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
University of Liège
Liège, Belgium

Academic Year 2018 - 2019

"*Sometimes it seems as though each new step towards Artificial Intelligence, rather than producing something which everyone agrees is real intelligence, merely reveals what real intelligence is not.*"

Douglas Hofstadter

Abstract

A multi-modal neural network exploits information from different channels and in different terms (e.g., images, text, sounds, sensor measures) in the hope that the information carried by each mode is complementary, in order to improve the predictions the neural network. Nevertheless, in realistic situations, varying levels of perturbations can occur on the data of the modes, which may decrease the quality of the inference process. An additional difficulty is that these perturbations vary between the modes and on a per-sample basis. This work presents a solution to this problem. The three main contributions are described below.

First, a novel attention module is designed, analysed and implemented. This attention module is constructed to help multi-modal networks handle modes with perturbations.

Secondly, two new regularizers are developed to generalize the robustness to more intensive failing modes (relative to the training set).

Lastly, a unified multi-modal attention module is presented, combining the main types of attention mechanisms in the deep learning literature with our module. We suggest that the unified module could be coupled with a prediction model to enable the latter face unexpected situations, and extract the most relevant information in the data.

Acknowledgements

I would like to thank everybody who kept me busy this year. In particular, my thesis advisor Dr. Raphaël Marée for always encouraging my research, and Romain Mormont who gave me valuable feedback on the writing of this Master thesis. I would also like to thank the jury for reading the text.

Moreover, I would like to acknowledge the work of all the Professors at the University of Liège who helped me become an engineer.

I am also very grateful to my good friends Mathias Berger and Lucas Fuentes. They took time of their busy schedules to provide me with very valuable comments and suggestions, which have drastically improved the quality of this thesis. Thank you.

Finally, I must express my profound gratitude to my family, especially my grandparents for providing me with unfailing support and continuous encouragement throughout the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Aurélien Werenne
Liège, Belgium 2018-2019

Contents

List of Figures

Notation and Acronyms

$\triangleq$	Is defined as
N	Number of samples
M	Number of modes
k_B	Boltzmann constant
$M_\odot$	Solar mass
e	Euler's number, base of the natural logarithm (2.71828)
$\mathcal{L}$	Loss function
$\boldsymbol{\theta}$	Set of parameters of the specified model
$\nabla_{\boldsymbol{\theta}}$	Gradient with respect to $\boldsymbol{\theta}$
λ_c	Weight of capacity penalty
λ_e	Weight of energy penalty
Ω	Energy regularizer
Ψ_i	Potential energy of mode i
E_{total}	Total energy
E_i	Modal energy of mode i
e_i	Self-energy of mode i
e_{ij}	Shared energy of mode j on mode i
α_i	Importance score of mode i
β_i	Attention score of mode i
ρ	Coldness in Boltzmann distribution
T	Temperature in Boltzmann distribution

AE	**A**uto**e**conder
BP	**B**ack-**p**ropagation
CNN	**C**onvolutional **N**eural **N**etwork
DAE	**D**enoising **A**uto**e**conder
DL	**D**eep **L**earning
DM	**D**ispersion **M**easure
EMMA	**E**nergy-based **M**ulti-**M**odal **A**ttention
ISM	**I**nter**s**tellar **M**edium
IP	**I**ntegrated **P**rofile
LSTM	**L**ong **S**hort **T**erm **M**emory
MMDL	**M**ulti **M**odal **D**eep **L**earning
MMN	**M**ulti **M**odal **N**etwork
NLL	**N**egative **L**og-**L**ikelihood
RNN	**R**ecurrent **N**eural **N**etwork
SGD	**S**tochastic **G**radient **D**escent
SNR	**S**ignal-to-**n**oise **R**atio
WER	**W**ord **E**rror **R**ate

Chapter 1

Introduction

1.1 Motivation

In recent years, tremendous progress has been made in the field of Artificial Intelligence (AI), especially in Deep Learning (LeCun, Bengio, and Hinton, 2015; Fan, Ma, and Zhong, 2019). Deep Learning has helped AI systems reach and sometimes surpass human-level perception, mainly in computer vision (He et al., 2016) and natural language processing (Wu et al., 2016). This has given rise to amazing industrial applications such as autonomous driving, early cancer detection, enhanced machine translation, etc. In safety-critical contexts, a key concern of engineers is to make sure the trained models are error-free, which can be challenging if the input data does not hold sufficient information to reduce the uncertainty on the predictions to an admissible level.

One possible solution researchers have been exploring is to use multiple modalities[1], which has largely been inspired by the often multi-modal nature of information gathering processes in humans, i.e., we see objects, hear sound, feel the texture, smell odours, and taste flavours. Multi-modal deep learning (MMDL) essentially consists in exploiting information from different channels and in different forms in the hope that the information carried by each mode is (partially) complementary, in order to improve the predictive performance of deep learning models. For example, in (Caltagirone et al., 2018) sensorial inputs from wide-angle cameras and LIDAR[2] sensors are combined for road detection. Cameras provide dense information over a long range under good illumination conditions and fair weather, whereas LIDARs are only marginally affected by the external lighting conditions but have a limited range. Thus, merging the complementary information of the two sensors improves the accuracy of the road detection process. Despite its improvements on the predictions, MMDL still suffers from a major drawback. Indeed, no systematic mechanisms exist to handle failing modes. In the present report, a mode is said to be failing if a) it has a high noise to signal ratio, b) the data is much different from the training data, c) the data is missing. Failing modes a) and b) generally degrade the quality of the predictions because they introduce perturbations in the neural network.

While a solution has not been found for neural networks, humans seem to handle these situations robustly on a daily basis. Analysing human strategies and translating them as much as possible into the framework of AI can provide interesting heuristics to solve this crucial issue. A famous example showing this human ability is called the cocktail-party effect (Cocktail party effect, 2010). It refers to the difficulty we sometimes

[1] The term modality, also called mode, is generally understood to mean "the way in which something happened or is experienced" (Baltrušaitis, Ahuja, and Morency, 2019)

[2] Laser Detection and Ranging

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. Similarly, our attention is shifted from vision to touch when we are in a room where the lights suddenly go out. These examples indicate that humans handle modes with perturbations (first example) or missing information (second example) by shifting their attention on to the other more relevant modes (Driver and Spence, 1998).

Inspired by this behaviour, this report presents a new approach to tackle failing modes. More precisely, a novel attention mechanism[3], dubbed *Energy-based Multi-Modal Attention* (EMMA), is introduced. This mechanism determines how much attention to devote to each mode, so that the relevant information is kept while masking out the perturbations. Additionally, this work offers some insight into how other attention mechanisms in the deep learning literature resemble to the ones observed in humans.

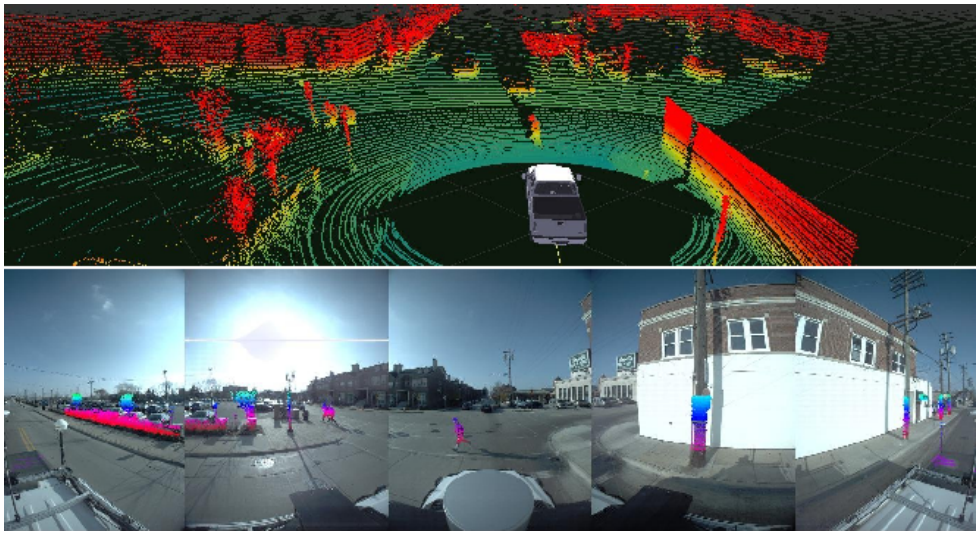


FIGURE 1.1: Same environment, different modes (top: LIDAR view, bottom: camera view).

1.2 Proposed solution

In essence, the EMMA module is placed in front of and coupled with the deep learning model. The module multiplies each mode of the input sample by a weight so that the modes with the most useful features are amplified, while the modes that are unnecessary or contain too much perturbations are masked out. The amount of attention allocated to each mode is determined based on its importance, which is defined in terms of three intrinsically related properties, namely

- *relevance*: the intrinsic informativeness of the mode for the predictive task at hand.
- *failure intensity*: the propensity of a mode to trigger undesirable activations in the neural network.[4]

[3] "Attention mechanisms in deep learning aim to highlight specific regions of the input space" (Chaudhari et al., 2019)

[4] We suggest that a mode of the input sample that is significantly different from the training distribution may cause undesired activations in the neural network, thereby negatively affecting the predictions.

- *coupling*: the interdependencies between the modes, which describe the extent to which the mode provide independent, complementary, redundant or conflicting information.

The module is designed in such a way that it is able to learn these three properties (and their interactions) for each mode. In other words, the module determines the attention to allocate to each mode on its general predictive power, the amount of perturbations it contains and the relationship it has with other modes. For example, this allows the module to mask out a specific failing mode provided that another mode can compensate for its failures. Let us stress that the determination of importance is done on a sample-per-sample basis.

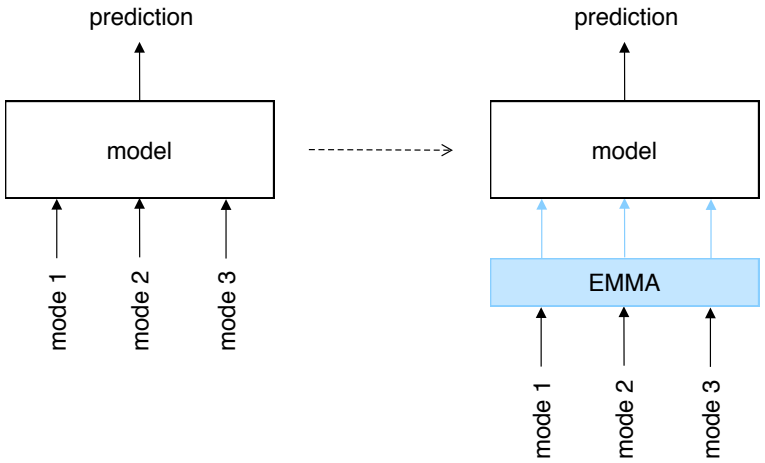


FIGURE 1.2: A multi-modal model with three input modes, without EMMA (left), augmented with EMMA (right).

Software Implementation

All the implemented models and experiments are available at this[5] repository, with a wiki explaining how to run the experiments; PyTorch[6] (Paszke et al., 2017) was the main framework used for the Machine Learning part.

1.3 Contributions

The contributions of this Master thesis can be summarised as follows

Contribution 1: an attention module improving the robustness against failing modes. In Chapter 5, the design of a new attention mechanism based on energy models (LeCun et al., 2006) is discussed, that can be added to any multi-modal model.

Contribution 2: a simple yet powerful regularizer applying to attention mechanisms. A common attention function is modified, establishing a link to the concept of capacity of psychology (Kahneman, 1975), which pertains to the amount of attention allocated among inputs. Subsequently, a new regularizer is

[5] https://github.com/Werenne/energy-based-multimodal-attention
[6] https://pytorch.org/

introduced to control the capacity, whose purpose is to help generalise against unexpected situation.

Contribution 3: a unified model for multi-modal attention. In Chapter 3, a review of the literature on attention in humans helps us identify how to construct a more complete multi-modal attention module.

1.4 Thesis Outline

The remainder of this work is organised as follows.

Chapter 2 explains the background (i.e. deep learning and energy models) this work is based upon.

Chapter 3 reviews the literature on attention in psychology and deep learning, and the similarities between them.

Chapter 4 describes a method for the estimation of the failure intensity of a mode.

Chapter 5 presents the ideas behind the architecture of the Energy-based Multi-Modal Attention module (Contribution 1 & 2).

Chapter 6 presents an evaluation and analysis of the module outlined in Chapter 5 (Contribution 1 & 2).

Chapter 7 proposes a unified multi-modal attention module (Contribution 3).

Chapter 8 concludes this work and suggests possible directions for future research.

Chapter 2

Background

2.1 Machine Learning

Machine Learning is a subfield of Artificial Intelligence (see Figure 2.1) concerned with the design of algorithms that allow machines (e.g. computers, robots, embedded systems) to learn. For a task **T**, a performance measure **P** and an amount of data **D**, the system is said to be learning if it improves its performance **P** at the task **T** by increasing **D** (gain experience). Moreover, there are three main types of learning paradigms, namely supervised, unsupervised and reinforcement learning. In supervised learning (Loog, 2017), the model learns on a labeled dataset, providing an answer that the algorithm can use to evaluate its accuracy on training data. An unsupervised model (Ghahramani, 2004), on the contrary, extracts features and patterns from unlabelled data. Lastly, reinforcement learning (Sutton and Barto, 1998) is typically used to train agents in dynamic environments, where the agent is able to act upon the environment. Reinforcement learning is best explained by an analogy. The learning algorithm is like a dog trainer, which teaches the dog (agent) how to respond to specific signs, like a whistle for example. Whenever the dog responds correctly, the trainer gives a reward to the dog, reinforcing the correct behaviour of the dog. Based on these three paradigms, several families of algorithms have been invented. Deep learning (Fan, Ma, and Zhong, 2019) is one of those families and is particularly powerful on perception tasks.

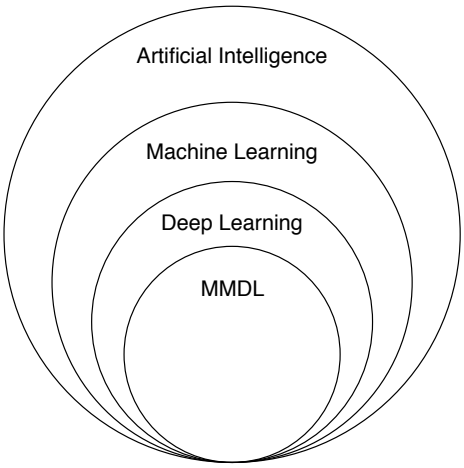


FIGURE 2.1: Venn diagram of the Artificial Intelligence field.

2.2 Deep Learning

Deep learning models, also called Deep Neural Networks, offer the significant advantage of being able to learn their own feature representation for the completion of a given task. A neural network is loosely inspired from our own brains, but can best be seen as a series of stacked non-linear parametric functions, enabling the network to learn multiple levels of representation with increasing abstraction. The parameters are tuned by optimizing a loss function with Stochastic Gradient Descent (SGD) or one of its many enhancements (Ruder, 2016). Let $\boldsymbol{\theta}$ be the set of parameters, $\mathcal{L}$ the loss function, y the groundtruth (labels) and $\hat{y}$ the predictions. First, the SGD algorithm estimates the gradient of the cost function on a randomly sampled batch of size N as

$$\mathbf{g} = \frac{1}{N} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{N} \mathcal{L}(\hat{y}^{(i)}, y^{(i)}) \tag{2.1}$$

where the computation of the gradient itself is done using back-propagation (BP) (Chauvin and Rumelhart, 1995). The SGD algorithm then follows the estimated gradient downhill, $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \mathbf{g}$ where ϵ is the learning rate, in the hope of minimizing the loss.

Optimizing the parameters to represent all valid inputs of a task, where the data is often very high-dimensional (e.g., images, sounds, text), may seem hopeless. However, neural networks surmount this obstacle by assuming that these high-dimensional data are lying along low-dimensional manifolds[1] (Goodfellow, Bengio, and Courville, 2016). An intuitive observation in favour of this claim is that uniform noise essentially never resembles structured inputs from these tasks. More rigorous experiments supporting the manifold hypothesis are (Cayton, 2005; Narayanan and Mitter, 2010; Weinberger and Saul, 2006).

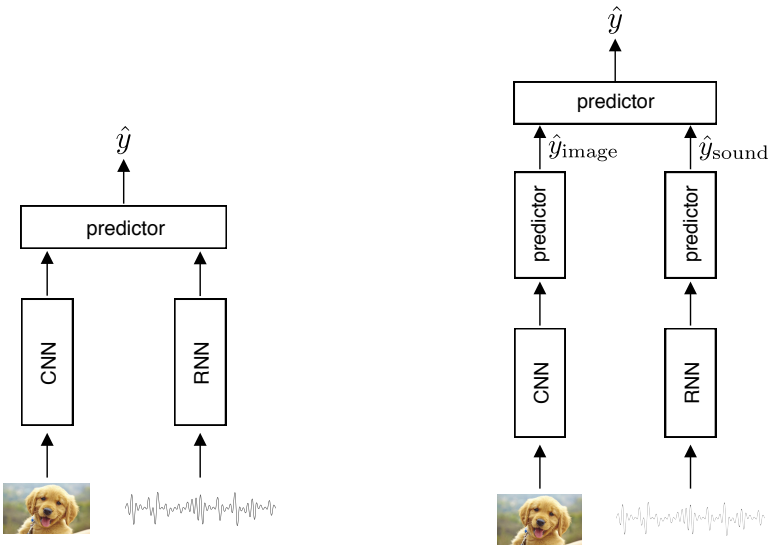
Multi-Modal Deep Learning

As a reminder, a modality refers to "the way in which something happened or is experienced" (Baltrušaitis, Ahuja, and Morency, 2019). Multi-Modal Deep Learning (MMDL) is simply the research area of neural networks using input samples consisting of multiple modes. Baltrušaitis et al. identified five non-exclusive use-cases of MMDL,

- *Representation*: learning how to represent and summarize multi-modal data in a way that exploits the complementarity and redundancy
- *Translation*: learning how to map data from one modality to another (e.g., image captioning)
- *Alignment*: learning to identify the direct relationships between elements from two or more different modalities (e.g. alignment of sound and video)
- *Fusion*: learning to join information from two or more modalities to perform predictions
- *Co-learning*: learning to transfer knowledge between modalities and their respective predictive models (e.g., zero shot learning)

[1] A manifold designates a connected set of points that can be approximated well by considering only a small numbers of degrees of freedom.

