) 2) reduces burden on MLP 3) EMMA is specifically designed to express complex optimal attention MLP alone is not 4) just add, no model architecture modification + no retrain from scratch 5) interpretability + easier to identify weakness: robustness (EMMA) or prediction (MLP). in analogy we can improve both separately. due to modularity.

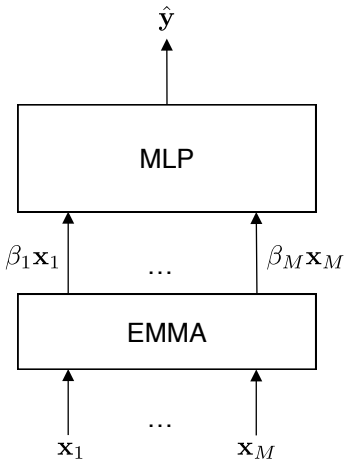


FIGURE 5.2: Summary

How compute/determine β, see next section

Hypotheses

Hypothesis. What and why.

- MLP (no image, no sequence)
- Information gain multiple modes > information gain modes separate
- Different modes, thus independent to noise or at least semi-independent

5.2 General Framework

The importance of a mode is influenced by three intrinsically tied factors:

- *relevance*: how much it helps improve the accuracy of the predictions
- *outlyingness*: the more outlying the input, the less information it carries
- *coupling*: a mode i and a mode j are said to be strongly coupled if they influence each other. They can be coupled in two ways: negatively, if one mode is outlying, the other takes over and becomes more important. Or positively if one mode is outlying, the other becomes also less important

To capture those three factors, we introduce E_i, the modal energy of mode i. Modal energies are learnable parametric functions of potential energies of the following structure

$$E_i = f(\Psi_i) + \sum_{k \neq i}^{M} g[f(\Psi_i), f(\Psi_k)] \tag{5.1}$$

The potential energy Ψ_i being a measure of outlyingness of mode i and f beign learnable via the loss function, f captures the relationship between relevance and outlyingness. The function g is constructed to learn the optimal coupling between modes. Using the Gibbs distribution to normalize modal energies, we obtain the importance scores α_i. Interpretability.

In most works of psychology, attention is seen as a selection process between inputs. Attention in Deep Learning is based on this view. Daniel Kahnemann, a famous psychologist models attention as a shared resource with a limited capacity (brain) being allocated between the inputs (tasks). This model of attention can be achieved by adding a small constraint to a classical attention mechanism. This is how we go from importance scores to attention scores.

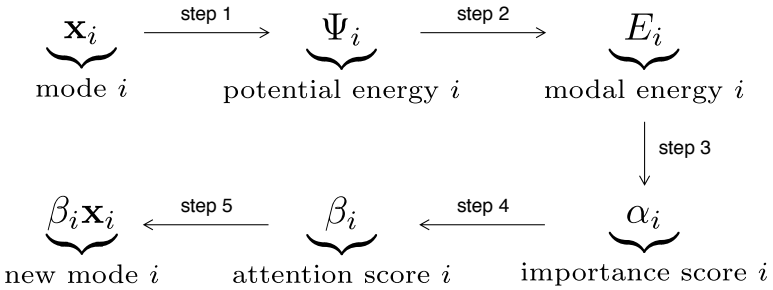


FIGURE 5.3: Summary of main steps in EMMA. Step 1 was explained in chapter 4. Step 2,3 and 4 are detailed in the rest of this chapter. Step 5 is trivial.

5.3 From Potential to Modal energies (step 2)

A problem we overlooked so far is that we neglected the integration constant when computing the potential energy. Leading to possible negative values of energy. Moreover, during the backpropagation the gradient computation involves taking the logarithm of Ψ_i[1], which is undefined for negative values. To solve this issue, we subtract the potential function by the lowest value of potential energy $\Psi_i^{(\min)}$ in the training set minus Euler's number:

$$\Psi_i' = \max(\mathrm{e}, \Psi_i - \Psi_i^{(\min)} + \mathrm{e}) \tag{5.2}$$

This sets the bias of the potential energy such that all values are positive and avoid exploding gradients. The reason we use a max-operator is because lower energy values than $\Psi_i^{(\min)}$ can occur during inference.

