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

TODO

Multi-modal deep learning, especially in safety-critical areas offers the advantage of having more information to extract permitting models to make better predictions. Attention mechanisms are commonly used to combine multiple modalities. However, only a few researchers considered cases where modes could be noisy or missing. Lot of constraints, explicit, not generalize with out-of-distribution.

This thesis presents a novel attention mechanism able to adapts its focus on the most useful modes for the prediction on a per-sample basis. Present results outperform standard data augmentation techniques by up to 10% confident it could generalize to more complex datasets and architectures. Further using our definition of capacity in attention, we introduce a simple regularizer inducing the model to let less information pass enabling it to reduce the perturbations on more intensive situations.

Acknowledgements

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

Contents

List of Figures

Notation and 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
NLP	**N**atural **L**anguage **P**rocessing
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, there has been tremendous advances 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, mostly in computer vision (He et al., 2016) and natural language processing (Wu et al., 2016). Giving rise to amazing industrial application such as autonomous driving, early cancer detection, enhanced machine translation, etc. A primary concern of engineers is to make sure the trained models are error-free, which can be challenging if the input data does not carry enough information.

One possible solution researchers started to explore is to use multiple modalities[1], which makes sense since our experience of the world is multi-modal, i.e., we see objects, hear sound, feel the texture, smell odours, and taste flavours. Multi-Modal Deep Learning (MMDL) is used in the hope that the information carried by each mode is additive, such that the model can learn to make more accurate predictions. 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 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 road detection. Despite its efficacy, MMDL suffers from a major drawback: no explicit mechanisms exists to handle failing modes. In the present report, a mode is said to be failing if a) it contains a significant amount of noise, 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 network.

On the other hand, humans seem to handle these situations robustly on a daily basis. A famous example showing this ability is called the cocktail-party effect (Cocktail party effect, 2010), it refers to the difficulty we sometimes have understanding speech in noisy social settings. As a subconscious response, we tend to look at the mouth of our interlocutor i.e. we shift some attention from the auditory to the visual senses. Similarly, our attention is shifted from vision to touch when we are wandering in a room where the lights suddenly switch off. These examples indicate that humans handle modes with perturbations (first example) or missing information (second example) by shifting their attention on the other more relevant modes (Driver and Spence, 1998).

[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

Inspired from this behaviour, this report presents a new approach to tackle failing modes. I introduce a novel attention mechanism, named *Energy-based Multi-Modal Attention* (EMMA), able to decide how much attention to devote to each mode, such that the relevant information is kept while masking out the perturbations. Additionally, this work offers some important insight into how current attention mechanisms in deep learning are surprisingly similar in some ways to attention in humans.

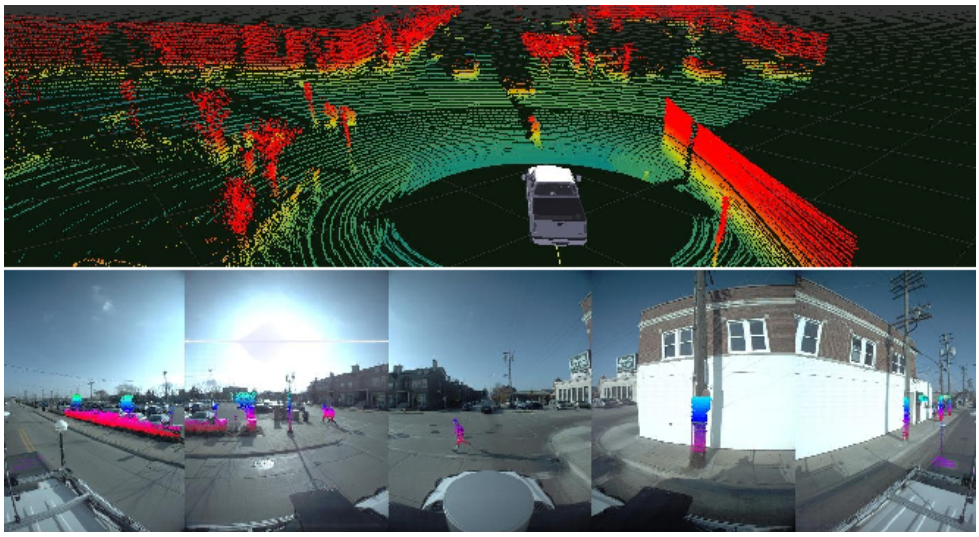


FIGURE 1.1: Same environment, different modes (top: LIDAR view, bottom: camera view)

1.2 Proposed solution

The attention module EMMA is inserted in front of the model, focusing its attention on the modes such that the most useful information passes through while the perturbations are filtered out. The amount of attention distributed to a mode is based on its importance, encompassing three intrinsically tied properties:

- *relevance*: the quantitative influence of the mode over the predictions
- *failure intensity*: a measure proportional to the outlyingness[3] of the mode
- *coupling*: how does the information carried by the mode relate to the other modes? Is it redundant, complementary or conflicting?

Let us emphasize that determining the importance is sample dependent, and is thus not easily solved by learning the global tendency.

Software Implementation

All the implemented models and experiments are available at this repository[4], with a wiki explaining how to run the experiments; PyTorch[5] was the main framework used regarding the Machine Learning part.