The EMMA module is applied to multi-modal networks performing fusion. Furthermore, networks doing fusion can combine their modalities in three different ways: by early-fusion, late-fusion and an hybrid of the first two. Early-fusion architectures have uni-modal encoders extracting the features of each mode, the obtained features are then concatenated altogether and fed into a common decoder making the predictions (see Figure 2.2a). In contrast, late-fusion has uni-modal predictors for each mode, followed by a decoder weighting the uni-modal predictions to compute the final prediction.



(A) Early-fusion (B) Late-fusion

FIGURE 2.2: Fusion of images and sounds for a classification task with a Convolutional Neural Network (CNN) (He et al., 2016), Recurrent Neural Network (RNN) (Wu et al., 2016).

2.3 Physics meets Deep Learning

Modelling complex probability distributions by parametric functions such as deep learning models is a difficult task, because all the probabilities must be positive and sum up to one. At its origins, many researchers in deep learning had an academic background in physics, from which they regularly found inspiration to solve problems. An example of this is the distribution of kinetic energies among molecules of gas, called the Boltzmann distribution, and given by

$$p(E_i) = \frac{1}{Z} e^{-E_i/k_B T} \quad \text{with the partition function} \quad Z = \int e^{-E_j/k_B T} \tag{2.2}$$

where E_i is the kinetic energy of molecule i, k_B the Boltzmann constant and T the temperature of the environment. The first thing to notice is that all the probabilities are positive and sum up to one for any set of combinations of energies E_i. Another observation to make is that high values of energies are unlikely ($E_i \propto -\log p(E_i)$), unless the temperature is sufficiently high enough. To sum it up, the Boltzmann distribution can be used to normalize any function to a distribution, where the temperature T is a parameter influencing the entropy of the distribution. The Boltzmann distribution has

two major applications in deep learning. First, it corresponds to the soft-max activation function (Duan et al., 2003), employed commonly for the purpose of outputting probabilities in multi-category classification tasks. Secondly, it was also used by deep learning researchers to construct energy-based models (LeCun et al., 2006). These types of neural networks optimize an energy function to be low on the data manifold and high everywhere else (see Figure 2.3), which is the mapped to probabilities via the Boltzmann distribution. A few examples of efficient energy-based models are Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), Variational Autoencoders (VAE) (Kingma and Welling, 2013) and Denoising Autoencoders (DAE) (Vincent et al., 2008). The latter will be used in this work to measure the outlyingness of the data (see Chapter 4).

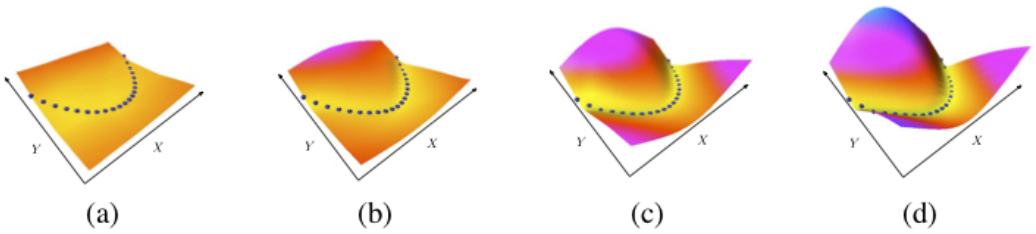


FIGURE 2.3: The shape of the energy surface at four intervals. Along the x-axis is the variable X and along the y-axis is the variable Y . The shape of the surface at (a) the start of the training, (b) after 15 epochs over the training set, (c) after 25 epochs, and (d) after 34 epochs. The energy surface has attained the desired shape: the energies around the training samples are low and energies at all other points are high. *Image and caption from* (LeCun et al., 2006).

Chapter 3

Literature Review

The purpose of this chapter is to review the state-of-the-art literature of multi-modal attention. The first section describes attention in humans from both a psychological and a neurological point of view. We argue this will give the reader more intuition about attention in deep learning. The second part moves on to the different attention mechanisms in deep learning, in particular self-attention and cross-modal attention.

3.1 Attention in Humans

"The most profound effect of attention is its capacity to bring the attended stimuli into the forefront of our conscious experience while unattended stimuli fades into the background, increasing the processing efficiency at every stage of perception" (Watzl, 2017). "A widely held assumption in the psychology literature is that the most fundamental function of attention is selection. At the level of single neurons, neuroscientists typically thought of attention in terms of selection between stimuli competing for the same neural receptive field" (Desimone and Duncan, 1995). Daniel Kahneman, an authorithy in psychology and economy, investigated the way in which humans perform multi-tasking (i.e., solve a multi-modal problem). Kahneman claimed that attention was more than selection, that it could be viewed as a limited resource being shared among the different modes, but he could not generalize his findings to the intra-modal level[1]. Moreover, the selection theory has been vigorously challenged in recent years by the amplification theory, where attention is an additional activity that interacts with built-in perceptual mechanisms by amplifying some of the input signals (Fazekas and Nanay, 2018). Furthermore, the absolute intensity of amplification is not important, in contrary it is the relative intensity between the inputs that matters (*the contrast effect*). Notice that the amplification theory generalizes the concept of capacity to the intra-modal level and neural level. Interestingly, we will see that the basic principles of attention mechanisms in deep learning has significant similarities with amplification.

Regarding multi-modal attention, three types can be distinguished: endogenous, exogenous and cross-modal attention (Driver and Spence, 1998). "People orient their attention endogenously whenever they voluntarily choose to attend to something, such as when listening to a particular individual at a noisy cocktail party, or when concentrating on the texture of the object that they happen to be holding in their hands. By contrast, exogenous orienting occurs when a person's attention is captured reflexively by the sudden onset of an unexpected event, such as when a mosquito suddenly lands

[1] Intra-modal attention manifests itself only in a subset of the mode, whereas inter-modal attention is between modes.

on our arm" (Driver and Spence, 1998). Lastly, cross-modal attention refers to the interaction of attention between two or more modes such as using visual clues (e.g. lip movements) to focus on the voice of a particular individual at a noisy cocktail party.

3.2 Attention in Deep Learning

"Attention mechanisms in deep learning aim to highlight specific regions of the input space" (Chaudhari et al., 2019). The most common way to do this, is by multiplying the input by an attention mask, where the attention mask consist of normalized continous values between zero and one. Observe the similarity with the amplification theory described in the previous section. In self-attention (Bahdanau, Cho, and Bengio, 2014), the attention mask is computed from the same mode on which it is applied. Conversely, for cross-modal attention mechanisms (Li et al., 2019), the attention mask is computed from multiple modes.

Self-attention was first introduced in natural language processing (NLP) for machine translation tasks by (Bahdanau, Cho, and Bengio, 2014). It helped the translation task by enabling the model to automatically search for parts of a source sentence that are relevant to predicting the next target work. With this approach, Bahdanau et al. achieved a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-French translation. Since then it has become a prominent tool in NLP but has also been used in a variety of other tasks such as image classification. (Hoogi et al., 2019) uses self-attention to learn to suppress irrelevant regions in images and highlight salient features useful for the specific classification task. The authors in (Hoogi et al., 2019) reduced the computation load and were able to compensate the absence of a deeper network by using the self-attention, without having a decreased classification performance. For a detailed review on this self-attention mechanisms, see (Galassi, Lippi, and Torroni, 2019).

Turning now to cross-modal attention, (Ephrat et al., 2018) presents an audio-visual model for isolating a single speech signal form a mixture of sounds such as other speakers and background noise (see Figure 3.1). Cross-modal attention is used to focus on certain parts of the audio with respect to an image of the desired speaker. The authors showed superior results compared to state-of-the-art audio-only methods. Similar works (Libovický, Helcl, and Mareček, 2018; Li et al., 2019; Wang, Wang, and Wang, 2018) are using cross-modal attention and have attained impressive results. However, most research using cross-modal attention has tended to focus on obtaining better predictions rather than improving the robustness. A few exceptions are discussed below.

A work investigating how multimodal fusion can help against failing modes is (Afouras et al., 2018). Their model fuses audio and video to obtain better speech-to-text. Interestingly, Afouras et al. use a combination of self-attention mechanisms followed by a cross-modal attention layer. The model was tested on thousands of natural sentences of British television. Furthermore, they added babble noise with 0dB signal-to-noise ratio to the audio streams, where the babble noise samples are synthesized by mixing the signals of 20 different audio samples from the dataset. The audio-visual model achieved a 13.7% word error rate (WER) on the dataset without noise, and a 33.5% WER on the dataset with noise whereas the audio-only model only achieved 64.7% WER. Despite obtaining great results, a major weakness with this experiment, however, is that their test set is corrupted in the exact same manner as their training set. In our experiments[2]

[2]See Section 6.3.2

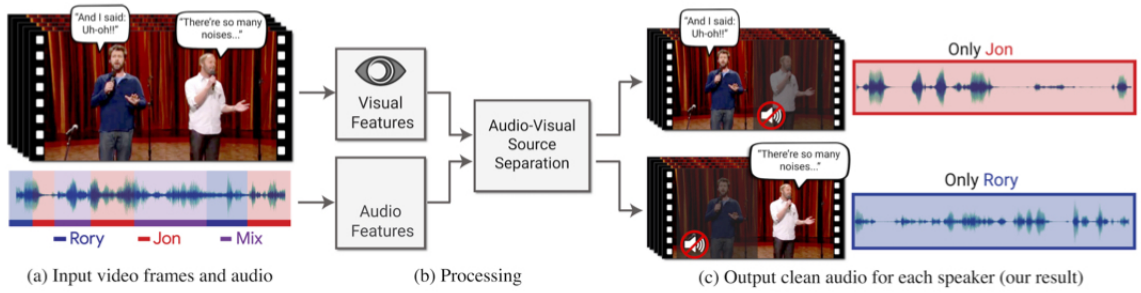


FIGURE 3.1: The authors of (Ephrat et al., 2018) present a model for isolating and enhancing the speech of desired speakers in a video. Their model was trained using thousands of hours of video segments from our new dataset, AVSpeech. *Image from* (Ephrat et al., 2018).

we will show that evaluating test data with the same noise as on the training data can significantly overestimate the robustness of the model. Additionally, the attention module in (Afouras et al., 2018) is presumably not able to detect and handle unseen samples. To summarize, the model was not tested against realistic failing modes situations.

The work that is most relevant to our proposed method is the attentive context proposed in (Shon, Oh, and Glass, 2018), which also incorporates attention on the inter-modal level to explicitly filter perturbations out (see Figure 3.2). The model is evaluated on a face-verification task, receiving a **v**oice sound and a **f**ace image. The attention mask $[\alpha_v, \alpha_f]$ is computed via a linear function, $f_{\text{att}} = \mathbf{W}^T[\mathbf{e}_v, \mathbf{e}_f] + \mathbf{b}$, on the embeddings $\mathbf{e}_v$ and $\mathbf{e}_f$. Several defects of the attention function f_{att} can be observed:

1. The function is unlikely to be expressive enough to capture complicated dependencies between the modes, and to recognize out-of-distribution[3] data.
2. A design constraint of this attention function is that all extracted embeddings must be of the same size, which may be a significant constraint when combining modes from low and high-dimensional data.
3. Similarly to (Afouras et al., 2018), the test set is corrupted in the same manner as the training set.

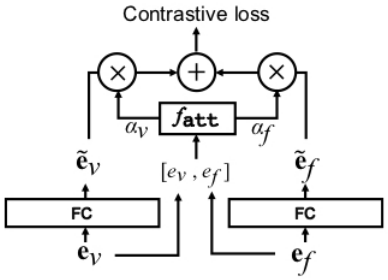


FIGURE 3.2: Neural network based fusion approaches. $\mathbf{e}_v$: speaker embedding, $\mathbf{e}_f$: face embedding. FC denotes a fully connected layer. *Image from* (Shon, Oh, and Glass, 2018).

[3] Relative to the training data.

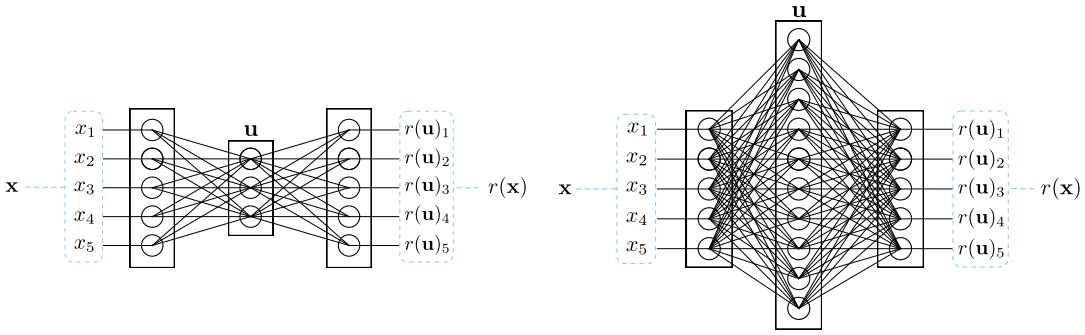
Chapter 4

Energy Estimation

In the introduction we suggested that a mode of the input sample that is significantly different from the training distribution may cause undesired activations in the neural network, i.e. the failure intensity of a mode is correlated with its likelihood relative to the training distribution. Furthermore, in Section 2.3 we explained that the energy function in energy-based models is proportional to the negative log-likelihood (NLL), which could thus be used to evaluate the failure intensity of a mode. This chapter discusses how to derive the energy function of a denoising autoencoder, which will be used as a metric for the failure intensity (discussed in Chapter 1).

4.1 Autoencoders

Autoencoders (AE) are models trained to reproduce their inputs to their outputs. An autoencoder is composed of two main parts, the encoder f and the decoder g. The input $\mathbf{x} \in \mathbb{R}^L$ is passed through the encoder $f : \mathbb{R}^L \mapsto \mathbb{R}^U$ as $f(\mathbf{x}) = h(W_f\mathbf{x} + \mathbf{b}_f) = \mathbf{u}$, where $h(\cdot)$ is an activation function applied element-wise and $\mathbf{u}$ represents the hidden layer. The decoder $g : \mathbb{R}^U \mapsto \mathbb{R}^L$ is then in charge of reconstructing the input, $g(\mathbf{u}) = W_g\mathbf{u} + \mathbf{b}_g$. The output is often called the reconstruction and is written $r(\mathbf{x}) = g(f(\mathbf{x}))$ with $r : \mathbb{R}^L \mapsto \mathbb{R}^L$. Autoencoders are trained in an unsupervised manner, most of the time using the mean-squared error between input and output as a loss function, $\mathcal{L}_{\text{MSE}} = \|r(\mathbf{x}) - \mathbf{x}\|_2^2$. Training a model to copy its input may seem useless. To answer this point, we need to distinguish two families of autoencoders, namely undercomplete and overcomplete autoencoders.



(A) Undercomplete AE (B) Overcomplete AE

FIGURE 4.1: The architecture of the two families of autoencoders

Undercomplete autoencoders

An autoencoder is said to be undercomplete when the size of the hidden layer $\mathbf{u}$ is smaller than the size of the input/output layers, i.e. $U < L$ (see Figure 4.1a). This amounts to constructing a low-dimensional representation of the input, and information is therefore lost in the process. It can be thought of as a non-linear principal component analysis (Scholz, Fraunholz, and Selbig, 2008; Ladjal, Newson, and Pham, 2019) as the values formed in the hidden layer are a non-linear representation in latent space of the input. As can be seen in Figure 4.2, minimizing the mean squared error is similar to minimizing the Euclidean norm of the vector $r(\mathbf{x}) - \mathbf{x}$.

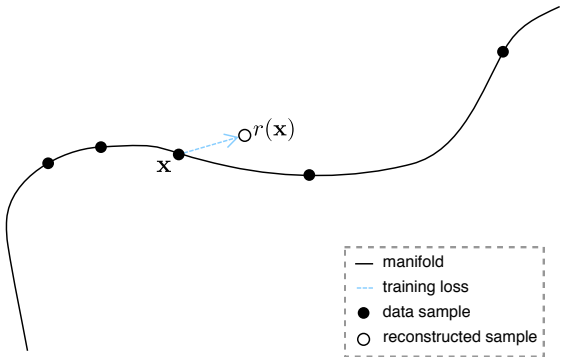


FIGURE 4.2: Vectorial representation of an undercomplete reconstruction process.

Overcomplete autoencoders

Conversely, an overcomplete AE has more hidden units than its input/output layer, i.e. $U > L$ (see Figure 4.1b). Hence, the model could thus learn to perfectly copy its input through the U hidden units and reproduce it at the output. However, the input is corrupted before being passed through the encoder, whereas the decoder is forced to reconstruct the original input, i.e. the model learns to denoise signals. This type of AE is called a denoising autoencoder (DAE). More formally, the input is corrupted with some small isotropic noise $\tilde{\mathbf{x}} = \mathbf{x} + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, with the training loss

$$\mathcal{L}_{\text{MSE}} = \|r(\tilde{\mathbf{x}}) - \mathbf{x}\|_2^2 \tag{4.1}$$

It is worth noticing the difference with the loss function of the undercomplete AE. We verify in Figure 4.3 that minimizing the loss implies that the *reconstruction error* $r(\tilde{\mathbf{x}}) - \tilde{\mathbf{x}}$ will converge to $-(\tilde{\mathbf{x}} - \mathbf{x})$ i.e. the model learns to invert the corruption process.

4.2 Energy in Autoencoders

The authors in (Alain and Bengio, 2012) found that the reconstruction error of a trained denoising autoencoder is proportional to the score (gradient of log-likelihood)

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

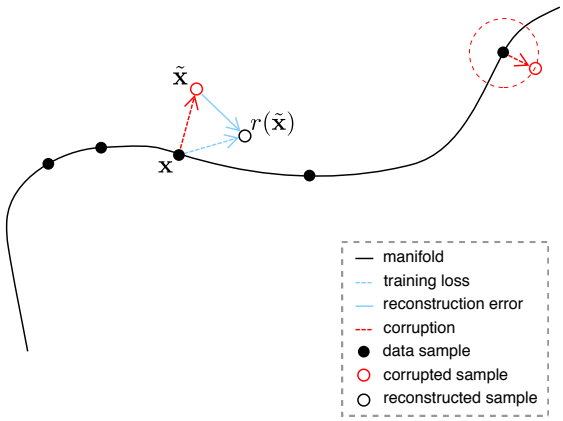


FIGURE 4.3: Vectorial representation of an overcomplete reconstruction process.