Another problem we did not mention is that potential energies are proportional to the NLL (see Eq. 4.7). Thus we could have potential functions of different modes on very different scales. We define *self-energy* of a mode i as

$$e_i = w_i \Psi_i' + b_i \ \ (= f(\Psi_i)), \qquad w_i, b_i \in \mathbb{R}^+ \tag{5.3}$$

[1] See appendix

The model is now able to put them on the same scale and at the same time the parameters are influenced by the loss and thus captures the relevance. Furthermore, the *shared energy* between mode i and j is constructed from the self-energies i and j as follows

$$e_{ij} = w_{ij} e_i^{\gamma_{ij}} e_j^{1-\gamma_{ij}} \quad (= g[f(\Psi_i), f(\Psi_k)]), \qquad w_{ij} \in [-\frac{1}{M-1}, \frac{1}{M-1}], \gamma_{ij} \in [0,1] \tag{5.4}$$

By adding the constraint $\gamma_{ij} = \gamma_{ji}$ the model can learn the coupling as described in Sec. 5.2. Indeed, if $\gamma_{ij} \to 0$ the two modes are strongly coupled, whereas if $\gamma_{ij} \to 1$ the two modes become independent. Notice that the modes can be positively coupled ($w_{ij} > 0$) and negatively coupled ($w_{ij} < 0$). We do **not** impose w_{ij} to be equal to w_{ji}, such that modes can influence each other asymmetrically (if one mode is missing, other can take over but not optimal the other way around).

Finally, modal energies are the summation of their self-energy and shared energies with all other modes:

$$E_i = e_i + \sum_{k \neq i}^{M} e_{ij} \tag{5.5}$$

We can recognize equation 5.1. Interestingly, an intuitive quantity for how uncertain the model is, is the total energy $E_{\text{total}} = \sum_i E_i$. Interpretability. We will verify this in experiments.

5.4 From Modal energies to Importance scores (step 3)

The importance scores are computed from the modal energies via the Gibbs distribution:

$$\alpha_i = \frac{1}{Z} e^{-\rho E_i} \quad \text{with the partition function} \quad Z = \sum_{k=1}^{M} e^{-\rho E_k} \tag{5.6}$$

This guarantees that the scores sum to one and that α_i are high for low energies, as desired. The hyperparameter ρ represents the coldness and is the inverse of the temperature, it controls the entropy of the importance scores distribution. At high temperature ($\rho \to 0$) the distribution becomes more uniform, and at low temperature ($\rho \to +\infty$) the importance scores corresponding to the lowest energy tends to 1, while the others approach 0 (see Fig. 5.4).

As we can see in Fig. 5.4, the hyperparamets ρ plays a key role on how the model will behave. It can either force the model to be more favorising (high ρ) or more indifferent (low ρ).

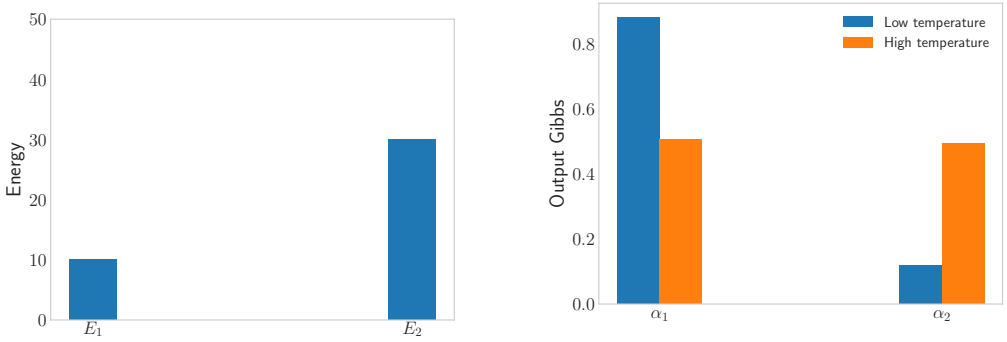
5.5 From Importance to Attention scores (step 4)

The attention scores are computed from importance scores as

$$\beta_i = \max[0, \tanh(g_a \alpha_i - b_a)] \quad \text{with} \quad g_a > 0,\ b_a \in [0,1] \tag{5.7}$$

The design choices are listed and justified below:

- hyperbolic tangent: introduces non-linearity
- max-operator (ReLU): by definition $\beta_i \in [0,1]$, not necessary but is a more general attention for $\alpha \in \mathcal{R}$
- the gain g_a controls the capacity



(A) Energies

(B) Corresponding importance scores for different temperatures.