[3] The outlyingness of a data point tells us how far the observation lies from the center of the training set distribution

[4] https://github.com/Werenne/energy-based-multimodal-attention

[5] https://pytorch.org/

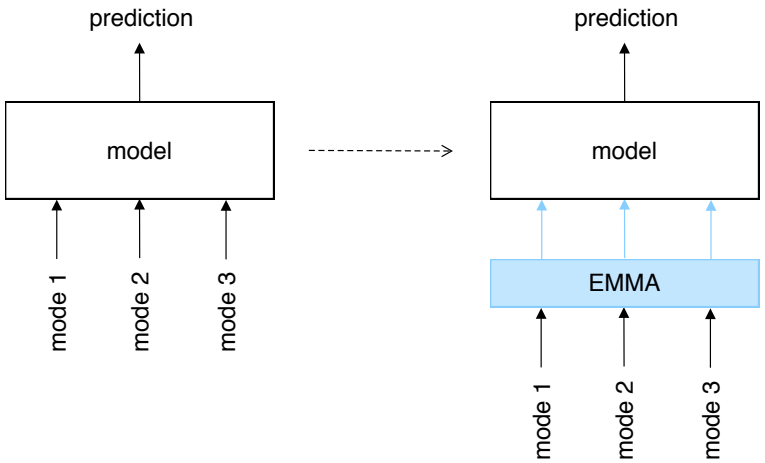


FIGURE 1.2: A multi-modal model with three input modes, without EMMA (left), improved with EMMA (right)

1.3 Contributions

The work presented in this Master thesis has led to three novel contributions.

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

Contribution 2: a simple yet powerful regularizer on attention mechanisms. We slightly modify a common attention function permitting us to establish a link to the concept of capacity in psychology (Kahneman, 1975); Capacity is the amount of attention distributed among the inputs. Subsequently, a new regularizer is introduced to control the capacity, which we claim can help generalize against unexpected situations.

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 about 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 and architecture of the Energy-based Multi-Modal Attention module (Contribution 1 & 2).

Chapter 6 presents a thorough evaluation and analysis of the module outlined in Chapter 5.

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 (Li, 2017) 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 were invented. Deep learning (Fan, Ma, and Zhong, 2019) is one of those families and is particularly powerful on perception tasks.

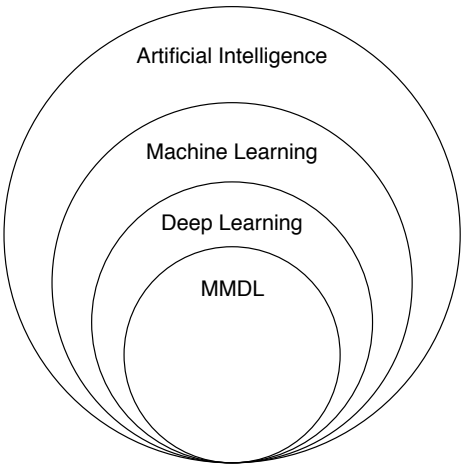


FIGURE 2.1: Venn diagram of the Artificial Intelligence field

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).

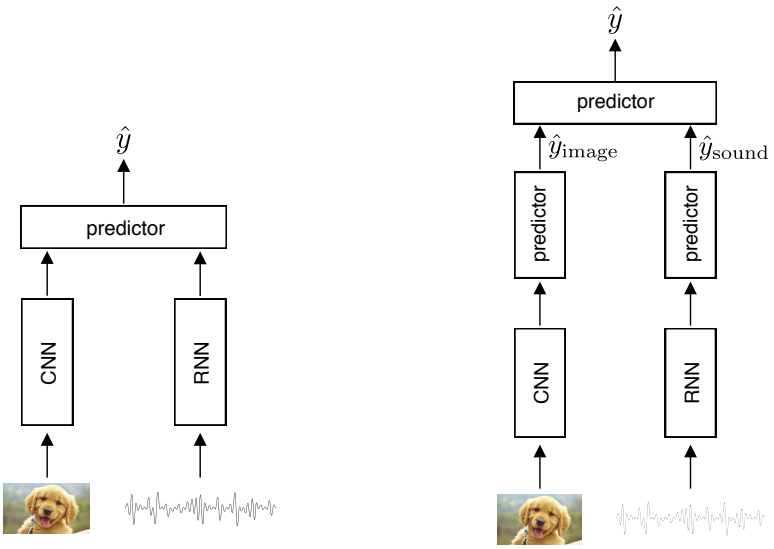
Multi-Modal Deep Learning

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).

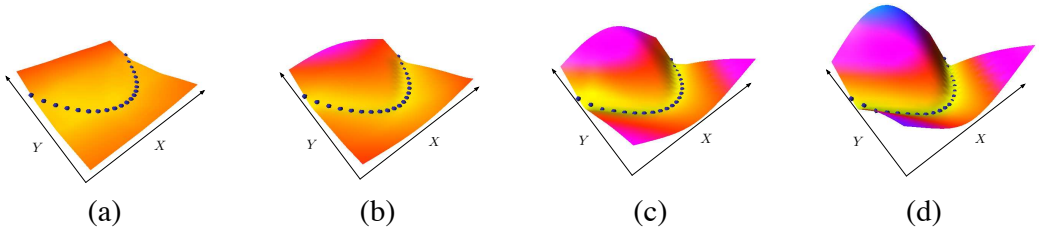


FIGURE 2.3: The shape of the energy surface at four intervals. Along the x-axis is the variable X and along the y-axis is the variable Y . The shape of the surface at (a) the start of the training, (b) after 15 epochs over the training set, (c) after 25 epochs, and (d) after 34 epochs. The energy surface 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 crossmodal 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 crossmodal 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. Lastly, crossmodal 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