To put it differently, the reconstruction error points towards the corresponding most likely datapoint. This result is not particularly suprising, as denoising a signal is essentially equivalent to finding the most likely datapoint among nearby samples in the distribution (see Figure 4.3). To illustrate Equation (4.2), we train a DAE on a generated circle manifold (more details about this experiment in Section 4.3). As we can see below, the vector field of the reconstruction error does indeed point towards the data manifold[1]. Alternatively, Figure 4.4 may be interpreted in terms of forces deriving from

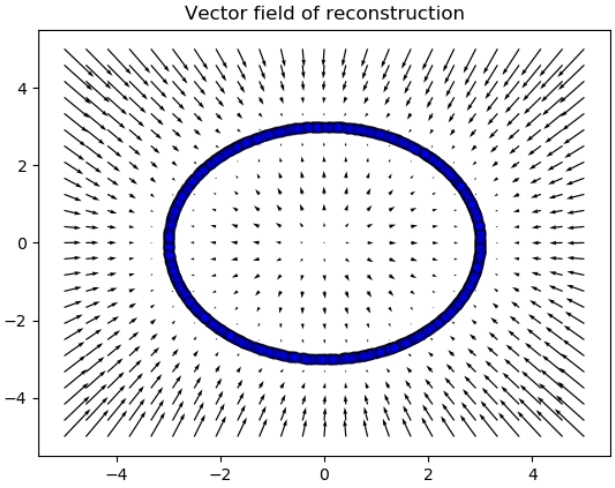


FIGURE 4.4: Vector field of reconstruction error on circle manifold. No corruption is applied at test time, the reconstruction error vector is simply the output minus the input. The experimental setup is described in Section 4.3.

potential fields as observed in physics. This interpretation can be useful to distinguish the manifold since it acts as a sink in the vector field i.e., has a low potential energy.

A vector field is the gradient of a potential energy field if it satisfies a simple condition called the integrability criterion[2]. In (Kamyshanska and Memisevic, 2014), the

[1] We refer broadly to a data manifold as a connected set of points that can be approximated well by considering only a small numbers of degrees of freedom.

[2] See Appendix C.1

authors found that using tied weights ($W_f = W_g^T = W$), produces an autoencoder configuration whose reconstruction error field satisfies the aforementioned integrability criterion. Indeed, one successively finds

$$\begin{aligned}\frac{\partial(r(\tilde{\mathbf{x}})_i - \tilde{x}_i)}{\partial \tilde{x}_j} &= \frac{\partial\left[W^T h(W\tilde{\mathbf{x}} + \mathbf{b}_f) + \mathbf{b}_g\right]_i}{\partial \tilde{x}_j} - \frac{\partial \tilde{x}_i}{\partial \tilde{x}_j} \\ &= \sum_k W_{ik} \frac{\partial h(W\tilde{\mathbf{x}} + \mathbf{b}_f)}{\partial(W\tilde{\mathbf{x}} + \mathbf{b}_f)} W_{jk} - \delta_{ij} \\ &= \frac{\partial(r(\tilde{\mathbf{x}})_j - \tilde{x}_j)}{\partial \tilde{x}_i}\end{aligned} \tag{4.3}$$

where δ_{ij} denotes the Kronecker delta. Hence, under this assumption, the reconstruction error can be expressed as the gradient of a scalar field, the potential energy $-\Psi$, such that $r(\tilde{\mathbf{x}}) - \tilde{\mathbf{x}} = -\partial\Psi(\tilde{\mathbf{x}})/\partial\tilde{\mathbf{x}}$. Thereupon, the reconstruction process can be seen as a gradient descent in the potential energy space (Kamyshanska and Memisevic, 2014). A key insight to gain from these developments and Equation (4.2) is that the aforementioned potential energy is in fact proportional to the NLL,

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

Since the gradient of the potential energy can be expressed in terms of the reconstruction error, the potential energy Ψ can be computed by integrating the latter,

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

Now, expressing r in terms of $f = h(W\mathbf{x} + \mathbf{b}_f)$ and $g = W^T f(\mathbf{x}) + \mathbf{b}_g$ in (4.5), and following the developments made in (Kamyshanska and Memisevic, 2014), we obtain

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

In this work only the sigmoid will be used as an activation function h, so that the expression of f can be written explicitly and

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

where $W_{.k}^T$ is the k^{th} column of W^T, and b_k^f the k^{th} entry of $\mathbf{b}_f$. All intermediate steps between (4.5) and (4.7) are detailed in (Kamyshanska and Memisevic, 2014). It is worth remarking that the potential energy can be negative by construction.

4.3 Experiment I

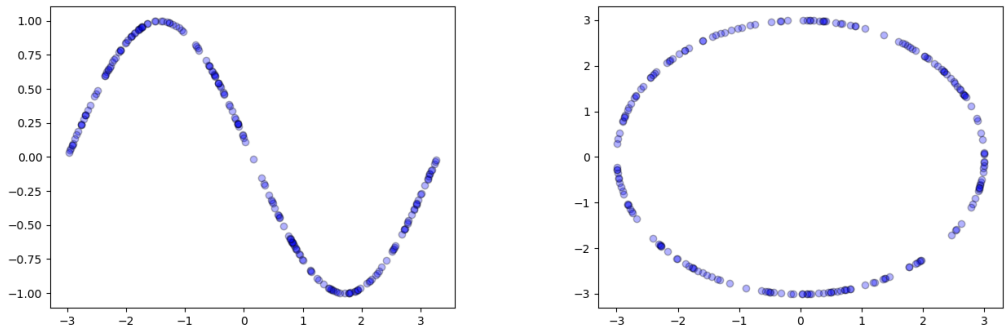
In this experiment, two simple data manifolds are generated, on which separate denoising autoencoders are trained. In order to evaluate the suitability of the potential energy as a proxy for the NLL, the former is plotted on a grid for each of these autoencoders.. As a comparison, we also compute the reconstruction error, $\|r(\tilde{\mathbf{x}}) - \tilde{\mathbf{x}}\|_2^2$, which is sometimes used in the Machine Learning community as a way to detect outliers.

Manifolds

The manifolds maps a set of N scalar values $\{t_1, \ldots, t_N\}$ drawn uniformly at random in $[0, 2\pi]$ into a set of vectors $\{\mathbf{x}^1, \ldots, \mathbf{x}^N\} \subset \mathbb{R}^2$. The entries of these vectors are obtained as

$$\text{wave} \begin{cases} x_1^k = t_k - \pi \\ x_2^k = \sin(t_k) \end{cases} \qquad \text{circle} \begin{cases} x_1^k = 3\sin(t_k) \\ x_2^k = 3\cos(t_k) \end{cases}$$

The resulting structures are illustrated in Figure 4.5.



(A) Wave (B) Circle

FIGURE 4.5: Manifold generation with 200 samples.

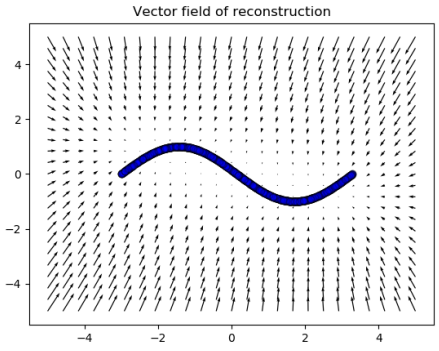
Setup

Each autoencoder has 8 hidden units, is trained for 25 epochs, with a batch size of 100, a corruption noise $\sigma = 0.008$ and a learning rate of $1e^{-3}$. The optimizer used is *Adam* (Kingma and Ba, 2014).

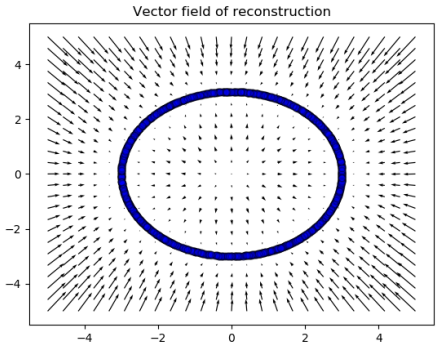
Results

As expected the vector fields of the reconstruction errors are directed towards the manifolds (see Figure 4.6), the manifolds acting as sinks. Notably, observe the presence of a source at the origin for the circle manifold (see Figure 4.6b).

The energy function and the norm of the reconstruction error are computed and plotted onto heatmaps (see Figure 4.7). We can see that the two estimators have low values in the neighbourhood of the manifold and are high everywhere else. However, the norm of the reconstruction error has also low values at the origin, due to the source in its vector field. This issue makes it difficult in general to use it as a robust criterion to detect and quantify out-of-distribution samples.

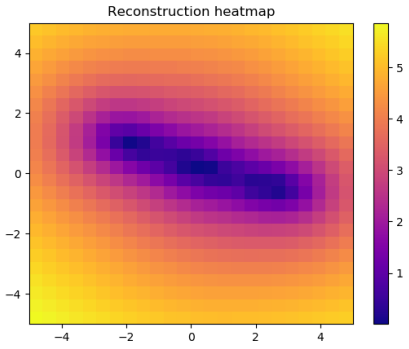


(A) Wave

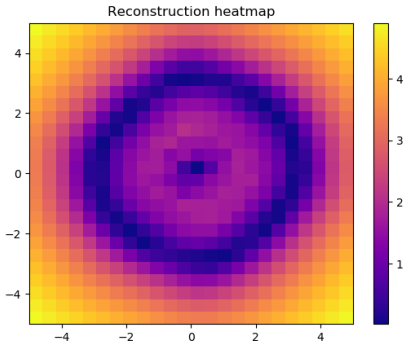


(B) Circle

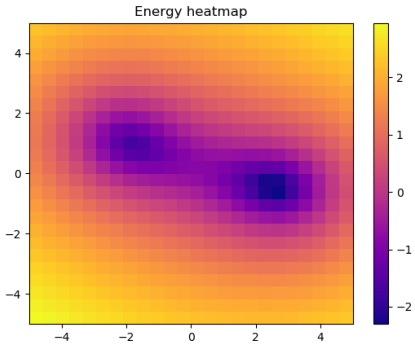
FIGURE 4.6: Vector fields of the reconstruction error evaluated on a mesh grid.



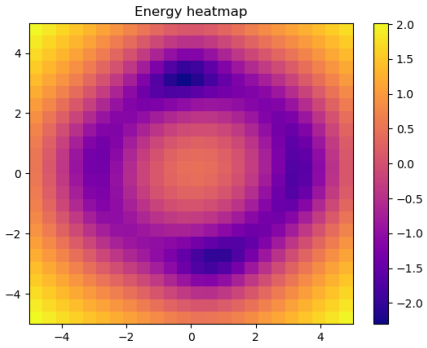
(A) Wave - Norm reconstruction error



(B) Circle - Norm reconstruction error



(C) Wave - Potential energy



(D) Circle - Potential energy

FIGURE 4.7: Estimators evaluated on wave (left) and circle (right) manifolds.

4.4 Limitations

Many interesting data structures are difficult to reproduce with shallow denoising autoencoders. For example, sequential data (e.g. sound) is better modelled with LSTM-DAE[3]. Likewise, CNN-DAE[4] are more appropriate to model spatial structures, such as images. However, the integrability criterion for these models is not satisfied anymore, and thus the negative log-likelihood cannot be estimated. Alternative methods to efficiently learn energy functions for spatial or sequential data are presented in (Zhai et al., 2016; Kim and Bengio, 2016)

[3] Long-Short Term Memory denoising autoencoder (Chen et al., 2018)
[4] Convolutional Neural Network denoising autoencoder (Turchenko, Chalmers, and Luczak, 2017)

Chapter 5

Energy-based Multi-Modal Attention

The literature review (Chapter 3) showed that previous research in MMDL has been mostly focused on leveraging multi-modality to improve the accuracy of the predictions. In this chapter, a new attention module is presented to increase the robustness against failing modes: as long as at least one modality provides sufficient information for the task at hand, the prediction network will be able to perform well. First, we start by providing a conceptual general framework. Then, the design of each step of the framework is described. Finally, the training of EMMA is discussed, along with two novel regularizers.

5.1 General Framework

A typical multi-modal network (MMN) receives samples at its input, where each sample is composed of multiple modes such as images and sounds. In real-world applications, the relative informativeness of different modes may evolve over time, on a per sample basis, e.g. as a result of perturbations or sensor malfunctions.

In order to address this problem, an attention module is proposed that pre-processes each input sample and evaluates the relative informativeness, also referred to as importance, of each mode. More precisely, those modes deemed informative are assigned a high weight, typically close to 1, whereas the modes considered too uninformative are assigned a weight close to 0. The weighted modes are then fed to the MMN.

The interpretation of the role of EMMA is twofold. First, EMMA can be seen as a sort of gate filtering out perturbations. Indeed, failing modes can provoke high activations in the MMN, thus affecting its predictive performance negatively. Masking the failing modes allows to diminish these activations, thereby improving predictions quality. Another way to view it is to understand that the MMN model is able to extract the multiplied weight from the original input. The model can then learn to make more robust predictions based on the extra information provided by that weight. Notice that even tough the internal architecture of the MMN is often structured as a many-to-one encoder-decoder as discussed in Section 2.2, the EMMA module can be fitted to any MMN architecture.

The key concept on which EMMA relies to produce scores for different modes is that of modal importance[1]. As a reminder, the modal importance is defined in terms of three intrinsic and related properties of each mode, namely

- *relevance*: the intrinsic informativeness of the mode for the predictive task at hand.
- *failure intensity*: the propensity of a mode to trigger undesirable activations in the neural network.
- *coupling*: the interdependencies between the modes, which describe the extent to which the mode provide independent, complementary, redundant or conflicting information.

The EMMA module will essentially try to learn the relationships between these properties for a specific dataset.

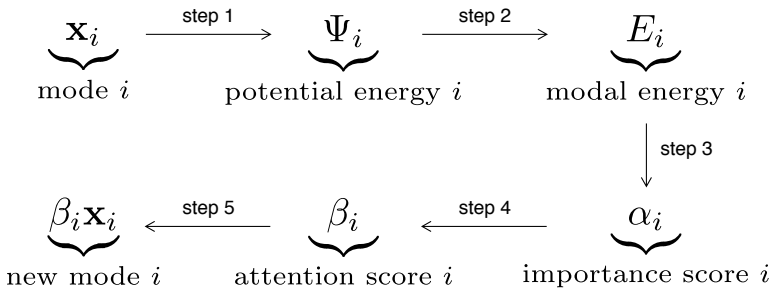


FIGURE 5.1: Summary of main steps in EMMA (step 2, 3 and 4 are detailed in the following sections, step 1 was explained in Chapter 4)

For the sake of clarity, making the exposition more formal is now in order. Let $\mathcal{D}^N = \{(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_N, y_N)\}$ be a dataset of i.i.d. samples, where the input $\mathbf{X}_k$ is composed of M modes $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$. The MMN tries to make predictions $\hat{y}_k$ as close as possible to the groundtruth y_k. The EMMA module computes the importance of mode i starting with the failure intensity, which is measured by the potential energy $\Psi_i = \Psi(\mathbf{x}_i)$ (step 1 in Figure 5.1, previously motivated in the introduction of Chapter 4). To capture the two other modal properties, namely the modal relevance and coupling, two additional functions are introduced. On the one hand, the *self-energy* e_i is designed to spontaneously learn the relationship between the relevance and the failure intensity. This comes as a result of the fact that it is a function of Ψ_i and its parameters are optimised with respect to the loss on the predictions. On the other hand, shared energies e_{ij} are designed to capture the optimal coupling between modes. Using these functions, the *modal energy* E_i can be constructed such that it takes a low value if mode i is important and a high value otherwise (step 2 in Figure 5.1, further discussed in Section 5.2). Next, the modal energies are normalized via the Boltzmann distribution[2] to form *importance scores* α_i, which are scalar values between zero and one, with more important modes corresponding to higher values (step 3 in Figure 5.1, further discussed in Section 5.3). From each importance score, the model then determines an *attention score* β_i, which is representative of the amount of attention that should be paid to each mode by the MMN (step 4 in Figure 5.1, further discussed in Section 5.4). Finally, each mode

[1] Introduced in Section 1.2
[2] See Section 2.3

is multiplied by its respective attention score (step 5 in Figure 5.1). Section 5.4 will clarify why the modes are not directly multiplied by the importance scores instead. A high-level view of the attention module EMMA is illustrated in Figure 5.2. The next sections detail the form and specificities of each function introduced previously.

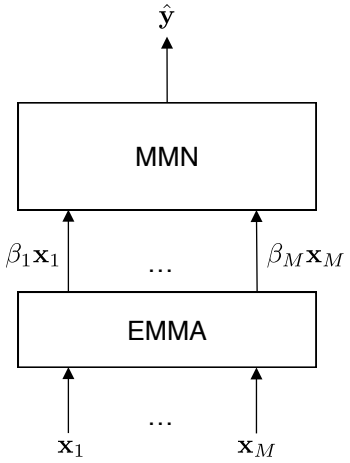


FIGURE 5.2: High-level view of a Multi-Modal Network with the EMMA module.

5.2 From Potential to Modal energies (step 2)

The self-energy is defined as an affine function of the potential energy,

$$e_i = w_i \Psi_i + b_i \qquad \text{with} \qquad w_i \in [1, +\infty],\ b_i \in \mathbb{R}^+, \tag{5.1}$$

where the parameters w_i and b_i are trained via a loss function on the predictions. Therefore, the model is able to capture both the relevance and failure via the self-energy. The second advantage of this transformation is that it helps the module to handle potentials on different numerical scales. Indeed, Equation (4.7) only guarantees being proportional to the NLL, thus potentials of different modes may not be on comparable scales. The reason the parameters are constrained to be positive will be justified below. Additionally, it will be shown below that self-energies are guaranteed to be positive at the end of this section.

After computing the self-energies, the shared energies can be determined. The expression e_{ij} denotes the shared energy of mode j on i and is constructed from the self-energies as follows

$$e_{ij} = w_{ij} e_i^{\gamma_{ij}} e_j^{1-\gamma_{ij}} \qquad \text{with} \qquad w_{ij} \in [-1, +1],\ \gamma_{ij} \in [0, 1] \tag{5.2}$$

where the parameter γ_{ij} learns the degree of coupling in the spectrum from strongly coupled ($\gamma_{ij} = 0$) to independent ($\gamma_{ij} = 1$). Indeed, if the model learns a value of γ_{ij} close to zero, mode j will influence mode i much more than for a γ_{ij} close to unity. Equally important is the direction of coupling between mode i and j, determined by the weights w_{ij} and w_{ji}. We verify that an increase/decrease of the self-energy e_j leads to an increase/decrease of the modal energy E_i for a positive weight w_{ij}, and a decrease/increase of E_i for a negative weight w_{ij}. This is valid since self-energies are guaranteed to be positive (see next paragraph). The direction of coupling is useful to

distinguish modes with redundant or conflicting information from those with complementary information. Notice that the degree and direction of coupling are asymmetric ($\gamma_{ij} \neq \gamma_{ji}$, $w_{ij} \neq w_{ji}$). This asymmetry is justified by the following example: take a multi-modal problem with three modes A, B and C. We want the model to learn that if mode A is failing, it is optimal that mode B "takes over". And if mode B is failing, it is optimal for C to "take over". This example can only be modelled with asymmetry. In conclusion, the model has the ability, through the use of shared energies, to discover the different interdependencies between the modes.