FIGURE 5.4: Input-output of Gibbs distribution for two temperatures: low temperature ($\rho = 0.1$) and high temperature ($\rho = 0.001$)

- the bias b_a controls how much the minimum importance is necessary to pass threshold

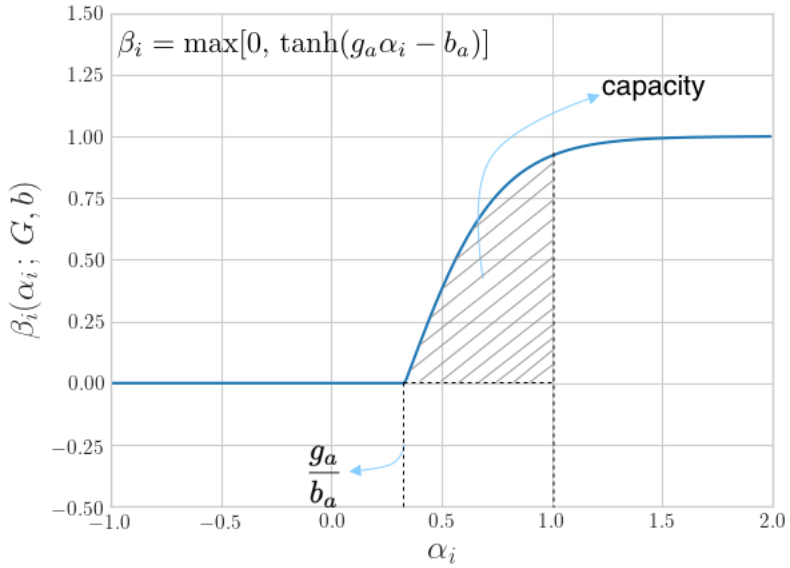


FIGURE 5.5: Attention function

Energy threshold

The maximal energy a mode i can contain in order to pass

$$\begin{aligned} &\Leftrightarrow g_a \alpha_i - b_a > 0 \\ &\Leftrightarrow \log(\alpha_i) > \log(b_a / g_a) \\ &\Leftrightarrow E_i \geq \frac{\log(g_a/b_a) - \log(Z)}{\rho} = E_{\text{threshold}} \end{aligned} \tag{5.8}$$

As we can see g_a/b_a controls threshold static because fixed. But the partition function is dynamic, if total energy is higher the minimal energy augments and thus less strict, which is good. The influence of the temperature is non-trivial because Z also depends on ρ.

Capacity

We can define the capacity, the shared resource, as

$$\text{capacity} \triangleq \int_0^1 \tanh(g_a\alpha + b_a)d\alpha \tag{5.9}$$

Define the auxiliary variable $u = g_a\alpha + b_a$. Now using

$$\frac{du}{d\alpha} = g_a \Leftrightarrow d\alpha = \frac{1}{g_a}du \tag{5.10}$$

we can write

$$\begin{aligned}\text{capacity} &= \frac{1}{g_a}\int_0^1 \tanh(u)du \\ &= \frac{1}{g_a}\log[\cosh(g_a\alpha + b_a)]\Big|_{\alpha=0}^{1} + \cancel{\text{constant}} \\ &= \frac{1}{g_a}\log\left[\frac{\cosh(g_a + b_a)}{\cosh(b_a)}\right]\end{aligned} \tag{5.11}$$

For a low capacity, only most important modes will pass. On the contrary, for a high capacity, all modes pass. If too high, will let noise pass. Too high capacity, not robust enough (will let noise pass). Too low, will lose to much information, accuracy drops. System learns optimal trade-off. Altough question is: does robustness generalize, do we need to regulate capacity. See experiments.