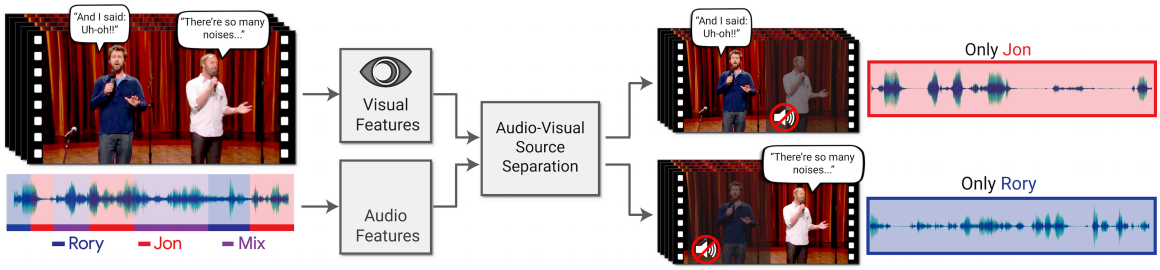
Attention mechanisms in deep learning aim to highlight specific regions of the input space. The most common way to do this, is by multiplying the input by an attention mask, where the attention mask consist of normalized continous values between zero and one. Observe the similarity with the amplication 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 crossmodal 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 crossmodal 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). Crossmodal 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 crossmodal attention and have attained impressive results. However, most research using crossmodal 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 crossmodal 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



(a) Input video frames and audio (b) Processing (c) Output clean audio for each speaker (our result)

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.

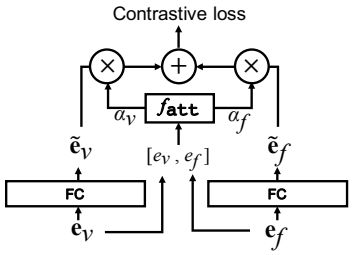


FIGURE 3.2: Neural network based fusion approaches. $\mathbf{e}_v$: speaker embedding, $\mathbf{e}_f$: face embedding. FC denotes a fully connected layer. *Image from* (Shon, Oh, and Glass, 2018)

[3] Relative to the training data

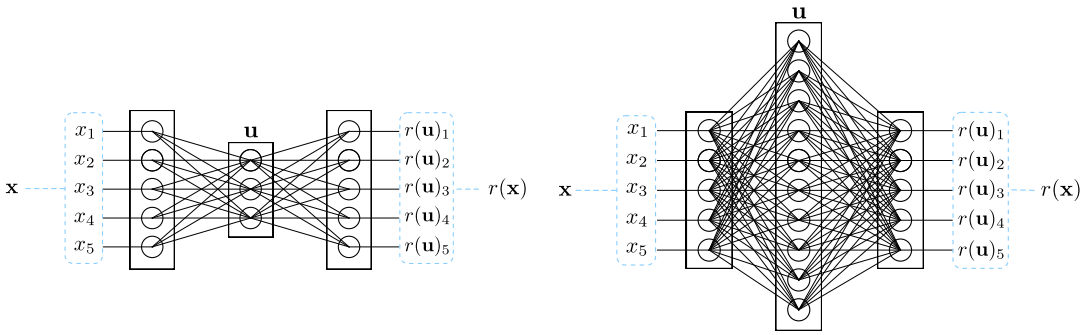
Chapter 4

Energy Estimation

As mentioned in the introduction, the EMMA module needs a measure of the outlyingness of each mode, in order to determine which modes are important and which are not. To this end we use an energy-based model (see Section 2), since their energy function is proportional to the negative log-likelihood (NLL). This chapter discusses how an energy function can be derived from an autoencoder.

4.1 Autoencoder

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 as $f(\mathbf{x}) = h(W_f\mathbf{x} + \mathbf{b}_f) = \mathbf{u}$ where $h(\cdot)$ is an activation function and $\mathbf{u}$ represents the hidden layer. The decoder is then in charge of reconstructing the input, $g(\mathbf{u}) = W_g\mathbf{x} + \mathbf{b}_g$. The output is often called the reconstruction and is written $r(\mathbf{x}) = g(f(\mathbf{x}))$. 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: the undercomplete and overcomplete autoencoders.



(A) Undercomplete AE (B) Overcomplete AE

FIGURE 4.1: The architecture of the two families of autoencoders

Undercomplete

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 (see Figure 4.1a). As a result, the input

$\mathbf{x}$ has to pass through a bottleneck, forcing the model to loose some information and keeping only the most relevant features. It can be thought of as non-linear principal component analysis (Scholz, Fraunholz, and Selbig, 2008; Ladjal, Newson, and Pham, 2019): 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 norm of the vector $r(\mathbf{x}) - \mathbf{x}$.

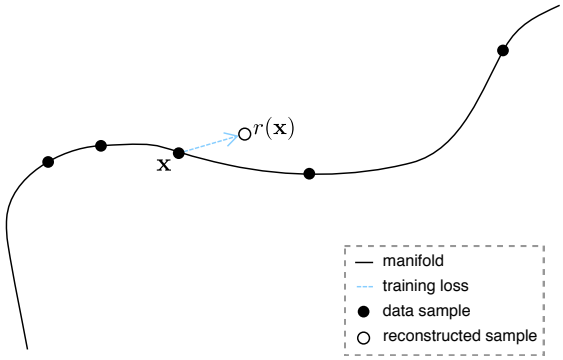


FIGURE 4.2: Vectorial representation of an undercomplete reconstruction process

Overcomplete

Conversely, an overcomplete AE has more hidden units than its input/output layer (see Figure 4.1b). Straightforwardly, the model can thus learn to perfectly copy its input to the output through the L hidden units. To spice things up, the input is corrupted before being passed through the encoder. If we force the AE to reconstruct the original input, we now have a model learning 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

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

Notice the difference with the loss function of the undercomplete AE. We verify on Figure 4.3 that minimizing the loss, $r(\tilde{\mathbf{x}}) \to \mathbf{x}$, is equivalent to learning to invert the corruption, $r(\tilde{\mathbf{x}}) - \tilde{\mathbf{x}} \to -(\tilde{\mathbf{x}} - \mathbf{x})$.