A consequence of the design of Equation (5.2) is that the evaluation of the gradient during the backpropagation step now involves taking the logarithm of e_i[3], which is undefined for negative values. As the weights in Equation (5.1) are positive, we only have to make sure the values of the potential energy are positive. The latter is done by lowering the potential Ψ_i to Euler's number e as

$$\Psi_i \leftarrow \max(\mathrm{e}, \Psi_i - \Psi_i^{(\min)} + \mathrm{e}) \tag{5.3}$$

where $\Psi_i^{(\min)}$ denotes the lowest value of Ψ_i in the training set. This correction avoids undefined values ($\Psi_i \geq 0$), exploding gradient[4] ($\Psi_i \geq \mathrm{e}$) and guarantees self-energies to be positive. The reason a max-operator is used is because lower energy values than $\Psi_i^{(\min)}$ can occur during inference. Clearly, this correction step (5.3) must be performed prior to the computation of self-energies (5.1).

5.3 From Modal energies to Importance scores (step 3)

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

$$\alpha_i = \frac{1}{Z} e^{-\rho E_i} \quad \text{with the partition function} \quad Z = \sum_{j=1}^{M} e^{-\rho E_j} \tag{5.4}$$

This guarantees the scores are normalized and sum up to one. A mode i will be said to be important if its score is close to one (low modal energy E_i). The hyperparameter ρ represents the coldness, 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. As can be observed in Figure 5.3, the coldness has a significant influence on the overall behaviour of the attention module. Hence, careful tuning of ρ is required.

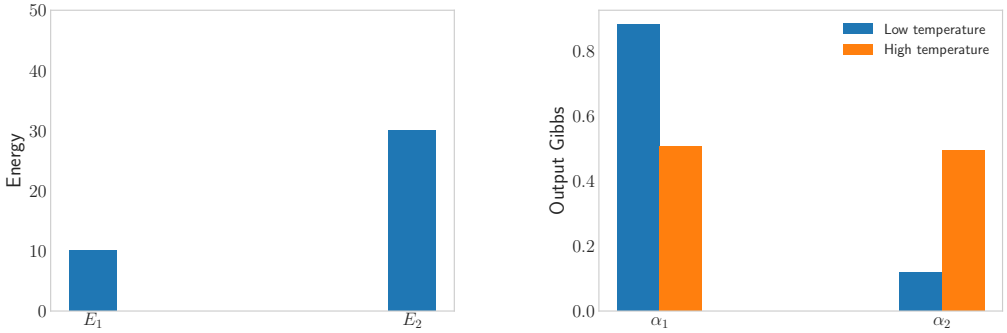
5.4 From Importance to Attention scores (step 4)

The attention scores are given by

$$\beta_i = \max[0, \tanh(g_a \alpha_i - b_a)] \quad \text{with} \quad g_a > 0,\ b_a \in [0, 1] \tag{5.5}$$

[3] See Appendix C.2

[4] Exploding gradients are very large gradients, which in turn results in large updates of the network weights, resulting in an unstable network. A good overview on this subject can be found in (Philipp, Song, and Carbonell, 2017)



(A) Energies (B) Importance scores

FIGURE 5.3: Input-output of Boltzmann distribution for two different temperatures, low temperature ($\rho = 0.1$) and high temperature ($\rho = 0.001$)

The hyperbolic tangent adds non-linearity while the gain g_a and bias b_a enable the model to control the threshold and capacity (see Figure 5.4). Those two concepts are detailed below.

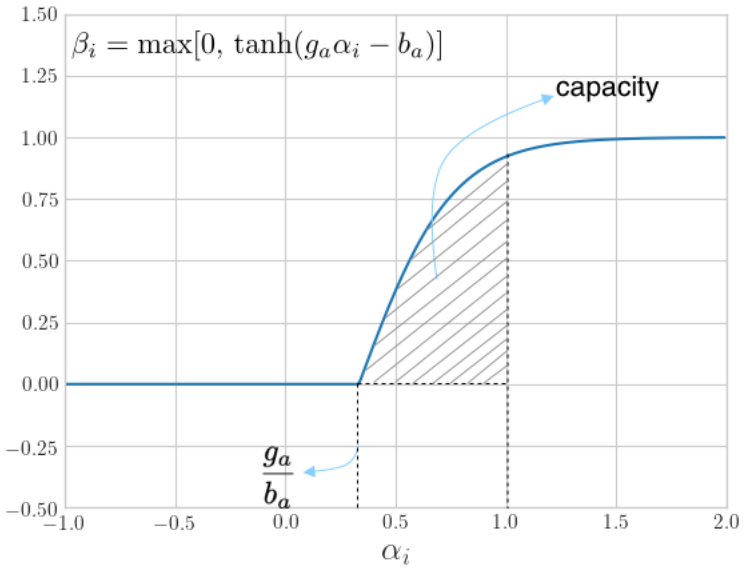


FIGURE 5.4: Attention function (the max-operator generalizes the attention function to cases where $\alpha \in \mathbb{R}$)

Energy threshold

The module will let the information of mode i pass by, only if $g_a\alpha_i - b_a > 0$

$$\begin{aligned} &\Leftrightarrow \log(\alpha_i) > \log(b_a/g_a) \\ &\Leftrightarrow E_i \leq \frac{\log(g_a/b_a) - \log(Z)}{\rho} = E_{\text{threshold}} \end{aligned} \tag{5.6}$$

where $E_{\text{threshold}}$ represents the maximum energy level for mode i to be taken into account (see Figure 5.4). We deduce that the learned gain and bias control this threshold. Notably, the threshold is varying on a per-sample basis due to the partition function Z. For example, an increase of the overall perturbation level on the entire input results in a reduction of the partition function, leading to a higher threshold $E_{\text{threshold}}$. It can thus be deduced that EMMA becomes less selective if the overall quality of the sample decreases.

Capacity

A more common way to write the attention function would be $\tanh(\mathbf{W}\boldsymbol{\alpha}+\mathbf{b})$, whereas we have $\tanh(g_a I\boldsymbol{\alpha} - b_a\mathbf{1})$, where I is the identity matrix. We argue the latter better mimics human's attention, permitting us to introduce the concept of capacity, which in psychology is viewed as the amount of resource that can be allocated (Kahneman, 1975). If we look at Figure 5.4, this can be translated as,

$$\text{capacity} \triangleq \int_0^1 \tanh(g_a\alpha - b_a)d\alpha \tag{5.7}$$

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

$$\frac{du}{d\alpha} = g_a \Leftrightarrow d\alpha = \frac{1}{g_a}du \tag{5.8}$$

we can write

$$\begin{aligned}\text{capacity} &= \frac{1}{g_a}\int_{-b_a}^{g_a-b_a} \tanh(u)du \\ &= \frac{1}{g_a}\log[\cosh(u)]\Big|_{-b_a}^{g_a-b_a} + \cancel{\text{constant}} \\ &= \frac{1}{g_a}\log\left[\frac{\cosh(g_a-b_a)}{\cosh(-b_a)}\right]\end{aligned} \tag{5.9}$$

When the capacity is too low, no sufficient amount of information is passed to the MMN, leading to wrong predictions. Similarly, if the capacity is too high, the perturbations of the failing modes will pass and cause a decrease in performances. It is expected that the model learns the optimal trade-off. However, if we want the attention module to be robust against failing situations it was not trained on, we suggest minimizing capacity would result in appealing properties for the system as a whole. The reasons for minimizing the capacity are two-fold: it forces the module to mask out more information, which may be sub-optimal on the training set but useful to generalize against more intensive failing modes. The second reason is to avoid the extreme case where the module leaves all the inputs unchanged (maximal capacity) and instead it is the MMN who learns to suppress perturbations. The ideal method for controlling the capacity, would be to add the derived expression (5.9) as a regularizer to the loss function, but this could potentially lead to instabilities[5] during training. A less precise but more stable way is to introduce the expression $g_a - b_a$ instead in the loss function. We verify that minimizing this expression will minimize the gain while maximizing the bias, i.e. lead to a decrease of the capacity. Nevertheless, Equation (5.9) can still be used to compute the learned capacity of the module. Notice that the concept of capacity can also be

[5]The gradient of Equation (5.9) can become very large for certain values of its weights g_a and b_a.

applied to $\tanh(W\boldsymbol{\alpha} + \mathbf{b})$, but in this case each mode would have his own capacity, making the importance scores less meaningful.

5.5 Training & Regularization

The training of the attention module and the prediction model is performed in two stages (see Figure 5.5). First, each mode is assigned a separate autoencoder, which is trained on the mode to learn the potential energy function. Once trained, the weights of the autoencoders are frozen. In the second phase, EMMA is inserted in front of the MMN and is trained end-to-end on both normal and corrupted data. By corrupted data, we mean samples on which a corruption process is applied in order to simulate one or more failing modes. Notably, the computational overload induced by the training of EMMA in the second stage is often negligible with respect to the MMN, provided that the number of parameters of EMMA[6] is in most cases far less than the number of parameters of the MMN.

Additionally, two regularizers are introduced in the loss function, written as

$$\tilde{\mathcal{L}} = \mathcal{L}(y, \hat{y}) + \lambda_c(g_a - b_a) - \lambda_e\Omega \quad \text{with} \quad \Omega = \sum_{k=1}^{M} \xi_k \log(\alpha_k) \quad \text{and} \quad \xi_k = \begin{cases} \xi_- = -1 & \text{if } \mathbf{x}_k \text{ is corrupted} \\ \xi_+ = +1 & \text{otherwise} \end{cases} \tag{5.10}$$

with λ_c and λ_e being positive real numbers used to set the relative importance of each regularizer, and the ξ_k vector being manually encoded. The first regularizer minimizes the capacity, where a higher λ_c pushes the module to let less information pass through in general (i.e., to be more "cautious"). The second regularizer ($\lambda_e\Omega$), which we call the energy regularizer, controls the trade-off between on the one hand the coupling and on the other hand the failure intensity. In use-cases with complex asymmetric interactions between the modes, the shared energies could potentially cause large discrepancies between modal energies E_i and their original potential energies Ψ_i. A major drawback of having such discrepancies is that this leads to a reduction of the influence of the failure intensity in the computation of the attention scores. This may result in poor generalization to samples with more intensive failing modes. We show that these discrepancies can be reduced to a certain degree by the use of the energy regularizer. Indeed, as an effect of this regularizer, modal energies with low/high potential energies (uncorrupted/corrupted modes) will be decreased/increased. Although the energy regularizer is relatively straightforward, we will demonstrate below that some care needs to be taken regarding the corruption process.

Energy regularization

Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_M\}$ be the set of all the parameters of the second step[7] of the attention module, where $\boldsymbol{\theta}_i = \{[\gamma_{ij}, w_{ij}]_{j=1}^{M}, w_i, b_i\}$ are the parameters composing the modal energy E_i. The effect of the energy regularizer in the SGD algorithm is isolated and written

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \epsilon\lambda_e \nabla_{\boldsymbol{\theta}}\Omega \tag{5.11}$$

[6] The number of parameters of EMMA scales up quadratically with the number of modes.
[7] See Section 5.2

Remember that the objective is to update the parameters such that for low/high potential energies Ψ_i the modal energies E_i are decreased/increased. To verify this let us compute[8] $\nabla_{\boldsymbol{\theta}}\Omega$,

$$\nabla_{\boldsymbol{\theta}}\Omega = \sum_{k=1}^{M} \xi_k \nabla_{\boldsymbol{\theta}} \log(\alpha_k) \tag{5.12}$$

The gradient of the logarithm can be developed as

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

We go further by expressing the equation above with respect to the subset of parameters $\boldsymbol{\theta}_i$:

$$\nabla_{\boldsymbol{\theta}_i} \log(\alpha_k) = \begin{cases} -\rho(1-\alpha_i)\nabla_{\boldsymbol{\theta}_i} E_i, & \text{if } i = k \\ \rho \alpha_i \nabla_{\boldsymbol{\theta}_i} E_i, & \text{if } i \neq k \end{cases} \tag{5.13}$$

The gradient of the regularizer can now be computed by plugging Equation (5.13) into the summation (5.12). Let M' be the number of uncorrupted modes. We obtain for an uncorrupted mode i,

$$\nabla_{\boldsymbol{\theta}_i}\Omega = \xi_+ \big[-\rho(1-\alpha_i)\nabla_{\boldsymbol{\theta}_i} E_i \big] + \big[(M'-1)\xi_+ + (M-M')\xi_-\big]\alpha_i \rho \nabla_{\boldsymbol{\theta}_i} E_i \tag{5.14}$$

and for a corrupted mode i,

$$\nabla_{\boldsymbol{\theta}_i}\Omega = \xi_- \big[-\rho(1-\alpha_i)\nabla_{\boldsymbol{\theta}_i} E_i \big] + \big[M'\xi_+ + (M-M'-1)\xi_-\big]\alpha_i \rho \nabla_{\boldsymbol{\theta}_i} E_i \tag{5.15}$$

We can summarize Equations (5.14) and (5.15) as

$$\boxed{\nabla_{\boldsymbol{\theta}_i}\Omega = -\big[(M-2M')\alpha_i + \xi_i\big]\rho \nabla_{\boldsymbol{\theta}_i} E_i} \tag{5.16}$$

[8] The batch is assumed to only contain one sample for the sake of simplicity. However, the demonstration can be generalized to any batch size.

Adding the constraint that $M' = \lfloor \frac{M+1}{2} \rfloor$, two cases can be distinguished. If the total number of modes M is even, then we have

$$\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \epsilon \lambda_e \rho \xi_i \nabla_{\boldsymbol{\theta}_i} E_i \quad \text{with} \quad \lambda_e \in \mathbb{R}^+ \tag{5.17}$$

The value of the modal energy function for the updated parameters can be developed with a first-order Taylor series approximation around the prior-to-update parameter set $\boldsymbol{\theta}_i^{(0)}$, with a fixed input $\mathbf{x}_i$:

$$E_i\big(\mathbf{x}_i; \boldsymbol{\theta}_i\big) \approx E_i\big(\mathbf{x}_i; \boldsymbol{\theta}_i^{(0)}\big) + \big(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{(0)}\big)^T \nabla_{\boldsymbol{\theta}_i} E_i \tag{5.18}$$

Substituting the updated parameters (5.17) into our approximation, we obtain

$$E_i\big(\mathbf{x}_i; \boldsymbol{\theta}_i^{(0)} - \epsilon \lambda_e \rho \xi_i \nabla_{\boldsymbol{\theta}_i} E_i\big) \approx E_i\big(\mathbf{x}_i; \boldsymbol{\theta}_i^{(0)}\big) - \epsilon \lambda_e \rho \xi_i \big(\nabla_{\boldsymbol{\theta}_i} E_i\big)^T \nabla_{\boldsymbol{\theta}_i} E_i \tag{5.19}$$

where the left-hand side corresponds to the energy value with the updated parameters, and the first term of the right-hand side corresponds to the energy value before the parameter update. Using the fact that ξ_i is negative (resp. positive) for a corrupted (resp. uncorrupted mode), it can be concluded from the equation above that the regularizer will update the parameters such that the values of the modal energy function E_i increases (resp. decreases) for a same sample of the corrupted (resp. uncorrupted mode) i.

In analogy, if M is odd we have

$$\boldsymbol{\theta}_i \leftarrow \begin{cases} \boldsymbol{\theta}_i - \epsilon \lambda_e \rho (1 - \alpha_i) \nabla_{\boldsymbol{\theta}_i} E_i, & \text{if } i \text{ is uncorrupted} \\ \boldsymbol{\theta}_i + \epsilon \lambda_e \rho (1 + \alpha_i) \nabla_{\boldsymbol{\theta}_i} E_i & \text{otherwise} \end{cases} \tag{5.20}$$

The principle is the same as in the even case with an additional effect: the correction will be proportional to the error. To put it in another way, high energies that must be low and low energies that have to be high will have stronger gradients than their counterparts. This is similar to the positive and negative phase in the optimization of Restricted Boltzmann Machines.

To conclude, let us notice that some undesired effects can appear if we do not add the constraint $M' = \lfloor \frac{M+1}{2} \rfloor$. As an illustration, take $M' = \lfloor \frac{M+1}{2} \rfloor + 1$, Equation (5.11) becomes

$$\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \epsilon \lambda_e \rho (\alpha_i + \xi_i) \nabla_{\boldsymbol{\theta}_i} E_i \tag{5.21}$$

which is unstable for uncorrupted modes leading to a collapse where all energies tend to decrease.

5.6 Advantages

The key advantages of using EMMA are:

- The generic design of EMMA permits it to be easily added to any type of architecture of a multi-modal model, without modifying nor EMMA nor the MMN.
- The burden on the MMN is reduced, it only has to learn to make good predictions from the received information. The MMN does not need anymore to learn to distinguish failing modes.

- Our attention module improves the interpretability of the overall model in two ways. First, it can be verified on a per-sample basis which modes are failing and important. Secondly, the *total energy*, $\sum_i E_i$, provides us with an approximate measure of the uncertainty on the predictions (see Section 6.3.2). A useful concrete application, would be to use these interpretable clues to trigger specific hardware/software recovery systems (e.g., luminosity calibration of camera in self-driving cars).