A classical attention mechanism is $\tanh(\mathbf{W}\alpha + \mathbf{b})$, but by forcing to be common... Notice that when we compare with our function, it does not have one capacity but multiple capacities, one for each mode. Interpretability loss and importance scores less meaningfull combined with capacity.

5.6 Training & Regularization

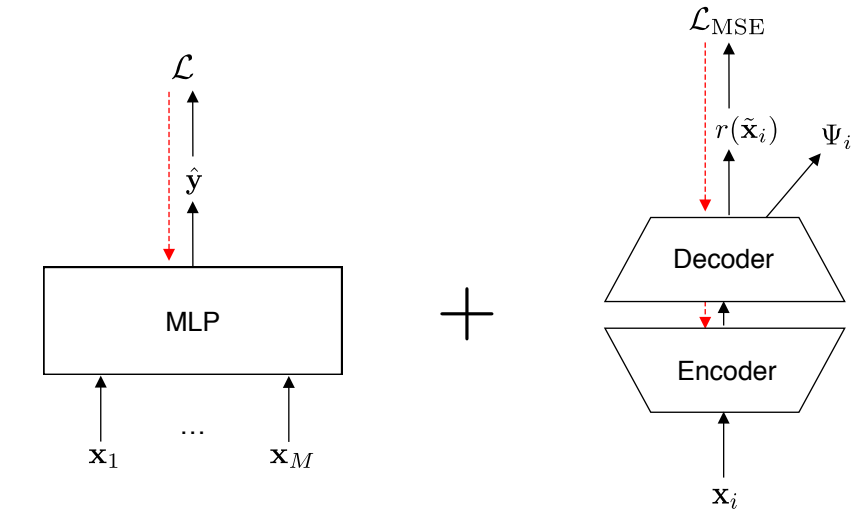
The training of the model happens in two phases (see Fig. 5.6). First, the MLP part and the autoencoders (one foreach mode) are trained in parallel. In the second phase, we insert EMMA in front of the MLP and train it end-to-end with normal and corrupted data. For example, when training with two modes: 50% normal data, 25% noisy data on mode 1, 25% noisy data on mode 2. Freezed autoencoders.

Moreover we add one regularizer for the capacity and one for the energy:

$$\tilde{\mathcal{L}} = \mathcal{L}(y, \hat{y}) + \lambda_c g_a - \lambda_e \Omega \quad \text{with} \quad \Omega = \sum_{k=1}^{M} \xi_k \log(\alpha_k) \quad \text{and} \quad \xi_k = \begin{cases} -1 & \text{if } \mathbf{x}_k \text{ is corrupted} \\ +1 & \text{otherwise} \end{cases} \tag{5.12}$$

where λ_c and $lambda_e$ respectively tune the weight of regularizing capacity and energy. In the experiments (chapter 8), we will see how λ_c and $lambda_e$ affect the performance and interpretability of the system. Intuitiveliy, trying to minimize the capacity while keeping good predictions may be a good idea to have a good generalized robustness.

Capacity can be seen as L_1 weight decay (cite, because $g_a > 0$). On the other hand, regularizing energy is done for interpretability purposes, if modal energies are only influenced by loss on predicitions, we could have a large discrepancy between modal energy E_i and the original potential energy Ψ_i. When the mode i is noisy, the corresponding potential energy is high, and with the regularizer we force α_i to be low and thus E_i to be also high.