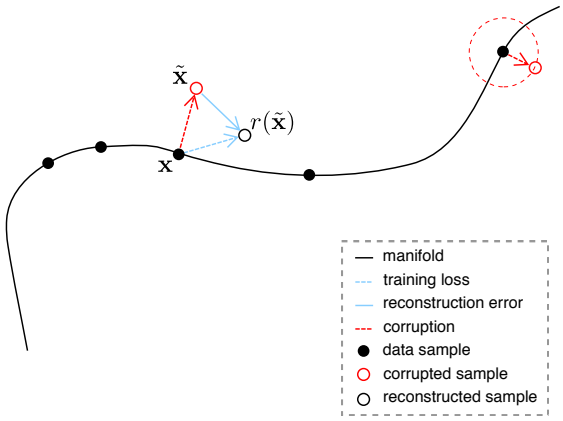


FIGURE 4.3: Vectorial representation of an overcomplete reconstruction process

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)

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

To put it differently, the reconstruction error points towards the corresponding most likely datapoint. This result is not particularly suprising, indeed, denoising a signal is essentialy equivalent to finding the most likely datapoint in the nearby neigborhood (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.

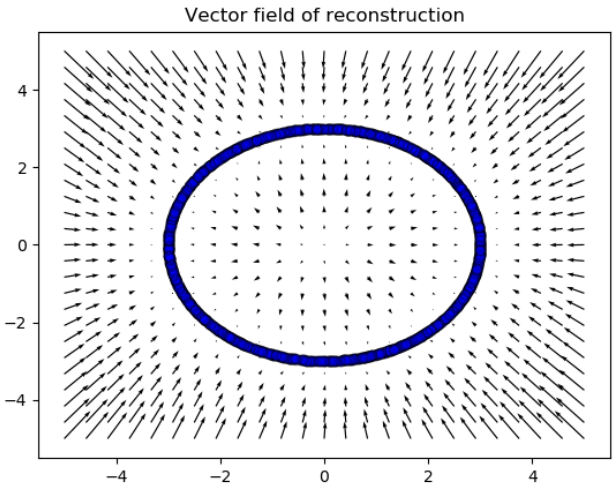


FIGURE 4.4: Vector field of reconstruction error on circle manifold. No corruption is applied at test time, the reconstruction error vector is simply the output minus the input.

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

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

where δ_{ij} denotes the Kronecker delta, $W = W_f$ and W_{ij} is the element on the i^{th} row and j^{th} columns. The vector field under those circumstances can be expressed as a gradient of a scalar field $-\Psi$, such that $r(\tilde{\mathbf{x}}) - \tilde{\mathbf{x}} = -\partial\Psi(\tilde{\mathbf{x}})/\partial\tilde{\mathbf{x}}$. In analogy to physics, the vector field can be interpreted as a force applied on the input and the scalar field as a potential energy. Thereupon, the reconstruction process can be seen as a gradient descent in the potential energy landscape (Kamyshanska and Memisevic, 2014). For our purpose, an important observation to make is that the potential energy is proportional

[1] See Appendix C.1

to the NLL,

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

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

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

Substituting $f = h(W\mathbf{x} + \mathbf{b}_f)$ and $g = W^T\mathbf{x} + \mathbf{b}_g$

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

In this work only the sigmoid activation function will be used, thus solving f

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

where $W_{.k}^T$ is the k^{th} column of W^T, and b_k^f the k^{th} element of $\mathbf{b}_f$. The intermediate steps between (4.5) and (4.7) are detailed in (Kamyshanska and Memisevic, 2014). In contrast to physics, notice 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. From those trained autoencoders, the energy function is computed on a grid mesh. 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

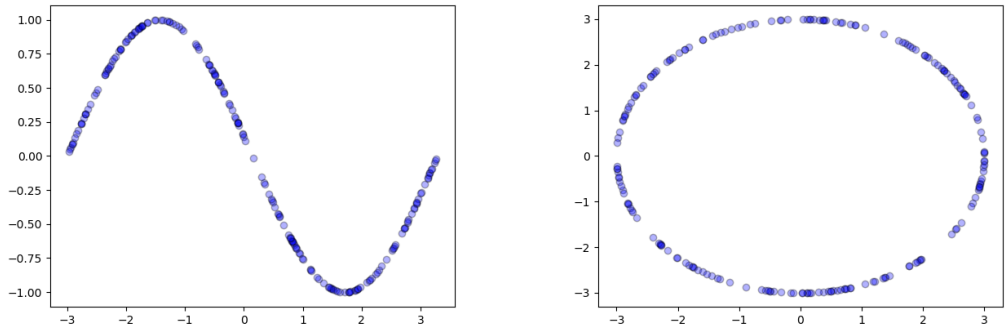
The manifolds consists of a set of N samples $\mathbf{x} \in \mathbb{R}^2$ in the form of a wave and a circle. The N samples, written $\mathbf{t}$, are randomly selected in an interval $[0, 2\pi]$, and are transformed to manifolds as

$$\text{wave} \begin{cases} \mathbf{x}_1 = \mathbf{t} - \pi \\ \mathbf{x}_2 = \sin(\mathbf{t}) \end{cases} \qquad \text{circle} \begin{cases} \mathbf{x}_1 = 3\sin(\mathbf{t}) \\ \mathbf{x}_2 = 3\cos(\mathbf{t}) \end{cases}$$

The result of this process can be viewed on Figure 4.5.