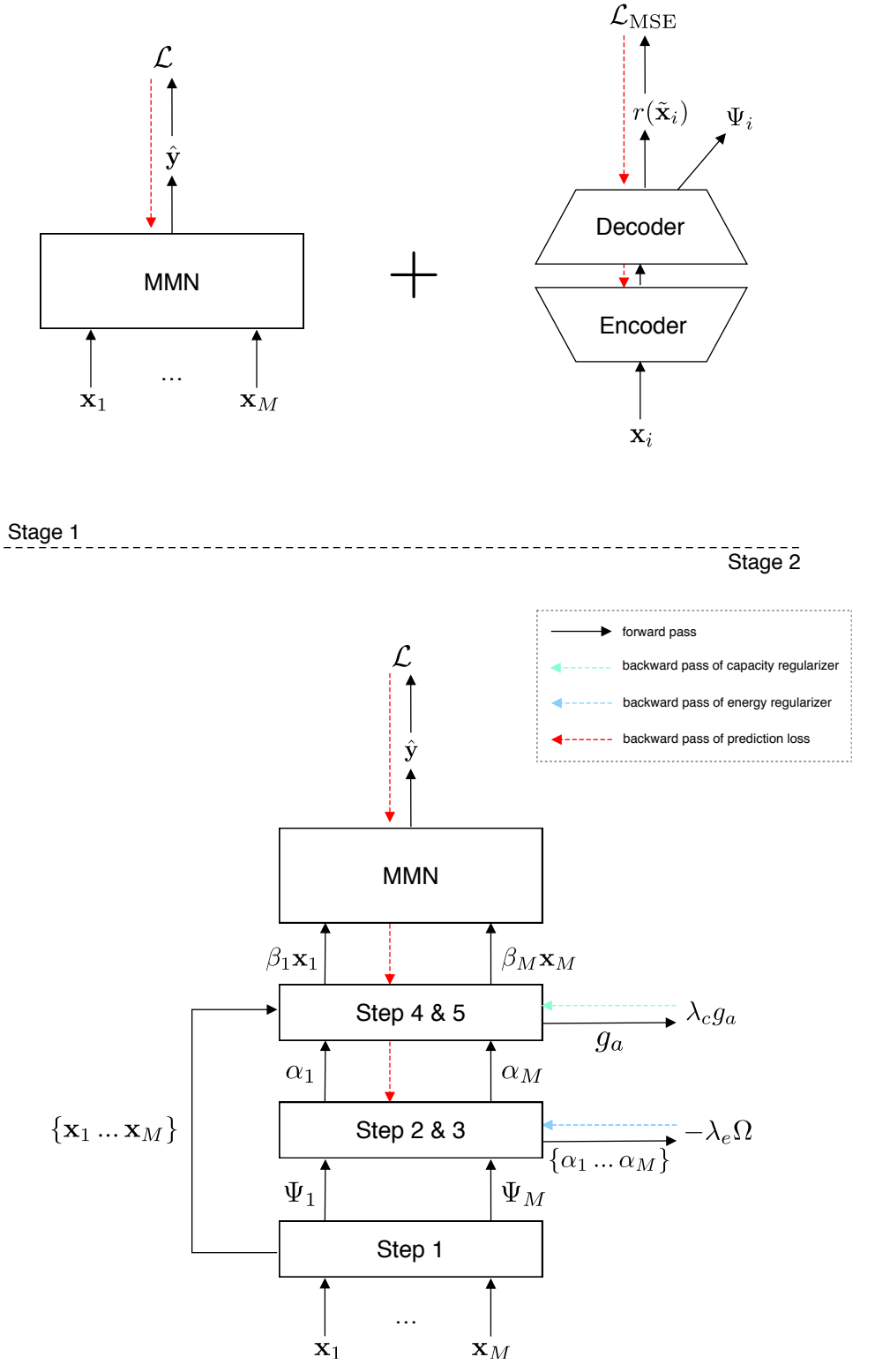


FIGURE 5.5: Summary of end-to-end training.

Chapter 6

Experiments & Results

This chapter experimentally verifies the main ideas discussed in the previous chapters 4 and 5. In chapter 4, the derivation of the potential energy from an autoencoder was shown, with the potential energy being proportional to the log-likelihood. We claimed that this metric could be used to distinguish normal samples from failing ones. On the other hand, chapter 5 discussed the design of the attention module constructed to change the allocation of attention in the face of varying levels of perturbations, and thus improving the robustness of the MMN.

6.1 Pulsar Stars

For the following experiments, our models will be trained on the HTRU2 dataset[1], containing features describing radio emissions measured on Earth. The positive class corresponds to radio emissions of pulsars, which are a rare type of Neutron star that produce radio emissions detectable here on Earth, whereas the negative class corresponds to the other radio emissions. Pulsars are of considerable scientific interest as probes of gravitational theories and time-keeping systems in spacecraft. A short summary of the seminal work of (Lyon, 2016) on this subject can be found in Appendix A.

The models will learn to distinguish pulsars from other radio emissions. The dataset consists of two modes:

- *integrated profile* (IP): each pulsar produces a unique pattern of pulse emissions. The integrated profile is an average of these patterns over many thousands of rotations (further details in Appendix A). The mode contains four features, namely the mean, standard deviation, excess kurtosis[2] and skewness of the integrated profile.
- *dispersion measure* (DM): the amount of dispersive smearing[3] a signal receives is proportional to a quantity called the dispersion measure, which is the integrated column density of free electrons between an observer and a pulsar (further details in Appendix A). Similarly, the mode contains four features, namely the mean, standard deviation, excess kurtosis and skewness of the dispersion measure.

An additional difficulty of this dataset is that it is skewed, there are approximately ten time less positive samples than negative ones (17.898 total samples, 1.639 positive

[1]The dataset can be found here, and was collected by (Keith et al., 2010)

[2]Kurtosis refers to the size of the tails of a distribution. Excess kurtosis is a measure of how prone the distribution is to extreme outcomes.

[3]See Appendix A.

	positive	negative	total
train	1.098	10.894	11.992
validation	270	2.683	2.953
test	271	2.682	2.953
total	1.639	16.259	17.898

TABLE 6.1: Number of samples per split/class of the pulsar dataset.

samples, 16.259 negative samples). This issue was handled by imposing a penalty in the loss function to fight the class imbalance (explained below).

In the experiments, the dataset will be split into training, validation and test sets (see Table 6.1). The validation set is used for tuning purposes, and to implement an Early Stopping[4] algorithm. Usually, the model is then retrained for a few iterations on both the combined training and validation set. In the case at hand, for the sake of simplicity, the model is not retrained on the combined sets. Prior to the training, the data is standardized for the purpose of having the same noise-to-signal ratio corruption effect on all features, since all their variances are equal to one. To avoid information leakage, the statistics (i.e., mean and variance) for the standardization are computed from the training set instead of the entire dataset, and applied to the three data subsets.

6.2 Experiment I

To verify whether the potential energy is a good measure of failure intensity, one autoencoder per mode was trained on the training set. Next, these autoencoders were evaluated on the test set, which partially contained noisy or out-of-distribution samples. The objective is to assess whether a clear difference in the potential energy values is recorded for the failing modes compared to that of regular ones. Notably, missing values are implicitly solved by replacing them by zeros[5]. With this in mind, we did not evaluate missing modes.

Noisy values

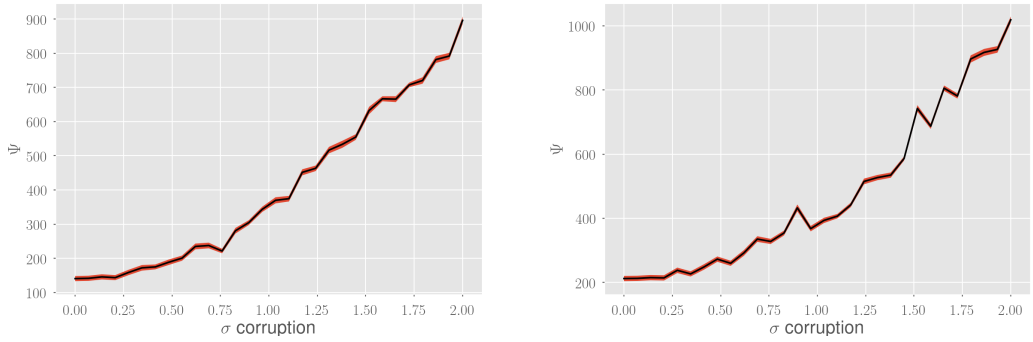
To simulate noise, Gaussian white noise $\mathcal{N} \sim (0, \sigma^2_{\text{corruption}})$ is added to the test set. Then, the mean and variance of potential energies obtained for all samples are computed. This process is then repeated for different noise intensities. The results for each mode are shown in Figure 6.1, which confirms that the potential energy can capture the noise in test data, relative to training samples.

Out-of-distribution samples

Since only two classes are considered in this experiment, the conventional procedure whereby autoencoders are trained on all samples is altered, such that the training set

[4] Early stopping (Prechelt, 1998) is a form of regularization to avoid overfitting, where the error on the validation set is used in determining when overfitting has begun i.e., when the validation error starts increasing.

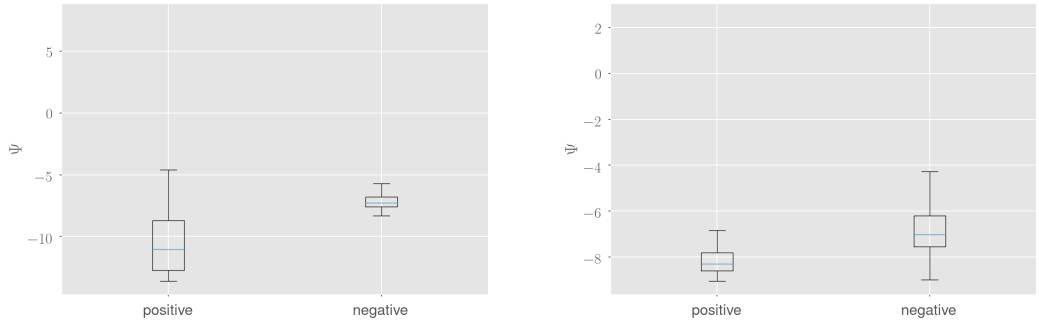
[5] $\beta \cdot 0 = 0, \forall \beta \in \mathbb{R}$



(A) IP mode (B) DM mode

FIGURE 6.1: Potential energy measured on noisy test samples (the mean corresponds to the black line, whereas the interval of two times the standard deviation error is displayed in red).

only includes positive samples. In this case, negative samples can therefore be considered as being out-of-distribution. As can be seen from Figure 6.2, the potential energies computed on the test set (containing both positive and negative samples) allow to distinguish fairly well between positive and negative samples. Indeed, in most cases, the potential energies of likely data (positive samples) are noticeably lower than that of unlikely data (negative samples). This pattern is consistent with energy-based models described in Section 2.3.



(A) IP mode (B) DM mode

FIGURE 6.2: Potential energy measured on positive and negative samples, derived from autoencoders only trained on positive samples.

6.3 Experiment II

6.3.1 Description

The motivation of this experiment is to show the effect of the attention module on the inputs, and its effectiveness in improving the robustness of the MMN. To this end, failing modes are simulated by adding white noise. The predictions of the model fitted with EMMA are compared with the predictions of models having the same architecture but without EMMA. The reason only noisy failing types are tested is two-fold. First, conventional models are generally not able to spontaneously learn to detect out-of-distribution samples. Secondly, it is straightforward to control the intensity of the failing mode by changing the noise-to-signal ratio.

To asses the extent to which fitting EMMA to a standard MMN can improve its predictive performance and robustness, the performance of the coupled models is compared with that of i) a stand-alone MMN ii) a stand-alone MMN trained via a standard data augmentation technique, which consists in adding noisy samples to the training set in the hope that the MMN learns to suppress the noise by itself. In summary, three types of models are evaluated:

- `base-model` is the MMN optimized on the training set (stage 1). Thus, without added noise neither using EMMA.
- `model-without` is the model initialized with the weights of the `base-model`, and finetuned on a mix of corrupted and uncorrupted samples of the training set (data augmentation technique).
- `model-with` is the combined `base-model` with the attention module EMMA. This model is trained end-to-end on a mix of corrupted and uncorrupted data (stage 2).

The corruption process is applied as follows to each data subset separately: 50% of the samples stay uncorrupted, 25% of the samples are corrupted only on the IP mode and for the remaining samples only the DM mode is corrupted. In particular, the samples are corrupted by adding Gaussian white noise with a $\sigma_{\text{corruption}} = 0.5$.

The chosen loss function is the binary cross-entropy with an imbalance penalty[6]. Moreover, the three types of models are trained for 30 epochs with Early Stopping, implemented in the following way: at each epoch the state of the current model is saved as the new optimal one if the validation error is lower than the previous minimum. Furthermore, the optimal classification threshold is determined on the validation set using the receiver operating characteristic (ROC) curve. Lastly, the F1-score is used as the criterion to evaluate the performance on the test set.

6.3.2 Results

Table 6.2 shows the F1-scores of the three types of models on both corrupted and uncorrupted samples of the test set. For the `model-with`, 125 combinations of hyperparameters were trained, from which the fifteen best ones were extracted and shown in Table 6.2. As expected, the `base-model` performs badly on samples with a noisy mode, since

[6]The imbalance penalty weights the loss of each sample with respect to the proportion of its corresponding class in the dataset

	Hyperparameters			F1-score		
	ρ	λ_c	λ_e	uncorrupted	IP noisy	DM noisy
base				0.8830	0.6441	0.6569
without				0.8671	0.7097	0.7683
with	10^{-4}	10^{-3}	10^{-2}	**0.8881**	**0.7333**	**0.8077**
with	10^{-4}	0	10^{-2}	**0.8849**	**0.7285**	**0.8183**
with	10^{-4}	10^{-4}	10^{-2}	**0.8945**	**0.7333**	**0.8182**
with	10^{-3}	10^{-3}	0	0.8809	**0.7347**	**0.8186**
with	10^{-4}	10^{-2}	10^{-3}	0.8736	**0.7383**	**0.7848**
with	10^{-1}	10^{-2}	0	0.8826	**0.7467**	**0.7925**
with	10^{-4}	10^{-3}	0	0.8786	**0.7190**	**0.7826**
with	10^{-3}	10^{-1}	10^{-2}	0.8800	**0.7432**	**0.8344**
with	10^{-4}	0	10^{-4}	0.8723	0.7051	**0.7853**
with	10^{-4}	10^{-4}	10^{-3}	0.8794	0.7053	**0.7853**
with	10^{-3}	10^{-3}	10^{-4}	0.8641	**0.7347**	**0.8129**
with	1	10^{-1}	10^{-1}	0.8683	**0.7237**	**0.8052**
with	10^{-3}	10^{-2}	10^{-4}	0.8705	**0.7105**	**0.8205**
with	10^{-4}	0	10^{-3}	0.8693	0.7051	**0.7901**
with	10^{-3}	10^{-1}	0	0.8817	0.7049	**0.8129**

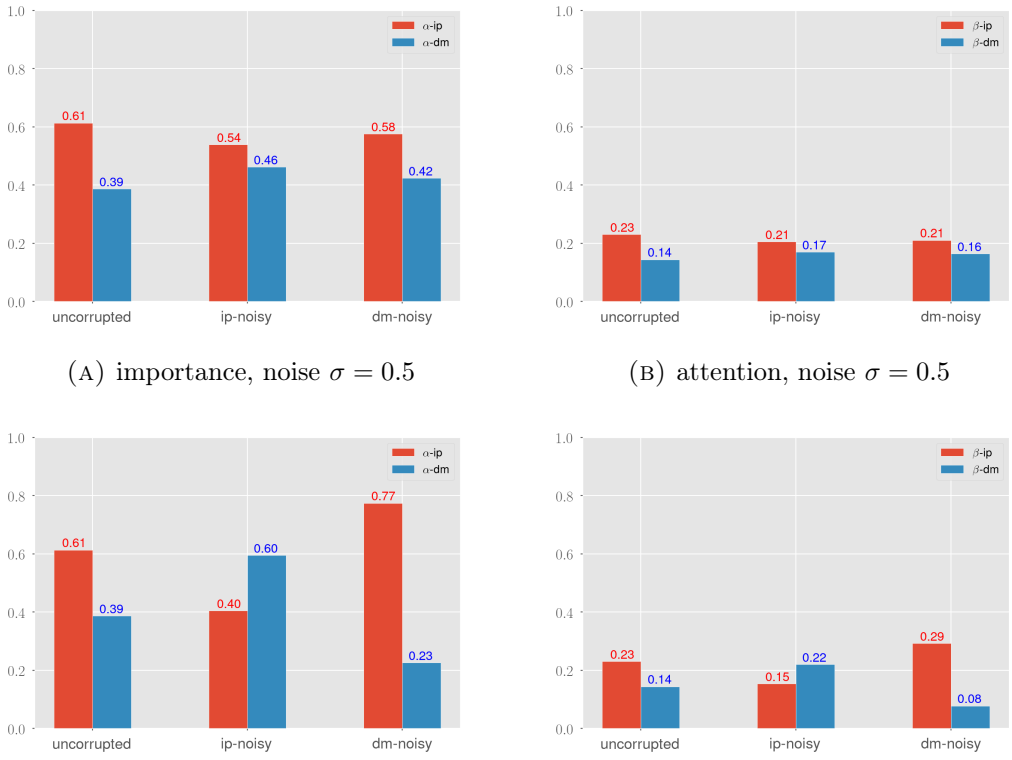
TABLE 6.2: F1-scores of the top-15 trained `model-with` (the models are ranked by their weighted average F1-score) along with `base-model` and `model-without`.

it was only trained on uncorrupted data. Furthermore, the results of `model-without` indicate that using data augmentation indeed improves the performance on noisy data, whereas the performance on uncorrupted samples slightly decreases. Moreover, it can be seen that the models equipped with the attention module (`model-with`) consistently outperform `model-without` and `base-model` on noisy data by a significant margin. Somewhat surprisingly, using the attention module appears to slightly improve the F1-score on uncorrupted data as well compared to the `base-model`. A possible explanation could be that EMMA is able to capture a certain spectrum of quality about the data, which would then be used by the MMN to improve the predictions. Nevertheless, further investigation is required to validate this hypothesis. Lastly, it is worth noticing that the models seems to be more robust against one failing mode (DM) then the other (IP). Since the noise-to-signal ratios are uniform across modes, it could be hypothesized that the latter mode is more informative.

Attention allocation and shifts

A key feature of EMMA is to allocate its attention toward the most important modes. The validity of this claim can be ascertained by monitoring the change in allocation

under various circumstances. As a reminder, the importance and attention scores introduced in Section 5.1 can be leveraged allowing us to evaluate the attention change. These scores are reported in Figure 6.3 for two levels of noise, and clear changes in allocation can be observed as a result of the application of noise onto the different modes. Interestingly, it can be seen in Figure 6.3a that for low levels of noise applied to the IP mode, it still receives more attention than the DM mode. This is consistent with the aforementioned observation that the IP-mode seems to be more informative. In contrast, if the noise level on this mode is sufficiently increased (Figure 6.3c), an attention shift occurs and the other mode becomes more important. A more complete view of the attention scores for varying levels of noise applied to different modes is given in Figure 6.4.