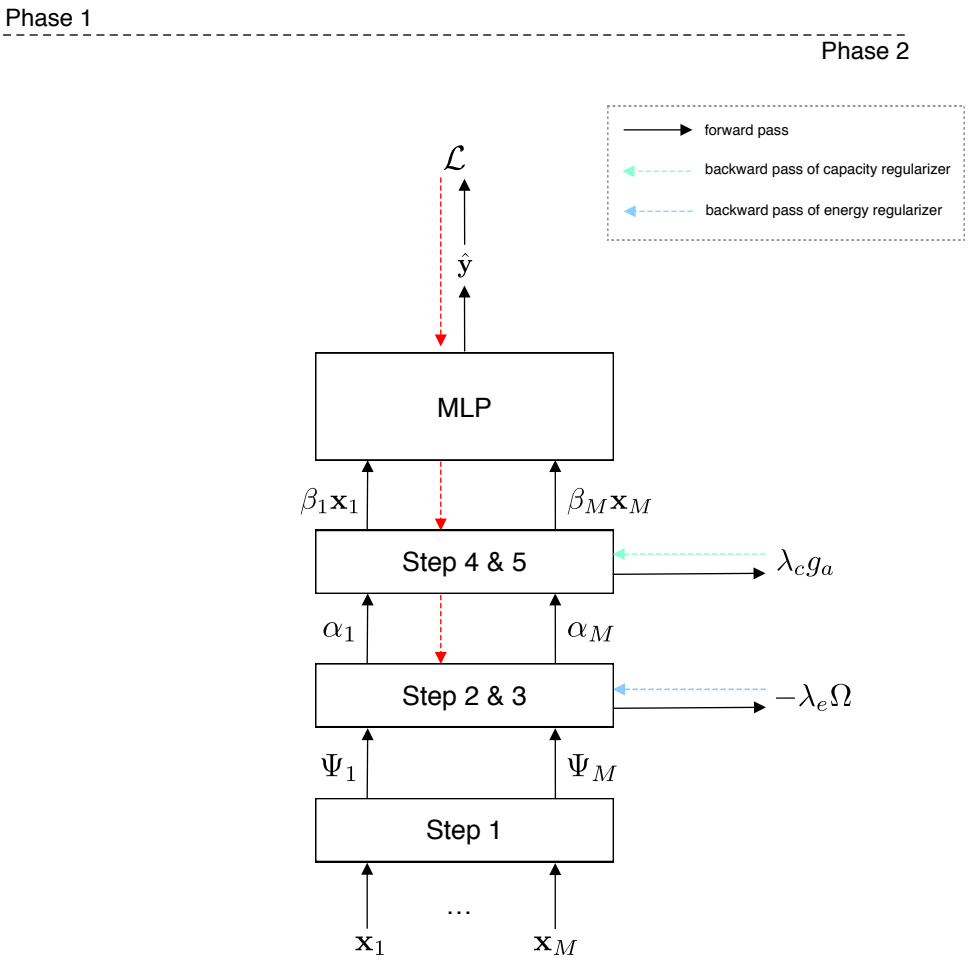


FIGURE 5.6: Summary of how EMMA is trained.

Backpropagation of regularizer

param step update $\theta - \epsilon\lambda\nabla_\theta\Omega$. 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,

$$\nabla_\theta\Omega = \sum_{k=1}^{M} \eta(\mathbf{m}_k)\nabla_\theta \log(\alpha_k) \tag{5.13}$$

develop $\nabla\Omega$, stop before M' introduction. The gradient of the logarithm becomes

$$\begin{aligned}
\nabla_\theta \log(\alpha_k) &= \nabla_\theta \log\left[\frac{e^{-\tau E_k}}{Z}\right] \\
&= \nabla_\theta(-\tau E_k) - \nabla_\theta \log \sum_{l=1}^{M} e^{-\tau E_l} \\
&= -\tau\nabla_\theta E_k - \frac{\sum_{l=1}^{M} \nabla_\theta e^{-\tau E_l}}{\sum_{l=1}^{M} e^{-\tau E_l}} \\
&= -\tau\nabla_\theta E_k + \tau\frac{\sum_{l=1}^{M} \nabla_\theta(E_l) e^{-\tau E_l}}{\sum_{l=1}^{M} e^{-\tau E_l}} \\
&= \tau\left[-\big(1 - \frac{e^{-\tau E_k}}{Z}\big)\nabla_\theta E_k + \sum_{l\neq k}^{M} \frac{e^{-\tau E_l}}{Z}\nabla_\theta E_l \right] \\
&= \tau\left[-\big(1 - \alpha_k\big)\nabla_\theta E_k + \sum_{l\neq k}^{M} \alpha_l\nabla_\theta E_l \right]
\end{aligned} \tag{5.14}$$