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 used optimizer is *Adam* (Kingma and Ba, 2014).

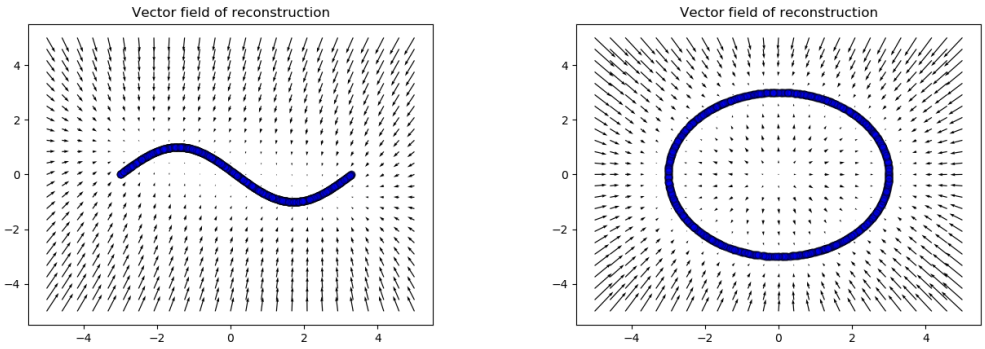


(A) Wave (B) Circle

FIGURE 4.5: Manifold generation of 200 samples

Results

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



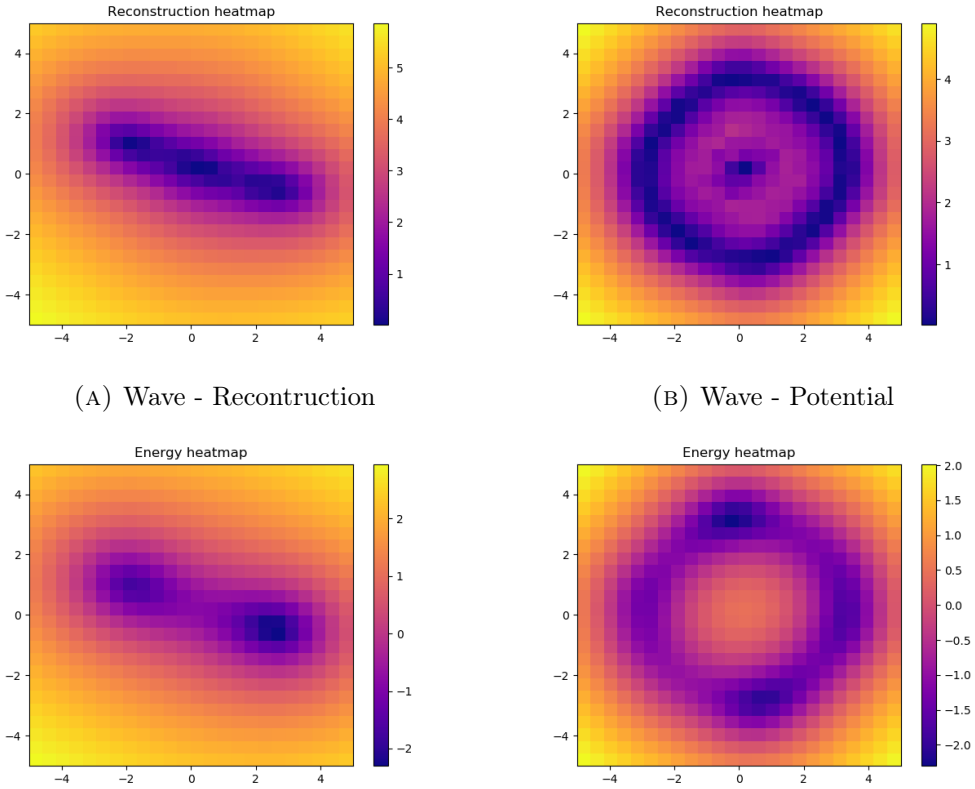
(A) Wave (B) Circle

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

The energy function and 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 reconstruction norm has also low values at the origin, which can be explained by the fact that the norm of the vectors is small at the source.

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. Likewise, CNN-DAE are more appropriate to model spatial structures, such as images. However, the integrability criterion for these models is not satisfied anymore,



(A) Wave - Recontruction

(B) Wave - Potential

(C) Circle - Recontruction

(D) Circle - Potential

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

and thus can not estimate the negative log-likelihood. Alternative methods to efficiently learn energy functions for spatial or sequential data are: (Zhai et al., 2016; Kim and Bengio, 2016)

Chapter 5

Energy-based Multi-Modal Attention

The literature review (Chapter 3) showed that previous research in MMDL has been mostly focused on leveraging the 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 enough information for the task, 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

We define the i.i.d. dataset $\mathcal{D}^{(N)}$ with N samples $(\mathbf{X}, y)$. The input $\mathbf{X}$ is composed of M modes $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ of possibly different dimensions, such as images and sounds. The multi-modal network will be abbreviated as MMN. This model tries to make predictions $\hat{y}$ as close as possible to the groundtruth y. The internal architecture of the MMN is often structured as a many-to-one encoder-decoder as discussed in Section 2.2. Nonetheless, the EMMA module is not constrained to any specific internal MMN architecture.