(A) importance, noise $\sigma = 0.5$

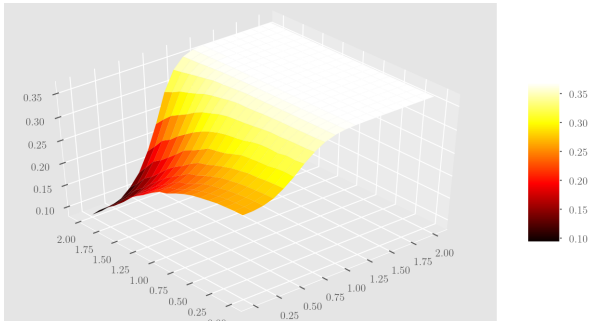
(B) attention, noise $\sigma = 0.5$

(C) importance, noise $\sigma = 2$

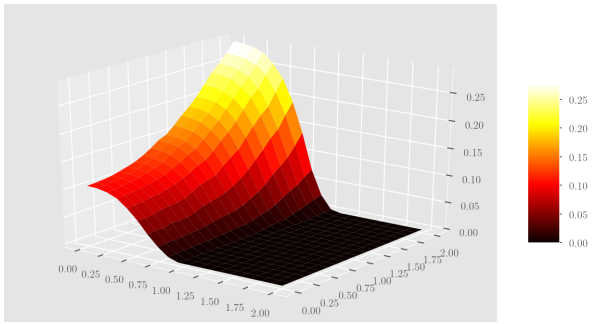
(D) attention, noise $\sigma = 2$

FIGURE 6.3: Importance and attention scores for the 1st ranked `model-with` ($\rho = 10^{-4}$, $\lambda_e = 10^{-3}$, $\lambda_c = 10^{-2}$), on two different levels of noises ($\sigma = 0.5$ and $\sigma = 2$).

To gain a better understanding of the attention allocation, it is worth further investigating the role of the coldness hyperparameter (ρ, see Section 5.3). A simple way of doing so consists in comparing the importance scores of models trained with significantly different temperatures as displayed in Figure 6.5a and 6.5c. As expected, the values of the importance scores are more uniformly distributed for the one with a high temperature (Figure 6.5a), permitting the module to stay more stable for higher levels of noise (Figure 6.5b). Indeed, it was proved that an increase of the overall perturbation level makes the module more selective (see the *energy threshold* explained in Section 5.4) On the contrary, the model trained with a lower temperature has more pronounced attention shifts (Figure 6.5c). As a result, the corrupted mode will be masked very rapidly, and discarded altogether for high levels of noise (seen in Figure 6.5d). Another



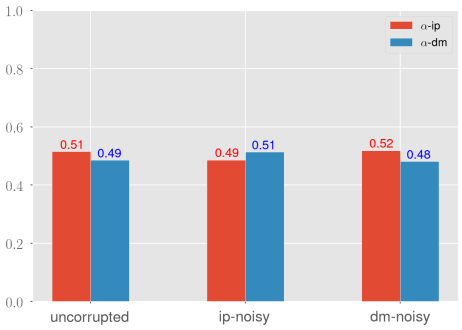
(A) β-ip



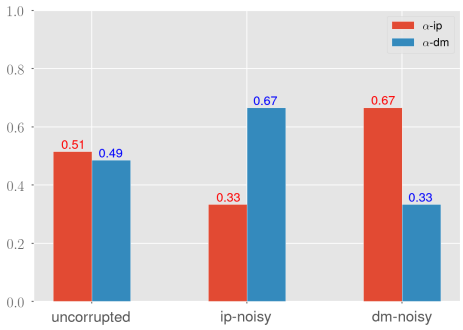
(B) β-dm

FIGURE 6.4: 3D visualization of attention scores on varying levels of noises between the modes of the same model as Figure 6.3. The axis in the horizontal plane represent the variance of noise on each modes, whereas the vertical axis corresponds to the attention score.

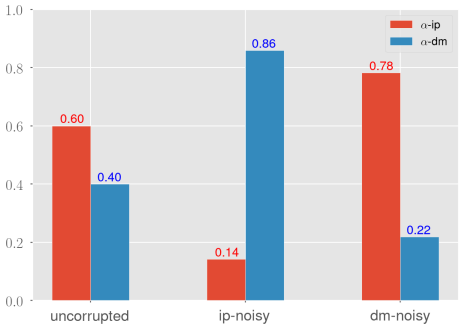
view of the same example is given in Figure 6.6, where instead the attention scores are visualised on a continuum of noise intensities. It can be seen that the model with the high temperature (Figure 6.6a and 6.6b) is able to learn a richer attention profile than the one with the low temperature (Figure 6.6c and 6.6d). Indeed, the crossing-point in Figure 6.6a corresponding to the attention shift, may be interpreted as the trade-off between the relevance and failure intensity. In contrast, the other module having a low temperature is naturally more biased towards the uncorrupted mode, limiting the training process and making it more difficult to learn the interactions between the modes. These results also visually confirm the utility of multiplying the modes by attention scores instead of importance scores. Indeed, in Figure 6.6c and 6.6d the module learns to allocate a high attention to both modes for low levels of noise and is still able to mask out the corrupted mode if the noise-to-signal ratio increases. This behaviour can not be modelled if we multiply the modes by the attention scores, since they sum up to one.



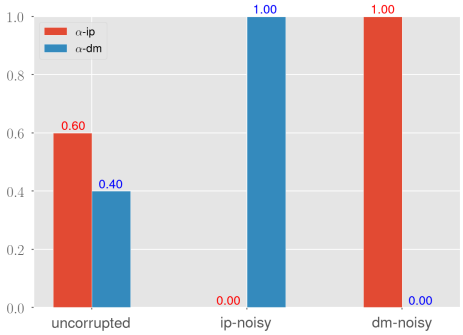
(A) importance, $\rho = 10^{-4}$, noise $\sigma = 0.5$



(B) importance, $\rho = 10^{-4}$, noise $\sigma = 2$

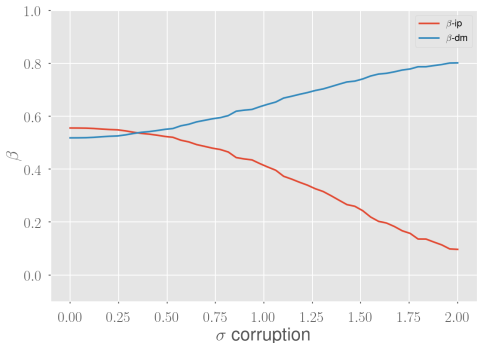


(C) importance, $\rho = 10^{-1}$, noise $\sigma = 0.5$

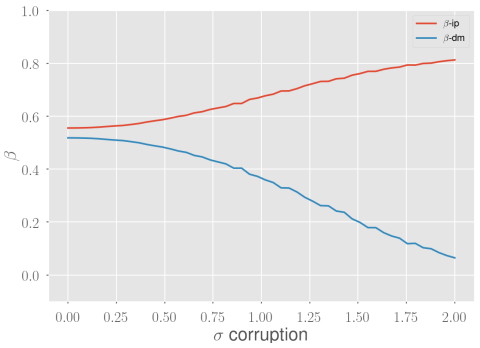


(D) importance, $\rho = 10^{-1}$, noise $\sigma = 2$

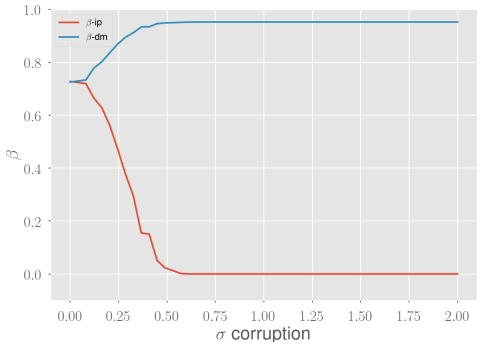
FIGURE 6.5: Importance scores comparison of models with different temperatures: `model-with` ($\rho = 10^{-4}$, $\lambda_e = 10^{-2}$, $\lambda_c = 10^{-3}$) and `model-with` ($\rho = 10^{-1}$, $\lambda_e = 10^{-2}$, $\lambda_c = 10^{-3}$).



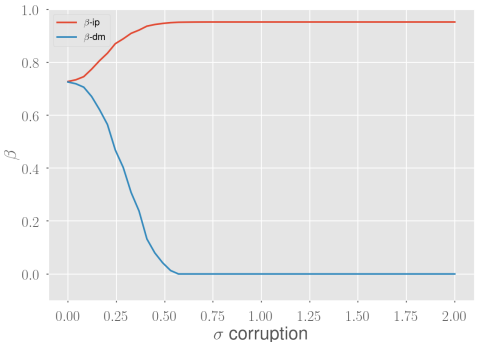
(A) attention, $\rho = 10^{-4}$, noise on IP



(B) attention, $\rho = 10^{-4}$, noise on DM



(C) attention, $\rho = 10^{-1}$, noise on IP



(D) attention, $\rho = 10^{-1}$, noise on DM

FIGURE 6.6: Attention scores comparison of models with different temperatures: `model-with` ($\rho = 10^{-4}$, $\lambda_e = 10^{-2}$, $\lambda_c = 10^{-3}$) and `model-with` ($\rho = 10^{-1}$, $\lambda_e = 10^{-2}$, $\lambda_c = 10^{-3}$). The learned capacity[7] for these models are respectively 0.49 and 0.63.

Total Energy

As mentioned in Chapter 5, one of the key advantages of the attention module is to facilitate/promote results interpretability. In particular, a quantity named the total energy, which can be computed as the sum of all modal energies of a given sample, provides readily interpretable clues pertaining to the workings of the model. By construction, the total energy captures the overall amount of perturbation in a given sample, and can thus serve as a good proxy for quantifying the uncertainty on the predictions at test time.

In order to evaluate the veracity of this claim, the following experiment is carried out. Firstly, both modes are corrupted in the test set. Then, the total energy and F1-score are computed for each sample, and averaged over the entire test set. This process is then repeated for increasing noise intensities. The results are shown in Figure 6.7. A significant correlation can indeed be observed between the total energy and the F1-score. Remarkably, this pattern is observed for all the modules, independently of the chosen set of hyperparameters. Hence, the total energy could indeed be used as a proxy for uncertainty on the predictions.

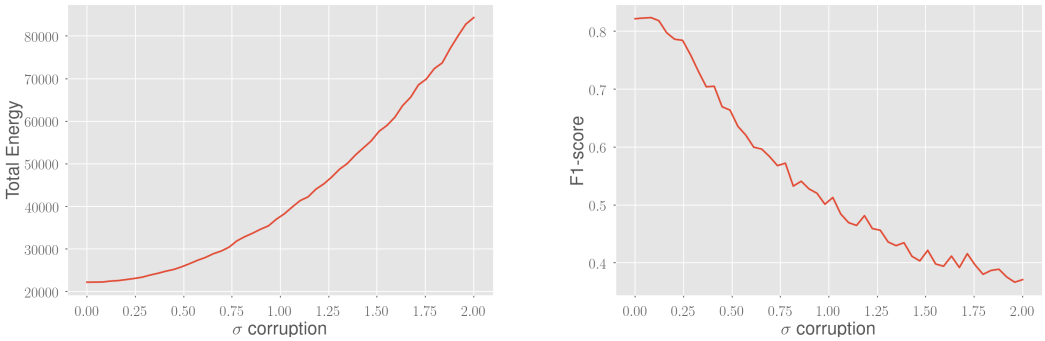


FIGURE 6.7: Total energy for `model-with` ($\rho = 10^{-4}$, $\lambda_e = 10^{-3}$, $\lambda_c = 10^{-2}$).

Robustness Generalisation

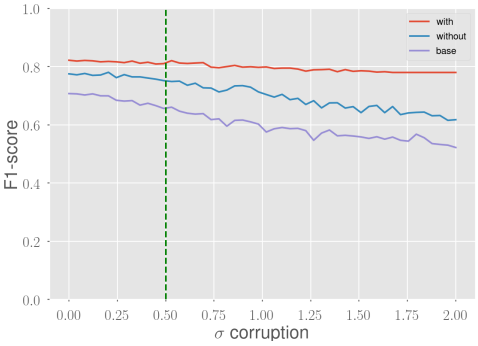
As discussed in Chapter 3, the robustness of a model is often evaluated under the same conditions as the training. This approach appears somewhat questionable. Indeed, the conditions under which the model will be deployed cannot, in most cases, be specified exactly in advance. In particular, the nature and occurrence of failing modes may be unknown *a priori*. As a result, it appears very likely that such an evaluation method provides an overly optimistic assessment of model robustness.

In this study, in order to carry out a more realistic robustness assessment of the combined model (`model-with`), the quality of the predictions is evaluated on test sets with varying levels of noises. In particular, *only one mode* of the test set is corrupted with a specific intensity of noise on which the model is evaluated. This process is then repeated for increasing noise intensity. The results of the robustness analysis of the 1st ranked model are shown In Figure 6.8. When the noise is applied on the DM-mode (Figure 6.8b), he performance of the `model-with` remain more stable than that of the other models. The masking of the noise by the attention module is verified in Figure 6.8d (blue line). Concerning, the results when the other mode (IP) is noisy (Figure

6.8a) are less convincing. However, the performance seems to stabilize for high noise levels (Figure 6.8a, from $\sigma \approx 1.35$). In fact, it can be shown that this high noise level corresponds to the attention switch in Figure 6.8c. This may indicate that the capacity was regularized too strongly, and as result does not let sufficient information pass. To support this claim, results for a module with a higher capacity are displayed in Figure 6.9. It can indeed be observed that the global amount of allocated attention (capacity) is higher, as the sum of the area under the curves of the attention scores is higher. In consequence, the module can instead learn to shift the attention sooner (Figure 6.9c and 6.9d), leading to a more consistent and stable performance (Figure 6.9a compared to Figure 6.8a).

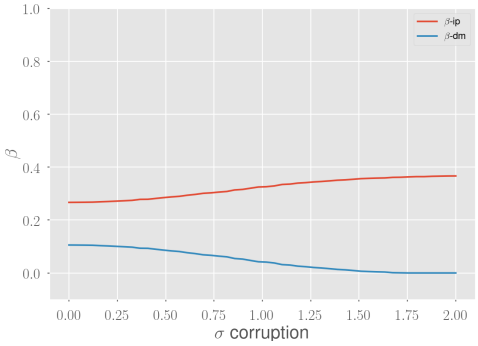
Finally, it must be stated that the capacity regularizer requires careful tuning. Indeed, an extreme case in which the capacity was regularized too heavily is reported in Figure 6.10. Strikingly, the model performance deteriorates significantly, and the stand-alone MMN with data augmentation training outperforms the combined models over a wide range of noise intensities.

(A) F1-score, noisy IP-mode



(B) F1-score, noisy DM-mode

(C) attention score, noisy IP-mode



(D) attention score, noisy DM-mode

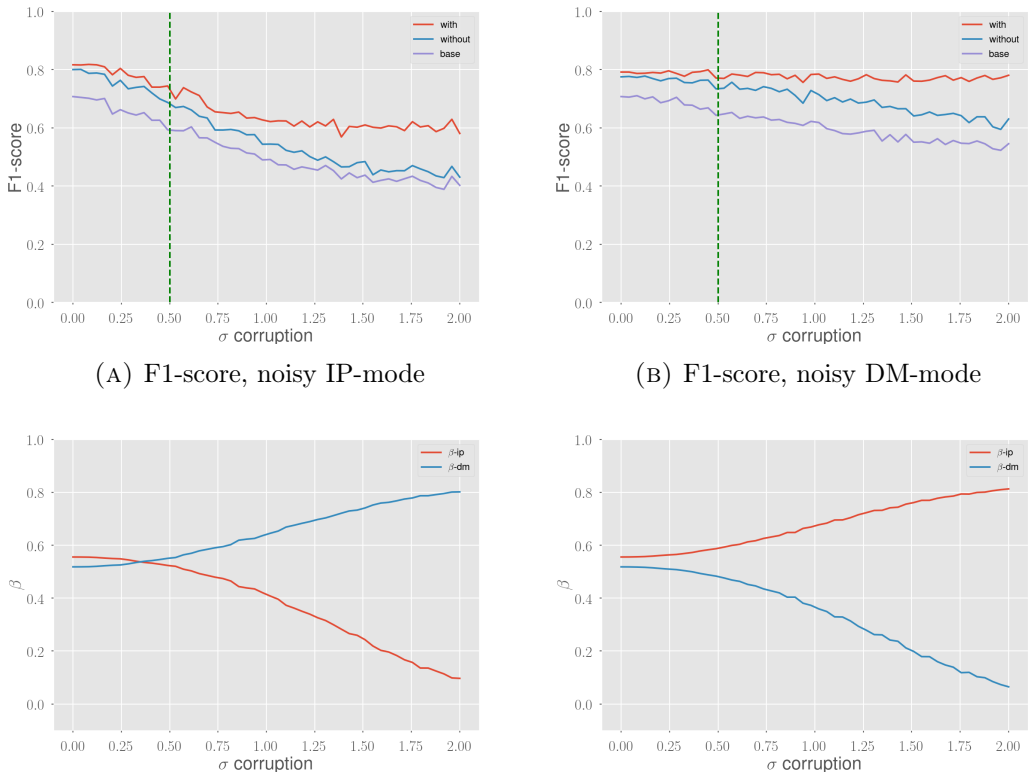
FIGURE 6.8: Noise generalisation of `model-with` ($\rho = 10^{-4}$, $\lambda_e = 10^{-3}$, $\lambda_c = 10^{-2}$). The learned capacity is 0.19.

6.4 Results discussion

In the preceding sections, a set of numerical experiments was carried out in order to evaluate the performance of a novel module (EMMA) specifically designed to tackle the problem of failing modes in multi-modal deep learning approaches.

The experiments made use of a real dataset drawn from the field of astrophysics, and each sample comprised two distinct modes. Encouraging results were observed and reported, as MMN models fitted with the attention module displayed improved predictive performance, robustness, interpretability and consistently outperformed i) a traditional stand-alone MMN ii) a stand-alone MMN trained via a standard data augmentation technique. The influence of custom regularizers and related hyperparameters was also investigated.

The main objective of this work was to develop ideas and a new framework to tackle the problem of failing modes in multi-modal data. In this chapter, we believe to have verified to validity of certain of the ideas outlined in Chapter 5. Indeed, it was shown that EMMA improves the robustness against failing modes by masking them out, these inputs were then fed to an MMN who was then able to make more stable predictions. In addition, the impact of some of the hyperparameters were discussed. However, all these results must be interpreted with caution as no general rigorous analysis was performed on a real-case dataset. Furthermore, interpreting the effect of the hyperparameters is difficult since their influence are not isolated. A limitation of the experiments that can be noticed is that the dataset only contains two modes. As a result, no complex dependencies between the modes had to be learn, minimizing the impact of the shared energies which could thus not be studied.