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 5.14 can be split as follows

$$\nabla_{\theta_i} \log(\alpha_k) = \begin{cases} -\tau(1-\alpha_i)\nabla_{\theta_i} E_i, & \text{if } i = k \\ \tau\alpha_i\nabla_{\theta_i} E_i, & \text{if } i \neq k \end{cases} \tag{5.15}$$

explain what happens for bad choice of M' and for good choice M', and why we choosed +1 and -1 to make math work. 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 (5.15) in (5.13) and extrapolating the sums we obtain the expression of $\nabla_{\theta_i}\Omega$ if mode i is normal

$$\eta_+\big[-\tau(1-\alpha_i)\nabla_{\theta_i} E_i\big] + \big[(M'-1)\eta_+ + (M-M')\eta_-\big]\alpha_i\tau\nabla_{\theta_i} E_i \tag{5.16}$$

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

$$\eta_-\big[-\tau(1-\alpha_i)\nabla_{\theta_i} E_i\big] + \big[M'\eta_+ + (M-M'-1)\eta_-\big]\alpha_i\tau\nabla_{\theta_i} E_i \tag{5.17}$$

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

$$\boxed{\nabla_{\theta_i}\Omega = -\big[(M-2M')\alpha_i + \text{sign}(\eta_i)\big]\tau\nabla_{\theta_i} E_i} \tag{5.18}$$

Contrastive divergence

Two cases can be distinguished: If the total number of modes M, is even then

$$\nabla_{\theta_i}\Omega = -\text{sign}(\eta_i)\tau\nabla_{\theta_i} E_i \tag{5.19}$$

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 (5.19) that gradient descent will update values of the energy function E_i downward/upward for normal/abnormal inputs.

If M is uneven,

$$\nabla_{\theta_i}\Omega = \begin{cases} -(1-\alpha_i)\tau\nabla_{\theta_i}E_i & \text{if input } i \text{ is normal} \\ (1+\alpha_i)\tau\nabla_{\theta_i}E_i & \text{otherwise} \end{cases} \tag{5.20}$$

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.

5.7 Research questions

- Are the robustness and interpretability increased using EMMA and MMDL?
- Does limiting the capacity help to generalize robustness better?
- How is the interpretability affected when regularizing energy? If it helps, is it at the cost of losing accuracy?

Chapter 6

Dataset

One whole chapter because quiet interesting however feel free to skip this part, just know it has two modes of each four features. thesis dataset link dataset

Chapter 7

Experiments & Results

Both are experiments on dataset described in previous chapter. Experiment II will be about energy estimation. Experiment III evaluates and analyzes the robustness of the model with and without EMMA.

7.1 Corruption

- standardize, why? split sets and apply one from train error standardize
- SNR (see good explanation in pulsar thesis). If greater than 1, signal non-distinguishable. White noise. Explain it is not the same than AE corruption. $10\log(\frac{1}{\sigma^2})$
- on signal and background because we corrupt the whole mode and not the class

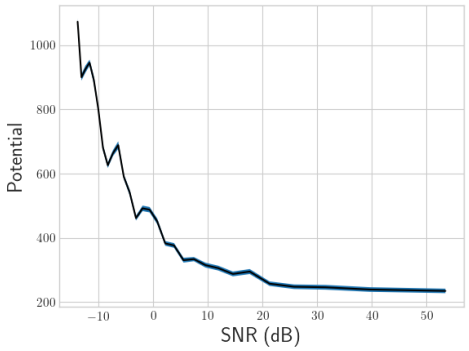
7.2 Experiment II

- train-test split
- AE trained on train-set and then test on test-set
- matrix with number of signals, ... (eda)

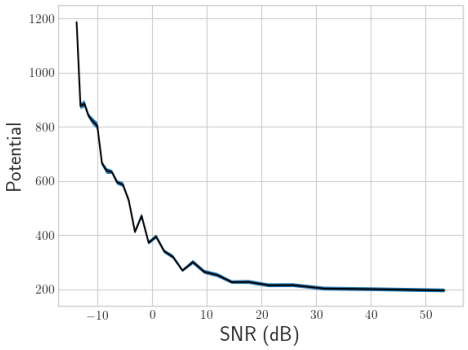
Setup: max epochs = 30, batch size = 64, noise DAE = 0.01, d input = 4, n hidden = 12, adam 0.001, sigmoid