Our attention module is placed in front of the MMN to highlight the most valuable modes, on a per-sample basis. This is done by computing an *importance score* α_i for each mode: a scalar value between zero and one, more important modes corresponding to higher values. The method used by the model to compute the importance is further detailed in the next paragraph. From each importance score, the model determines an *attention score* β_i (more details in Section 5.4), which is representative for the amount of information of the mode that can pass through. Each mode is thus multiplied by its respective attention score (see Figure 5.1). Section 5.4 will clarify why the modes are not multiplied by the importance scores instead.

In the introduction of this work, we identified three properties influencing the importance: relevance, failure intensity and coupling. The failure intensity of mode i can be measured by a potential energy function Ψ_i (see Chapter 4) obtained by training an autoencoder on mode i. To encompass the two other properties (relevance and coupling) as well, we introduce the *modal energy* E_i as

$$E_i = e_i(\Psi_i) + \sum_{k \neq i}^{M} e_{ik}\big(e_i(\Psi_i), e_k(\Psi_k)\big) \tag{5.1}$$

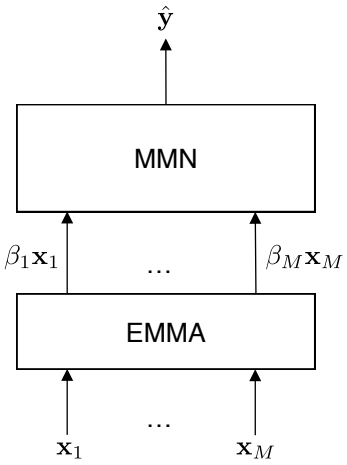


FIGURE 5.1: High-level view of a Multi-Modal Network with the EMMA module

which is constructed such that it takes, low values if the mode of the sample is important, and high values otherwise. Each modal energy is composed of its self-energy (e_i) and the shared energies (e_{ik}) with all the other modes. We expect the self-energy function to spontaneously learn the relationship between the relevance and the outlyingness (i.e. failure intensity), since the self-energy is optimized with respect to the loss on the predictions and is a function of the potential energy (more details in Section 5.2). Whereas, shared-energies are designed to capture the optimal coupling between the modes (more details in Section 5.2). Finally, the modal energies are normalized via the Boltzmann distribution[1] to the importance scores (more details in Section 5.3).

There are two ways of interpreting the proposed solution in this chapter. First, EMMA can be seen as a sort of gate filtering perturbations out. Indeed, failing modes can provoke high activations in the MMN, disturbing the predictions. But by masking the outlying modes we diminish those activations, making it easier for the MMN to make good preditions. Another way to view it, is to understand that the MMN model easily extracts β_i and $\mathbf{x}_i$ from the multiplication $\beta_i\mathbf{x}_i$. The model can then learn to make more robust predictions based on the extra inputs β_i.

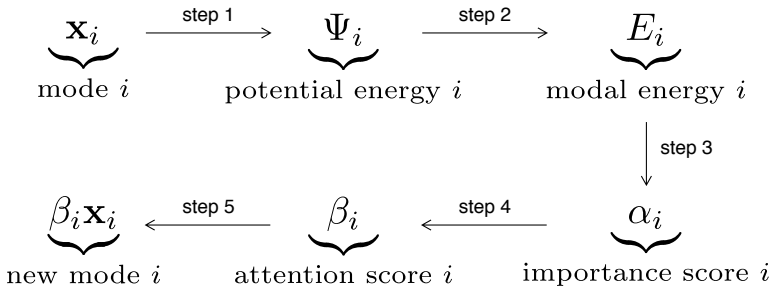


FIGURE 5.2: Summary of main steps in EMMA (step 2, 3 and 4 are detailed in the following sections, step 1 was explained in Chapter 4)

[1] See Section 2.3

5.2 From Potential to Modal energies (step 2)

We compute the *self-energy* as an affine function of the potential energy,

$$e_i = w_i \Psi_i + b_i \qquad \text{with} \qquad w_i, b_i \in \mathbb{R}^+ \tag{5.2}$$

where the parameters w_i and b_i are trained via a loss function on the predictions, in consequence the model is able via the self-energy to capture both the relevance and failure. The second advantage of this transformation is that it helps the module to face potentials on different scales, since Equation (4.7) only guarantees being proportional to the NLL, consequently potentials of different modes may not be on commensurate scales. The reason the parameters are constraint 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.

Once the self-energies obtained, we can now compute the *shared energies*. The expression e_{ij} denotes the shared energy of mode j on i and is constructed from the self-energies as follows

$$e_{ij} = w_{ij} e_i^{\gamma_{ij}} e_j^{1-\gamma_{ij}} \qquad \text{with} \qquad w_{ij} \in [-1, +1],\ \gamma_{ij} \in [0, 1] \tag{5.3}$$

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}. The direction of coupling is useful to distinguish modes with redundant or conflicting information from those with complementary information. The latter is true since self-energies are guaranteed to be positive (see next paragraph). 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 (5.3) is that the evaluation of the gradient during the backpropagation step now involves taking the logarithm of e_i[2], which is undefined for negative values. As the weights in Equation (5.2) 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

$$\Psi_i \leftarrow \max(\mathrm{e}, \Psi_i - \Psi_i^{(\min)} + \mathrm{e}) \tag{5.4}$$

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[3] ($\Psi_i \geq \mathrm{e}$) and guarantees self-energies

[2] See Appendix C.2

[3] 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)