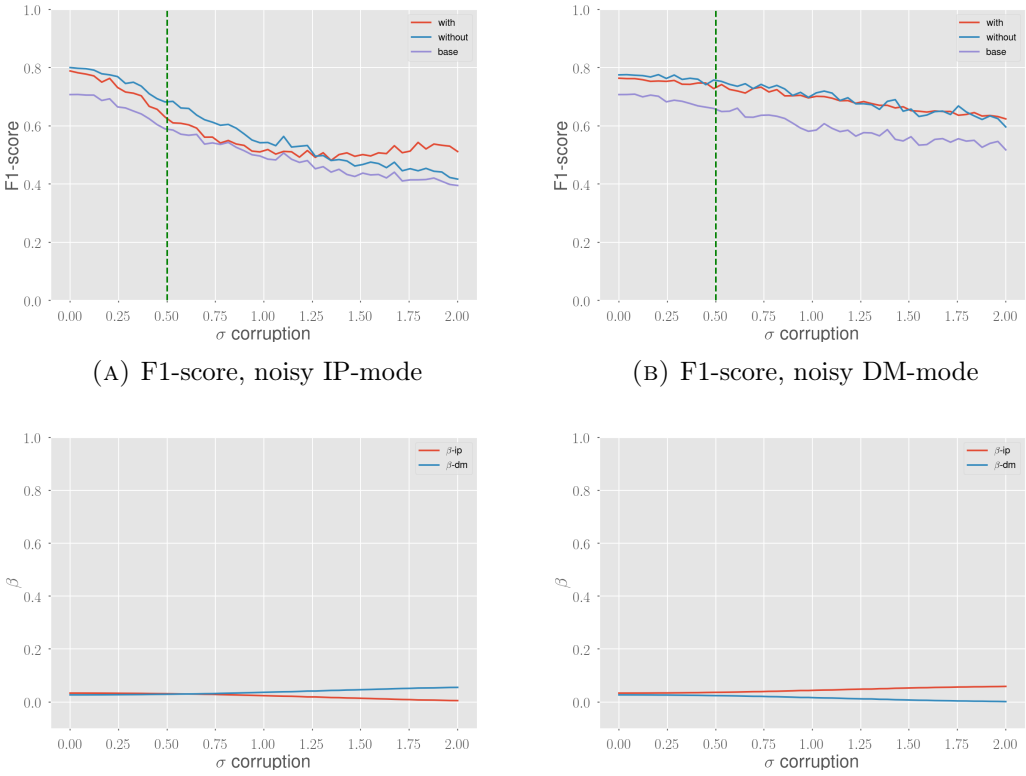
(A) F1-score, noisy IP-mode

(B) F1-score, noisy DM-mode

(C) attention score, noisy IP-mode

(D) attention score, noisy DM-mode

FIGURE 6.9: Noise generalisation of `model-with` ($\rho = 10^{-4}$, $\lambda_e = 10^{-2}$, $\lambda_c = 10^{-3}$). The learned capacity is 0.49.



(A) F1-score, noisy IP-mode

(B) F1-score, noisy DM-mode

(C) attention score, noisy IP-mode

(D) attention score, noisy DM-mode

FIGURE 6.10: Noise generalisation of `model-with` ($\rho = 10^{-4}$, $\lambda_e = 10^{-1}$, $\lambda_c = 10^{-1}$). The learned capacity is 0.04.

Chapter 7

A Unified Model for Multi-Modal Attention

The purpose of using EMMA is to help the multi-modal network (MMN) to handle failing modes, and more generally, to figure out the relative emphases to be placed on the different modes depending on their general contributions to the predictions. The literature review[1] discussed self-attention and cross-modal attention mechanisms, used to highlight information inside a specific mode, such as certain regions in an image or a set of frequencies in a sound. The difference between these two mechanisms is that self-attention only relies on information from the mode itself as a context, whereas cross-modal attention uses information from all the available modes. Now that we have EMMA, we claim to have all the ingredients to construct a complete multi-modal network like humans. As a reminder, human's complete multi-modal attention consists of three different components: exogenous, endogenous and cross-modal attention. The endogenous component of attention stems from a conscious, voluntary process, whereby a human elects to focus on a particular object (Driver and Spence, 1998). In contrast, the exogenous attention is triggered by the sudden onset of an unexpected event, and can thus be viewed as a reaction to an external stimulus (Driver and Spence, 1998). One way of constructing an endogenous module would be as a block of M self-attentions, where each self-attention is dedicated to one specific mode. On the other hand, the attention module developed in this work can be interpreted as reproducing the exogenous attention used by humans to robustly handle abnormal situations.

With this in mind, we present a unified model (see Figure 7.1) combining all the strengths of each type of attention. First, the attention masks β_i computed by the exogenous model, and the attention masks $\mathbf{m}_i$ computed by the endogenous module, are combined to obtain the resulting masks $\beta_i \mathbf{m}_i$. These mask will highlight the most important modes, and on an intra-modal level attend to the most relevant regions. The masks $\beta_i \mathbf{m}_i$ are then applied to the input sample, which is passed through the cross-modal module. Finally, the processed input $\{\mathbf{x}''_1 ... \mathbf{x}''_M\}$ is forwarded to the MMN. In addition, the complete module can further be refined by inserting feedback loops from the previously predicted output to the separate modules, as it is often done in the literature (Afouras et al., 2018; Vaswani et al., 2017; Bahdanau, Cho, and Bengio, 2014). For example, in a self-driving system, the attention module could improve its focus to different regions of the input image depending on the previously detected cars. It is worth mentioning that the proposed architecture is only a generalization of current attention modules such as in (Afouras et al., 2018), supplemented with an exogenous component.

[1] See Chapter 3

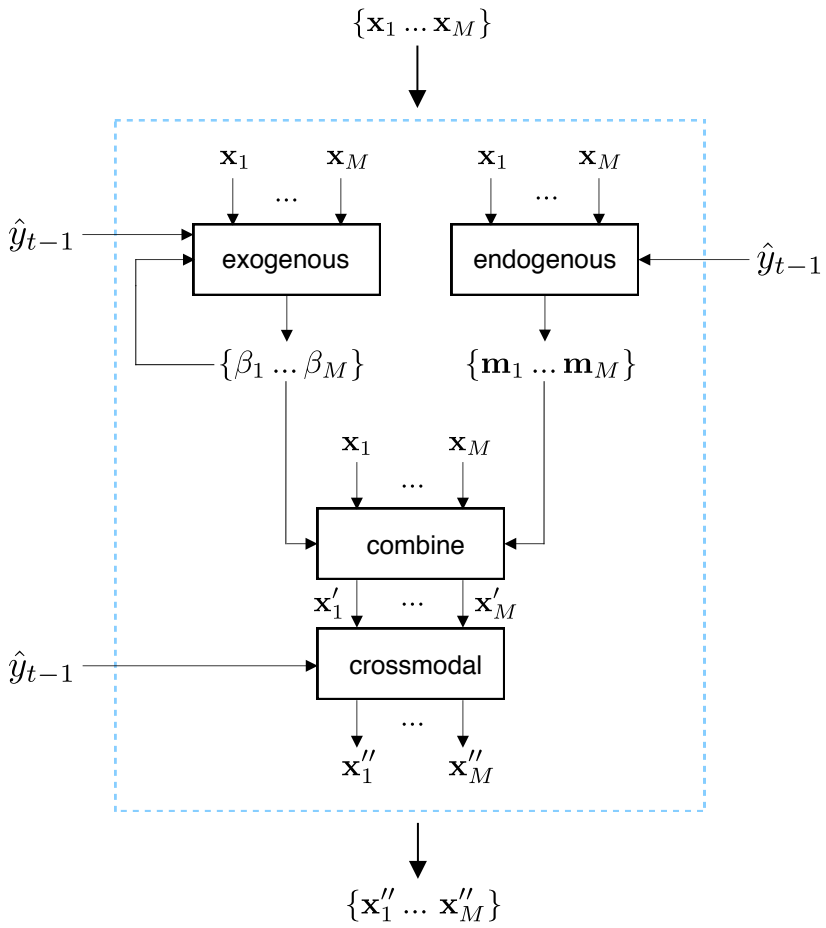


FIGURE 7.1: A possible architecture for a unified multi-modal attention.

Chapter 8

Conclusion

The primary objective of this work was to develop a deep learning module whose task is to pre-process multi-modal inputs to reduce the amount of perturbations. In the experiments, it was indeed shown that a masking of the perturbations actually occurs. As a result, the performance of the prediction model on samples with failing modes was improved. Additionally, we experimentally verified that this pre-processing step enables the performance gain to remain stable on more intensive failing modes. In contrast, models trained only with standard data augmentation experienced a decrease of the performance on modes containing more perturbations than in the training set. Despite these promising results, caution must be taken because we were unable to investigate the performance of the module on more complex datasets. Nevertheless, we believe that the ideas detailed in Chapter 5 are sufficiently general to be considered as a starting framework for the construction of more robust multi-modal neural networks.

Another main insight was to translate the concept of capacity from psychology to deep learning. This led to the idea of a regularizer forcing the module to limit the amount of extracted information from the input. The concept of capacity and its corresponding regularizer could eventually be applied to deep learning models with a limited amount of processing power as in embedded systems.

The last contribution proposed an architecture for a unified multi-modal attention, combining the two main types of attention mechanisms found in deep learning (i.e., self and crossmodal) with our attention module.

8.1 Future work

Our results are encouraging but should be validated on more complex datasets containing images, text, sounds, etc. Several points would need to be addressed such as finding efficient methods to approximate the log-likelihood of those data structures[1]. Another point is on which level to apply the attention masks, on the raw input data or on the features extracted by the encoders?

An idea that came up during this thesis but was not tested, is based on the following observation: the transition between the first and second stage of the training can be quiet "brutal" for the MMN. Indeed, weights of the MMN at the start of the second stage are optimal for uncorrupted modes. However, these weights will have to adapt immediately to weighted inputs, as an effect of EMMA. A smoother approach to consider is inspired from the process of *annealing* in metallurgy; "Annealing is a process in which

[1] Several alternatives were proposed in Chapter 5

a solid is first heated until all particles are randomly arranged in a liquid state, followed by a slow cooling process. At each cooling temperature enough time is spend for the solid to reach thermal equilibrium."[2] Applying this idea to our case, we could divide the weights of the first layer of the MMN by $\tanh(1/M)$ at the start of the second stage, and set the temperature high enough to obtain a uniform distribution of importance scores ($\alpha_i \approx 1/M, \forall i$). As a consequence, the effect of EMMA on the inputs of the MMN would be non-existent[3], keeping the latter in its local minima. Subsequently, a cooldown schedule would be applied to the temperature leading smoothly to more pronounced attention shifts on the input data. This guided approach has no guarantee to improve the results but in our opinion can be worth trying. Moreover, instead of fixing a final temperature[4] by tuning, a similar approach to Early Stopping could be used: the cooldown would be stopped when the validation error stops decreasing significantly.

[2] Explanation from here

[3] This assumption is only guaranteed if the parameters g_a and b_a are intialized as $g_a = 1$ and $b_a = 0$.

[4] Temperature at which the cooling schedule is stopped.

Appendix A

Dataset

This part of the thesis provides a brief overview of what pulsar stars are. It then goes on to explain the two modes of the dataset used for their detection: the integrated profile (IP) and dispersion measure (DM). The most part of this chapter is a direct summary of the background chapter in the doctoral thesis (Lyon, 2016)[1] . Some parts of the text are barely changed from (Lyon, 2016), nevertheless, many details have been voluntary ignored for the sake of simplicity as it is not the focus of this work.

"A star is a luminous ball of gas, mostly hydrogen and helium, held together by its own gravity. The nearest star of Earth is the Sun. Most mass is in the core, at the center of the star. Gravitational forces are by consequence directed inwards. During the majority of a star's life, it will fuse hydrogen to helium, generating an outwards pressure, balancing the gravity (Ghosh, 2007). By the time the hydrogen amounts become insufficient, the star starts to use other elements in the surrounding layers of the core as a fuel. As those elements diminish, the star's energy output drops rapidly, causing gravity to overcome the forces which had previously maintained the stars structure. The core of the star than undergoes a rapid and violent collapse (Ghosh, 2007). The collapse can lead to a number of potential evolutionary outcomes for the leftover core (see Figure A.1), depending on the stars birth mass measured in solar masses ($M_{\odot}$). Intuitively, the heavier the birth mass, the greater the inwards gravitational force are and the harder the collapse. The first outcome applies to low mass stars, which typically become white dwarfs following their collapse. Within white dwarfs, densely packed electrons are able to resist gravitational compression. Our own sun is likely to one day become a white dwarf star. Then there are stars between 8-20 $M_{\odot}$ at birth, electron degeneracy pressure can no longer prevent collapse as in white dwarfs, but they are not massive enough to undergo complete gravitational collapse, preventing the formation of a black hole. Instead the intense conditions within these stars cause electrons to combine with protons forming neutrons who resist against pressure; These stars are called neutron stars. The last evolutionary outcome applies to large stars with masses greater than 20 $M_{\odot}$. These stars can, under the right conditions, undergo complete gravitational collapse. This results in the formation of a black hole singularity otherwise known as a stellar mass black hole."

"A pulsar is a unique form of neutron star that retained most of its angular momentum of their progenitor star during collapse. Complex interactions between the surfaces of pulsars and their strong magnetic fields, helps to produce their defining feature, the emission of radio waves. The radio emission produced by pulsars originates from their magnetospheres (Ghosh, 2007). This is the area of space surrounding a pulsar in which charged particles are influenced by a co-rotating magnetic field, which has both open

[1] (Lyon, 2016) can be accessed here

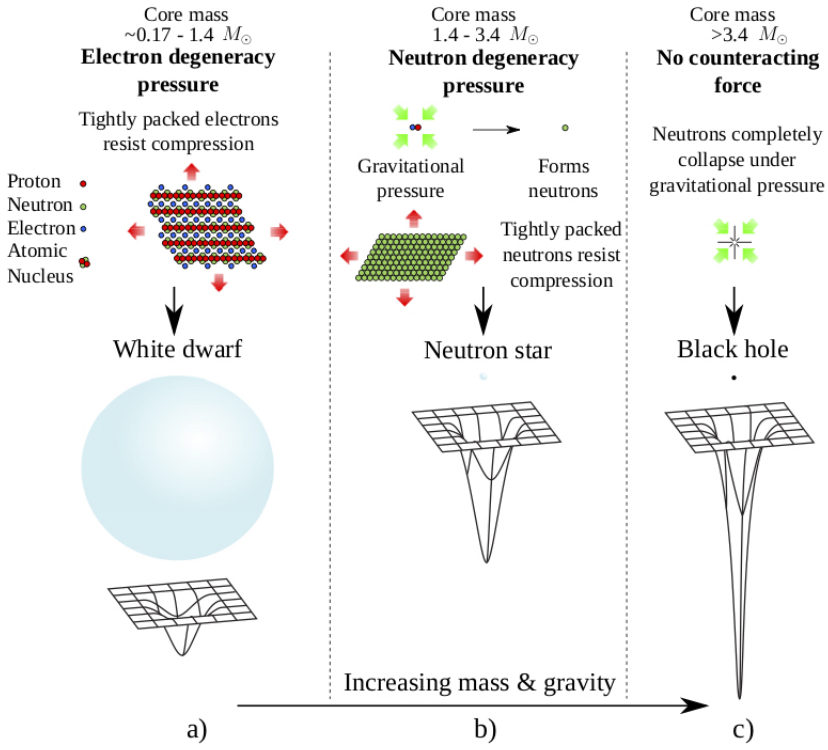


FIGURE A.1: Common evolutionary endpoints for main sequence stars. In a) electron degeneracy pressure prevents gravitational collapse, leading to the formation of a white dwarf star. In b) electron degeneracy pressure is no longer enough to counteract the inward force of gravity, however the gravitational pressure is insufficient to overcome neutron degeneracy pressure, allowing a Neutron star to form. Finally in c) the force of gravity is so great that gravitational collapse cannot be halted, resulting in the formation of a black hole. The depictions of the gravitational sinks above are based on diagrams by (Treat and Stegmaier, 2014). *Image and caption copied from* (Lyon, 2016).

and closed field lines (D.R. and M., 2005) (see Figure A.2). To maintain this co-rotation property, the velocity of the field lines must increase as they move further away from the pulsar. Eventually the distance becomes so great, that to maintain co-rotation, the velocity of the field lines must be greater than or equal to the speed of light c. This is not possible, thus the field lines are unable to close where the required velocity is c. The abstract cylinder aligned with the rotation axis, that synchronously rotates with the pulsar at a velocity c, is known as the light cylinder (see Figure A.2). The particles extracted from the surface are then believed to be accelerated along the co-rotating magnetic field lines of the magnetosphere (Lorimer, 2008), which endows the particles with increased energy. This additional energy causes the particles to emit radiation (Lorimer, 2008) to be emitted along the open field lines near a pulsar's magnetic pole. A pulsar's magnetic axis is usually inclined with respect to its rotational axis. Therefore each time a pulsar rotates, the radiation beam produced near the magnetic poles, is swept at an angle across the sky. If the beam crosses the line of sight of an observer here on Earth, the pulsar becomes detectable as a rise and fall in broadband radio emission. This pattern repeats periodically with each rotation of the pulsar. This is known as the lighthouse model of emission (Lorimer, 2008), because the beam of radiation is analogous to a lighthouse warning light rotating very quickly."

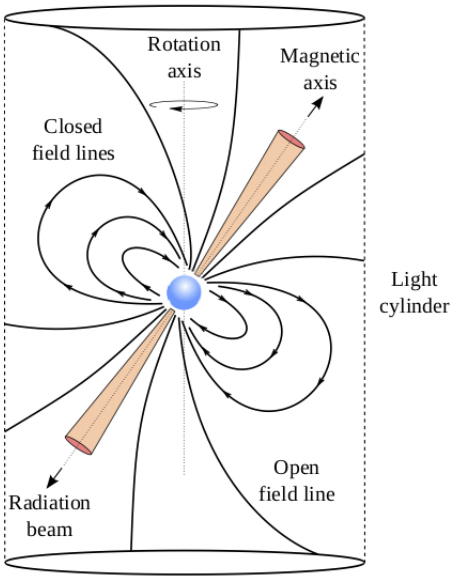


FIGURE A.2: Simplification of the lighthouse model of a radio pulsar, the pulsar is surrounded by a strong magnetic field comprising of open and closed field lines unable to close at the light cylinder. The light cylinder is an imaginary cylinder aligned with the pulsars rotational axis, that synchronously rotates with the pulsar at the speed of light. As the magnetic field cannot rotate at this velocity, the field lines cannot close at the light cylinder leading to open field lines. Radio pulses are emitted from the open field lines at a region near the magnetic poles in the pulsar's magnetosphere. *Image and caption copied from* (Lyon, 2016).