7.3 Experiment III

- BCE & F1 (not AUC – explain why). F1vsAUC1. F1vsAUC2. F1vsAUC3. F1
- train-valid-test
- threshold optimal choice via ROC. all on valid set
- 3 models: base (train normal, valid normal), without (train noisy, valid noisy), with (train noisy, valid noisy)
- 50-25-25 noisy mode. give detailed numbers. eda.
- trained with early stopping + retrain for .. epochs with valid+train. saved model.
- one subsection per plot: explain details experiment and how results are obtained. then analyze and conclusions.

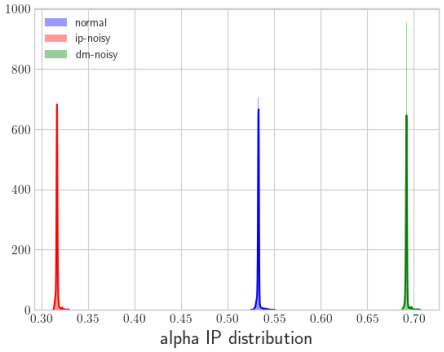


(A) Autoencoder trained on integrated-profile mode

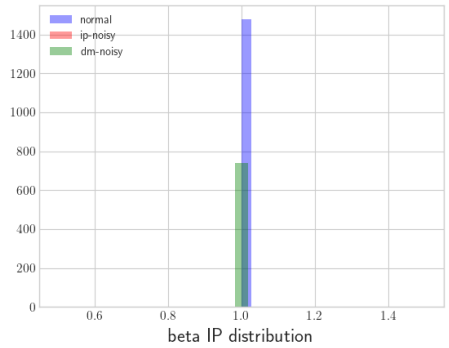


(B) Autoencoder trained on dm-snr mode

FIGURE 7.1: Potential energy vs signal-to-noise ratio on test set. Results as expected, the more corrupted, the more energy. As we can see, variance of energy is very small. Exponential grow is due to logarithmic scale of SNR



(A) ...



(B) Autoencoder trained on dm-snr mode

FIGURE 7.2: Potential energy vs signal-to-noise ratio on test set. Results as expected, the more corrupted, the more energy. As we can see, variance of energy is very small. Exponential grow is due to logarithmic scale of SNR

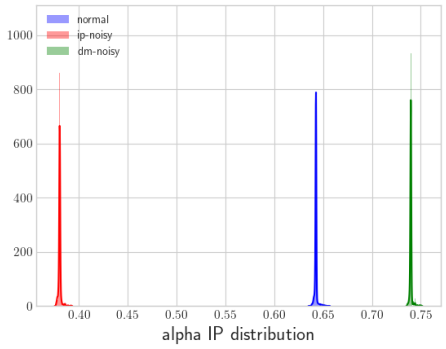
Attention-shift

Robustness generalisation

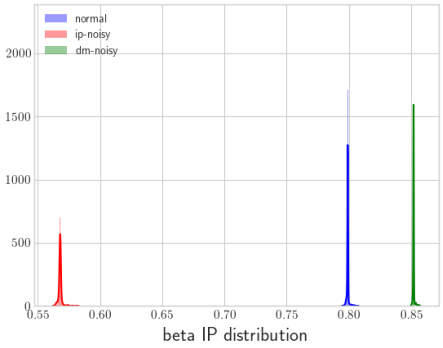
Yerkes-Dodson curve

over-under arousal. do on larger range.