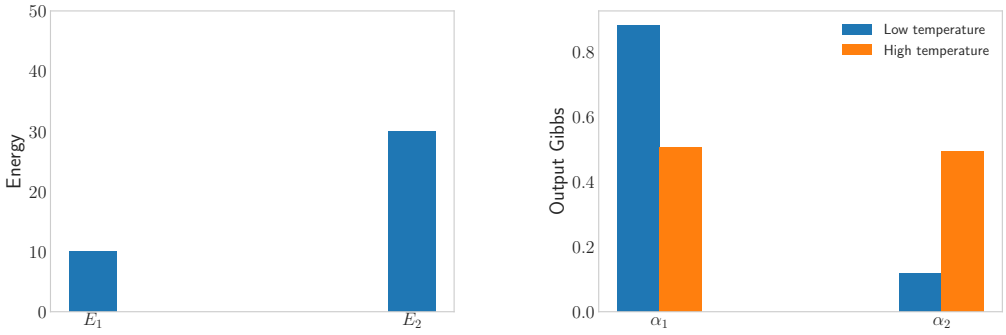
to be positive. The reason a max-operator is used is because lower energy values than $\Psi_i^{(\min)}$ can occur during inference. Clearly, this correction must be performed prior to the computation of self-energies i.e., prior to Equation (5.2).

5.3 From Modal energies to Importance scores (step 3)

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

$$\alpha_i = \frac{1}{Z} e^{-\rho E_i} \quad \text{with the partition function} \quad Z = \sum_{k=1}^{M} e^{-\rho E_k} \tag{5.5}$$

This guarantess the scores to be normalized and summing 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 on Figure 5.3, the coldness has a significant influence on the overall behaviour of the attention module; Careful tuning of ρ is thus necessary.



(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 ($\rho = 0.001$)

5.4 From Importance to Attention scores (step 4)

The attention scores are given by

$$\beta_i = \tanh(g_a \alpha_i - b_a) \quad \text{with} \quad g_a > 0,\ b_a \in [0, 1] \tag{5.6}$$

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). The latter two concepts are detailed below.

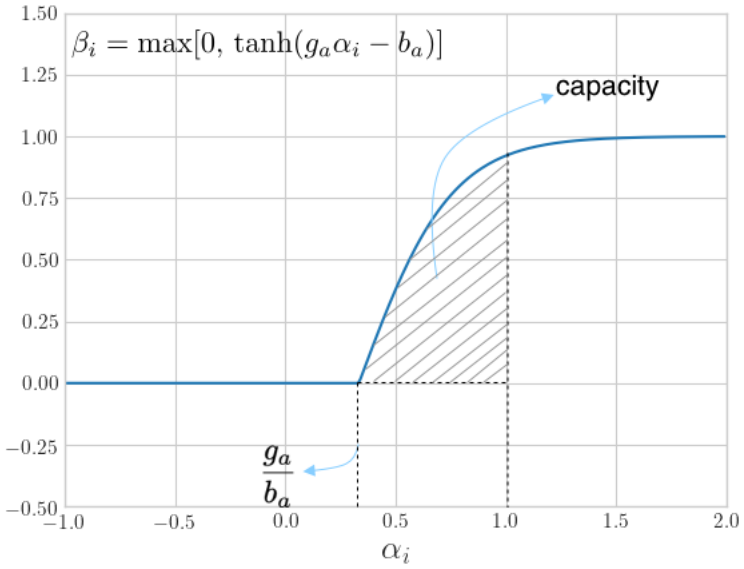


FIGURE 5.4: Attention function (the max-operator generalizes the attention function to cases where $\alpha \in \mathbb{R}$)

Energy threshold

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

$$\begin{aligned} &\Leftrightarrow \log(\alpha_i) > \log(b_a/g_a) \\ &\Leftrightarrow E_i \geq \frac{\log(g_a/b_a) - \log(Z)}{\rho} = E_{\text{threshold}} \end{aligned} \tag{5.7}$$

where $E_{\text{threshold}}$ represents the maximal amount of energy is allowed to have, to not be completely masked of (see Figure 5.4). We deduce that the learned gain and bias control this threshold. Nevertheless, the value of the partition function Z also controls the threshold, making it dynamic. The partition function will be higher if the total energy ($\sum_i E_i$) is higher, resulting in a diminshment of the threshold. To put it in another way, EMMA adapts the selectiveness with respect to the overall quality of the entire input sample. Notice that the influence of the temperature (ρ^{-1}) is non-trivial to analyse, because Z also depends on ρ.

Capacity

A more common way to write the attention function would be $\tanh(\mathbf{W}\boldsymbol{\alpha} + \mathbf{b})$, whereas we have $\tanh(g_a\mathbf{I}\boldsymbol{\alpha} - b_a\mathbf{u})$ with the unit vector $\mathbf{u} = (1 \dots 1)^T$. 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,

$$\text{capacity} \triangleq \int_0^1 \tanh(g_a\alpha - b_a)d\alpha \tag{5.8}$$

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

$$\frac{du}{d\alpha} = g_a \Leftrightarrow d\alpha = \frac{1}{g_a}du \tag{5.9}$$

we can write

$$\begin{aligned} \text{capacity} &= \frac{1}{g_a}\int_{-b_a}^{g_a-b_a} \tanh(u)du \\ &= \frac{1}{g_a}\log[\cosh(u)]\Big|_{-b_a}^{g_a-b_a} + \cancel{\text{constant}} \\ &= \frac{1}{g_a}\log\left[\frac{\cosh(g_a-b_a)}{\cosh(-b_a)}\right] \end{aligned} \tag{5.10}$$

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, it can be interesting to try to control this trade-off. To this end, we created a simple regularize which is discussed in the next section. Observe that the concept of capacity can also be applied to $\tanh(\mathbf{W}\boldsymbol{\alpha}+\mathbf{b})$, but 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 freezed. 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, in most cases the computational overload induced by the training of EMMA in the second stage will be negligible with respect to the MMN, since the number of parameters of EMMA will be far less than the number of parameters of the MMN. Moreover, the parameters of EMMA have a more contrained domain.