"Each pulsar produces a unique pattern of pulse emission known as its pulse profile (Lorimer, 2008). Two such profiles are shown in Figure A.3. However whilst pulsar rotational periods are extremely consistent, their profiles can deviate from one-period to the next. Whilst such changes in the pulse profile provide clues to what is happening in and around the pulsar, they make pulsars hard to detect. This is because their signals are non-uniform and not entirely stable overtime. However these profiles do become stable, when averaged over many thousands of rotations."

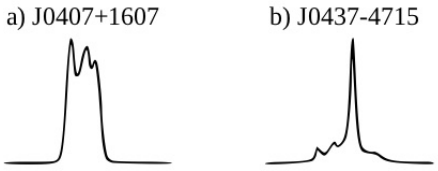


FIGURE A.3: Example pulse profiles of two separate pulsars. These profiles were adapted from those originally presented in (D.R. and M., 2005). *Image and caption copied from* (Lyon, 2016).

"Signals travelling through the interstellar medium (ISM) are affected, the most significant effect is known as dispersion. As pulsar signals travel through the ISM towards the Earth, they interact with charged particles (free electrons) on route. These interactions delay the arrival of the signal here on Earth. The low frequency components of the signal are delayed more than the corresponding high frequency counterparts. This has a dispersive effect that causes pulsar signals to become smeared in time. This

makes it difficult to detect pulsars, as their pulses become less pronounced as shown in Figure A.4. Manifesting itself as a reduction in the signal-to-noise ratio of a detected pulse. The amount of dispersive smearing a signal receives is proportional to a quantity called the dispersion measure (DM) (D.R. and M., 2005). The DM is the integrated column density of free electrons between an observer and a pulsar (Lorimer, 2008). The true column density, and thus the precise degree to which a signal is dispersed, cannot be known a priori. A number of dispersion measure tests or "DM trials", must therefore be conducted to determine this value as accurately as possible. An accurate DM can be used to undo the dispersive smearing, allowing the signal-to-noise ratio of a detected signal to be maximised (D.R. and M., 2005). For a single dispersion trial, each frequency channel is shifted by an appropriate delay. Subsequent trials increment the delay in steps, until a maximum DM is reached. This maximum will vary according to the region of sky being surveyed, the observing frequency, and bandwidth. The process produces one 'de-dispersed' time series per frequency channel. These are then summed to produce a single de-dispersed time series per trial (as shown at the bottom plot of Figure A.4 a). In total de-dispersion produces a number of time series equal to the total number of DM trials."

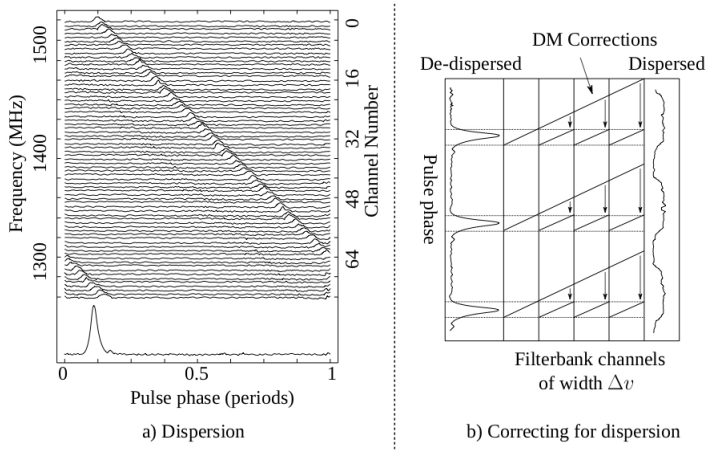


FIGURE A.4: An example of signal dispersion. Based upon diagrams originally presented in (D.R. and M., 2005). Plot a) shows how a signal is dispersed in time. Dispersion hides the true pulse shape and causes a lowering of the detected signal-to-noise. Plot b) shows the application of DM corrections to a dispersed signal. The DM correction is different in each frequency channel, since dispersion is proportional to frequency. *Image and caption copied from* (Lyon, 2016).

"By precisely measuring the timing of such pulses, astronomers can use pulsars for unique experiments at the frontiers of modern physics. Indeed, pulsars exist in strong-field gravitational environments due to their enormous mass. It is impossible to study such environments within Earth-based laboratories, or even within the confines of our own solar system which is lightweight by comparison. In the strong-field environment provided by pulsars, their immense gravitational fields directly affect the arrival times on Earth of the signals they produce, via special and general-relativistic effects. By studying these effects, tests of many gravitational theories can be accomplished. Another application of measuring the arrival time of pulses is that they are effective time keeping system, rivalling atomic clocks for accuracy. Such clocks are useful for spacecraft navigation and timekeeping here on Earth."

Appendix B

Experimental setups

B.1 Experiment of Chapter 4

Each autoencoder was constructed as in 4.1, with $L = 4$ and 12 hidden units. The training details are listed below

- number of epochs: 30
- batch size: 64
- σ noise denoising: 0.01 (to not be confused with corruption process to simulate noisy modes)
- Adam optimizer, with learning rate: 0.001

Appendix C

Miscellaneous

C.1 Integrability criterion

The integrability criterion (Santilli, 1982) is a sufficient condition for a vector field to be a gradient field as well. It states that for some open, simple connected set U, a continuously differentiable function $F : U \to R^L$ defines a gradient field if and only if

$$\frac{\partial F_j(\mathbf{x})}{\partial x_i} = \frac{\partial F_i(\mathbf{x})}{\partial x_j}, \quad \forall i, j = 1...L \tag{C.1}$$

In other words, integrability follows from the symmetry of the partial derivatives.

C.2 Gradient with respect to gamma

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

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \gamma_{ij}} = \frac{\partial \tilde{\mathcal{L}}}{\partial E_i} \cdot \frac{\partial E_i}{\partial \gamma_{ij}} + \frac{\partial \tilde{\mathcal{L}}}{\partial E_j} \cdot \frac{\partial E_j}{\partial \gamma_{ij}} \tag{C.2}$$

In particular,

$$\begin{aligned}
\frac{\partial E_i}{\partial \gamma_{ij}} &= \frac{\partial}{\partial \gamma_{ij}} \sum_{k=1}^{M} E_{ik} \\
&= \frac{\partial E_{ij}}{\partial \gamma_{ij}} + \frac{\partial E_{ji}}{\partial \gamma_{ij}} \\
&= \frac{\partial}{\partial \gamma_{ij}} (w_{ij} e_i^{\gamma_{ij}} e_j^{1-\gamma_{ij}}) + \frac{\partial}{\partial \gamma_{ij}} (w_{ji} e_j^{\gamma_{ij}} e_i^{1-\gamma_{ij}}) \\
&= w_{ij} e_i^{\gamma_{ij}} \frac{\partial}{\partial \gamma_{ij}} e_j^{1-\gamma_{ij}} + e_j^{1-\gamma_{ij}} \frac{\partial}{\partial \gamma_{ij}} e_i^{\gamma_{ij}} + ... \\
&= (\log e_i + \log e_j)(w_{ij} e_i^{\gamma_{ij}} e_j^{1-\gamma_{ij}} + w_{ji} e_j^{\gamma_{ij}} e_i^{1-\gamma_{ij}})
\end{aligned} \tag{C.3}$$

As we can see, the gradient with respect to γ_{ij} does indeed involve a natural logarithm of self-energies e_i and e_j. Thus, self-energies must be constrained to positive values.

Bibliography

Afouras, Triantafyllos et al. (2018). "Deep Audio-Visual Speech Recognition". In: *arXiv e-prints*, arXiv:1809.02108, arXiv:1809.02108. arXiv: 1809.02108 [cs.CV].

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].

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *arXiv e-prints*, arXiv:1409.0473, arXiv:1409.0473. arXiv: 1409.0473 [cs.CL].

Baltrušaitis, T., C. Ahuja, and L. Morency (2019). "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2, pp. 423–443. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2018.2798607.

Caltagirone, Luca et al. (2018). "LIDAR-Camera Fusion for Road Detection Using Fully Convolutional Neural Networks". In: *arXiv e-prints*, arXiv:1809.07941, arXiv:1809.07941. arXiv: 1809.07941 [cs.CV].

Cayton, Lawrence (2005). "Algorithms for manifold learning". In:

Chaudhari, Sneha et al. (2019). "An Attentive Survey of Attention Models". In: *arXiv e-prints*, arXiv:1904.02874, arXiv:1904.02874. arXiv: 1904.02874 [cs.LG].

Chauvin, Yves and David E. Rumelhart, eds. (1995). *Backpropagation: Theory, Architectures, and Applications*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc. ISBN: 0-8058-1259-8.

Chen, Mu-Yen et al. (2018). "Learning the Chinese Sentence Representation with LSTM Autoencoder". In: *Companion Proceedings of the The Web Conference 2018*. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee, pp. 403–408. ISBN: 978-1-4503-5640-4. DOI: 10.1145/3184558.3186355. URL: https://doi.org/10.1145/3184558.3186355.

Cocktail party effect (2010). *Cocktail party effect — Wikipedia, The Free Encyclopedia*. [Online; accessed 29-April-2019]. URL: https://en.wikipedia.org/wiki/Cocktail_party_effect.

Desimone, Robert and John Duncan (1995). "Neural Mechanisms of Selective Visual Attention". In: *Annual Review of Neuroscience* 18.1. PMID: 7605061, pp. 193–222. DOI: 10.1146/annurev.ne.18.030195.001205. eprint: https://doi.org/10.1146/annurev.ne.18.030195.001205. URL: https://doi.org/10.1146/annurev.ne.18.030195.001205.

D.R., Lorimer and Kramer M. (2005). *Handbook of pulsar astronomy*. Cambridge University Press.

Driver, Jon and Charles Spence (1998). "Crossmodal attention". In: *Current Opinion in Neurobiology* 8.2, pp. 245 –253. ISSN: 0959-4388. DOI: https://doi.org/10.1016/S0959-4388(98)80147-5. URL: http://www.sciencedirect.com/science/article/pii/S0959438898801475.

Duan, Kaibo et al. (2003). "Multi-category Classification by Soft-max Combination of Binary Classifiers". In: *Proceedings of the 4th International Conference on Multiple*

Classifier Systems. MCS'03. Guildford, UK: Springer-Verlag, pp. 125–134. ISBN: 3-540-40369-8. URL: http://dl.acm.org/citation.cfm?id=1764295.1764312.

Ephrat, Ariel et al. (2018). "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation". In: *arXiv e-prints*, arXiv:1804.03619, arXiv:1804.03619. arXiv: 1804.03619 [cs.SD].

Fan, Jianqing, Cong Ma, and Yiqiao Zhong (2019). "A Selective Overview of Deep Learning". In: *arXiv e-prints*, arXiv:1904.05526, arXiv:1904.05526. arXiv: 1904.05526 [stat.ML].

Fazekas, Peter and Bence Nanay (Oct. 2018). "Attention Is Amplification, Not Selection". In: *The British Journal for the Philosophy of Science*. ISSN: 0007-0882. DOI: 10.1093/bjps/axy065. eprint: http://oup.prod.sis.lan/bjps/advance-article-pdf/doi/10.1093/bjps/axy065/25875820/axy065.pdf. URL: https://doi.org/10.1093/bjps/axy065.

Galassi, Andrea, Marco Lippi, and Paolo Torroni (2019). "Attention, please! A Critical Review of Neural Attention Models in Natural Language Processing". In: *arXiv e-prints*, arXiv:1902.02181, arXiv:1902.02181. arXiv: 1902.02181 [cs.CL].

Ghahramani, Z. (2004). "Unsupervised Learning". In: *Springer*.

Ghosh, Pranab (2007). "Rotation and Accretion Powered Pulsars". In: *World Scientific Series in Astronomy and Astrophysics*.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Goodfellow, Ian et al. (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

He, K. et al. (2016). "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

Hoogi, Assaf et al. (2019). "Self-Attention Capsule Networks for Image Classification". In: *arXiv e-prints*, arXiv:1904.12483, arXiv:1904.12483. arXiv: 1904.12483 [cs.CV].

Kahneman, Daniel (1975). "Attention and effort". In:

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

Keith, M. J. et al. (Nov. 2010). "The High Time Resolution Universe Pulsar Survey – I. System configuration and initial discoveries". In: *Monthly Notices of the Royal Astronomical Society* 409.2, pp. 619–627. ISSN: 0035-8711. DOI: 10.1111/j.1365-2966.2010.17325.x. eprint: http://oup.prod.sis.lan/mnras/article-pdf/409/2/619/18579028/mnras0409-0619.pdf. URL: https://doi.org/10.1111/j.1365-2966.2010.17325.x.

Kim, Taesup and Yoshua Bengio (2016). "Deep Directed Generative Models with Energy-Based Probability Estimation". In: *arXiv e-prints*, arXiv:1606.03439, arXiv:1606.03439. arXiv: 1606.03439 [cs.LG].

Kingma, Diederik P. and Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization". In: *arXiv e-prints*, arXiv:1412.6980, arXiv:1412.6980. arXiv: 1412.6980 [cs.LG].

Kingma, Diederik P and Max Welling (2013). "Auto-Encoding Variational Bayes". In: *arXiv e-prints*, arXiv:1312.6114, arXiv:1312.6114. arXiv: 1312.6114 [stat.ML].

Ladjal, Saïd, Alasdair Newson, and Chi-Hieu Pham (2019). "A PCA-like Autoencoder". In: *arXiv e-prints*, arXiv:1904.01277, arXiv:1904.01277. arXiv: 1904.01277 [cs.CV].

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (May 2015). "Deep learning". English (US). In: *Nature* 521.7553, pp. 436–444. ISSN: 0028-0836. DOI: 10.1038/nature14539.

LeCun, Yann et al. (2006). "A tutorial on energy-based learning". In: *PREDICTING STRUCTURED DATA*. MIT Press.

Li, Guanbin et al. (2019). "Cross-Modal Attentional Context Learning for RGB-D Object Detection". In: *IEEE Transactions on Image Processing* 28.4, pp. 1591–1601. DOI: 10.1109/TIP.2018.2878956. arXiv: 1810.12829 [cs.CV].

Libovický, Jindřich, Jindřich Helcl, and David Mareček (2018). "Input Combination Strategies for Multi-Source Transformer Decoder". In: *arXiv e-prints*, arXiv:1811.04716, arXiv:1811.04716. arXiv: 1811.04716 [cs.CL].

Loog, Marco (2017). "Supervised Classification: Quite a Brief Overview". In: *arXiv e-prints*, arXiv:1710.09230, arXiv:1710.09230. arXiv: 1710.09230 [cs.LG].

Lorimer, R. Duncan (2008). "Binary and Millisecond Pulsars". In: *Living Reviews in Relativity* 11.1, p. 8. ISSN: 1433-8351. DOI: 10.12942/lrr-2008-8. URL: https://doi.org/10.12942/lrr-2008-8.

Lyon, R. J. (2016). "Why are pulsars hard to find". PhD thesis. The University of Manchester.

Narayanan, Hariharan and Sanjoy Mitter (2010). "Sample Complexity of Testing the Manifold Hypothesis". In: *Advances in Neural Information Processing Systems 23*. Ed. by J. D. Lafferty et al. Curran Associates, Inc., pp. 1786–1794. URL: http://papers.nips.cc/paper/3958-sample-complexity-of-testing-the-manifold-hypothesis.pdf.

Paszke, Adam et al. (2017). "Automatic differentiation in PyTorch". In:

Philipp, George, Dawn Song, and Jaime G. Carbonell (2017). "The exploding gradient problem demystified - definition, prevalence, impact, origin, tradeoffs, and solutions". In: *arXiv e-prints*, arXiv:1712.05577, arXiv:1712.05577. arXiv: 1712.05577 [cs.LG].

Prechelt, Lutz (1998). "Automatic early stopping using cross validation: quantifying the criteria". In: *Neural Networks* 11.4, pp. 761 –767. ISSN: 0893-6080. DOI: https://doi.org/10.1016/S0893-6080(98)00010-0. URL: http://www.sciencedirect.com/science/article/pii/S0893608098000100.

Ruder, Sebastian (2016). "An overview of gradient descent optimization algorithms". In: *arXiv e-prints*, arXiv:1609.04747, arXiv:1609.04747. arXiv: 1609.04747 [cs.LG].

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

Scholz, Matthias, Martin Fraunholz, and Joachim Selbig (2008). "Nonlinear Principal Component Analysis: Neural Network Models and Applications". In: *Principal Manifolds for Data Visualization and Dimension Reduction*. Ed. by Alexander N. Gorban et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 44–67. ISBN: 978-3-540-73750-6.

Shon, Suwon, Tae-Hyun Oh, and James Glass (2018). "Noise-tolerant Audio-visual Online Person Verification using an Attention-based Neural Network Fusion". In: *arXiv e-prints*, arXiv:1811.10813, arXiv:1811.10813. arXiv: 1811.10813 [cs.CV].

Sutton, Richard S. and Andrew G. Barto (1998). *Introduction to Reinforcement Learning*. 1st. Cambridge, MA, USA: MIT Press. ISBN: 0262193981.

Treat, J. and Stegmaier (2014). "Black Holes: Star Eater." In: *National Geographic*.

Turchenko, Volodymyr, Eric Chalmers, and Artur Luczak (2017). "A Deep Convolutional Auto-Encoder with Pooling - Unpooling Layers in Caffe". In: *arXiv e-prints*, arXiv:1701.04949, arXiv:1701.04949. arXiv: 1701.04949 [cs.NE].

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *arXiv e-prints*, arXiv:1706.03762, arXiv:1706.03762. arXiv: 1706.03762 [cs.CL].

Vincent, Pascal et al. (2008). "Extracting and Composing Robust Features with Denoising Autoencoders". In: *ICML 2008*.

Wang, Xin, Yuan-Fang Wang, and William Yang Wang (2018). "Watch, Listen, and Describe: Globally and Locally Aligned Cross-Modal Attentions for Video Captioning". In: *arXiv e-prints*, arXiv:1804.05448, arXiv:1804.05448. arXiv: 1804.05448 [cs.CL].

Watzl, Sebastian (2017). *Structuring Mind. The Nature of Attention and How It Shapes Consciousness*. Oxford, UK: Oxford University Press.

Weinberger, K.Q. and L.K. Saul (2006). "Unsupervised Learning of Image Manifolds by Semidefinite Programming". In: *Int J Comput Vision*.

Wu, Yonghui et al. (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". In: *ArXiv* abs/1609.08144.

Zhai, Shuangfei et al. (2016). "Deep Structured Energy Based Models for Anomaly Detection". In: *arXiv e-prints*, arXiv:1605.07717, arXiv:1605.07717. arXiv: 1605.07717 [cs.LG].