Energy generalisation

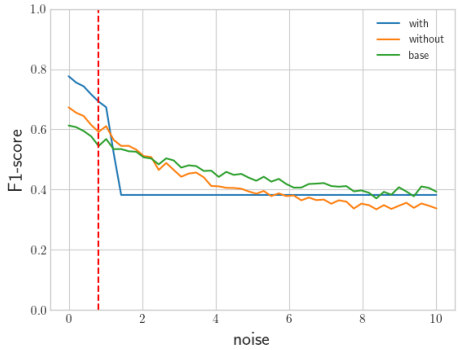


(A) Autoencoder trained on integrated-profile mode

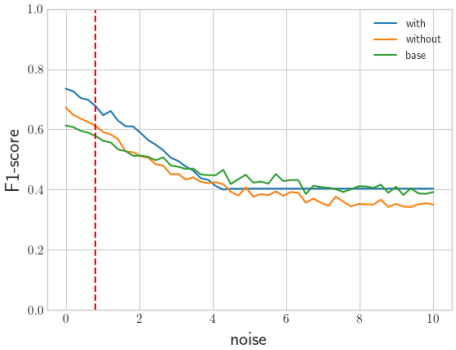


(B) Autoencoder trained on dm-snr mode

FIGURE 7.3: Potential energy vs signal-to-noise ratio on test set. Results as expected, the more corrupted, the more energy. As we can see, variance of energy is very small. Exponential grow is due to logarithmic scale of SNR



(A) ...



(B) Autoencoder trained on dm-snr mode

FIGURE 7.4: Potential energy vs signal-to-noise ratio on test set. Results as expected, the more corrupted, the more energy. As we can see, variance of energy is very small. Exponential grow is due to logarithmic scale of SNR

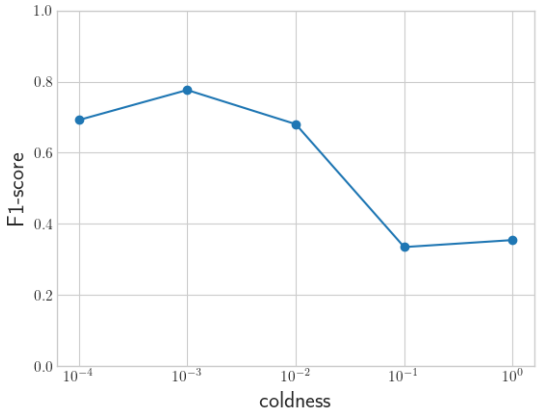


FIGURE 7.5: blabl

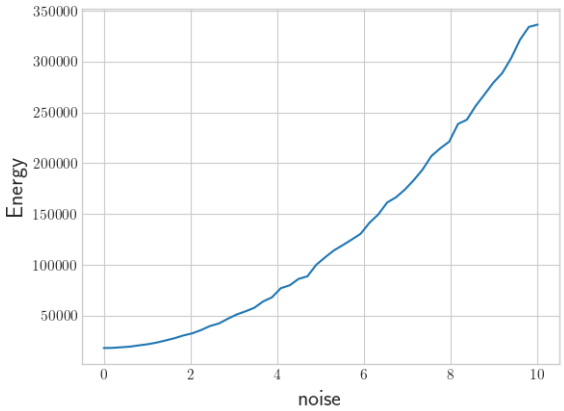


FIGURE 7.6: blabl

Chapter 8

Conclusion

Summary of what was seen/done during the master thesis from start to end.

8.1 Contributions

Summary of contributions

- EMMA adds robustness, mimicks crossmodal attention
- at best of out knowledge capacity & attention (Kahneman) never described in literature - first time link is explicitly done, and capacity DL introduced
- new regularizer

8.2 Research questions

Each research question + answer

8.3 Future work

- 2-level attention + diagram
- Annealing + init, end temperature + explain init of next layer Linke annealing
- different shared energies design
- Images/sound sequences
- multiple modes then blows up. Image 100 modes (altough not realistic in real-world problems) then tend to zero. Solution add a common gain? Analyze influence of multiple modes etc..
- it could be very easy to test on images/sounds (give even during test-time a true outlyingness measure, not specifically the NLL) and at the same time investigate general ways of getting such a measure

Appendix A

Code

A.1 Structure

A.2 Pytorch

A.3 Experiments

Appendix B

Experimental Setup

Appendix C

Miscellaneous

Bibliography

Alain, Guillaume and Yoshua Bengio (2012). "What Regularized Auto-Encoders Learn from the Data Generating Distribution". In: *arXiv e-prints*, arXiv:1211.4246, arXiv:1211.4246. arXiv: 1211.4246 [cs.LG].

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

Santilli, RM (1982). "Birkhofian generalization of Hamiltionian Mechanics". In: *Foundations of theoretical mechanics II*.