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

$$\tilde{\mathcal{L}} = \mathcal{L}(y,\hat{y}) + \lambda_c(g_a - b_a) - \lambda_e\Omega \quad \text{with} \quad \Omega = \sum_{k=1}^{M}\xi_k\log(\alpha_k) \quad \text{and} \quad \xi_k = \begin{cases} \xi_- = -1 & \text{if } \mathbf{x}_k \text{ is corrupted} \\ \xi_+ = +1 & \text{otherwise} \end{cases} \tag{5.11}$$

with λ_c and λ_e being positive real numbers used to set the relative importance of each regularizer. The first regularizer minimizes the capacity, where a higher λ_c pushes the module to let less information pass through (i.e., to be more "cautious"), which we suggest could hep generalize against more intensive failing modes. Another reason to mask out more information is to avoid extreme cases where the module lets all inputs unchanged and instead it is the MMN who learns to suppress perturbations. Secondly, the purpose of regularizing the energy ($\lambda_e\Omega$) is to control the trade-off between, on one

side the relevance and coupling, and on the other side the failure intensity. Without a regularizer, the parameters of the modal energy functions are optimized only regarding the predictions ($\mathcal{L}$), possibly leading to a large discrepancy between modal energies E_i and their original potential energies Ψ_i. Although the energy regularizer is relatively straightforward, we will show 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[4] 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

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \epsilon\lambda_e \nabla_{\boldsymbol{\theta}} \Omega \tag{5.12}$$

Remember the objective, we want this update to minimize the discrepancy, thus decrease/increase modal energies E_i for low/high potential energies Ψ_i. To verify this let us compute[5] $\nabla_{\boldsymbol{\theta}}\Omega$,

$$\nabla_{\boldsymbol{\theta}} \Omega = \sum_{k=1}^{M} \xi_k \nabla_{\boldsymbol{\theta}} \log(\alpha_k) \tag{5.13}$$

The gradient of the logarithm can be developed as

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

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

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

The gradient of the regularizer can now be computed by plugging Equation (5.15) into the summation (5.13). Let M' be the number of uncorrupted modes. We obtain

[4] See Section 5.2

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

for an uncorrupted mode i,

$$\nabla_{\boldsymbol{\theta}_i}\Omega = \xi_+\big[-\rho(1-\alpha_i)\nabla_{\boldsymbol{\theta}_i}E_i\big] + \big[(M'-1)\xi_+ + (M-M')\xi_-\big]\alpha_i\rho\nabla_{\boldsymbol{\theta}_i}E_i \tag{5.16}$$

and for a corrupted mode i,

$$\nabla_{\boldsymbol{\theta}_i}\Omega = \xi_-\big[-\rho(1-\alpha_i)\nabla_{\boldsymbol{\theta}_i}E_i\big] + \big[M'\xi_+ + (M-M'-1)\xi_-\big]\alpha_i\rho\nabla_{\boldsymbol{\theta}_i}E_i \tag{5.17}$$

Substituting ξ_k, we can summarize Equations (5.16) and (5.17) as

$$\boxed{\nabla_{\boldsymbol{\theta}_i}\Omega = -\big[(M-2M')\alpha_i + \xi_i\big]\rho\nabla_{\boldsymbol{\theta}_i}E_i} \tag{5.18}$$

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

$$\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \epsilon\lambda_e\rho\xi_i\nabla_{\boldsymbol{\theta}_i}E_i \quad \text{with} \quad \lambda_e \in \mathbb{R}^+ \tag{5.19}$$

Ignoring the second-order effects of the Taylor expansion of the modal energy function, it can be concluded from the equation above that the regularizer will update the parameters such that the values of the modal energy E_i increases/decreases if mode i is corrupted/uncorrupted.

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 (5.12) becomes

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

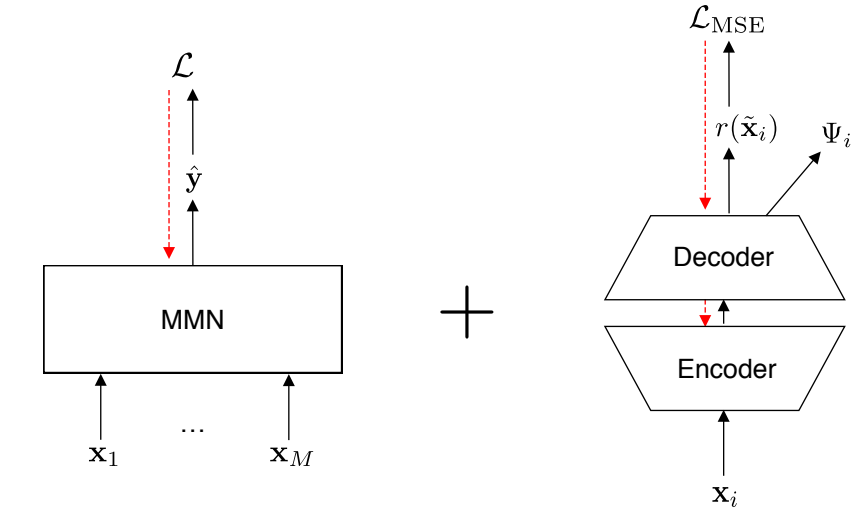
which is unstable for uncorrupted modes leading to a collapse where all energies tend to decrease.

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). 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).



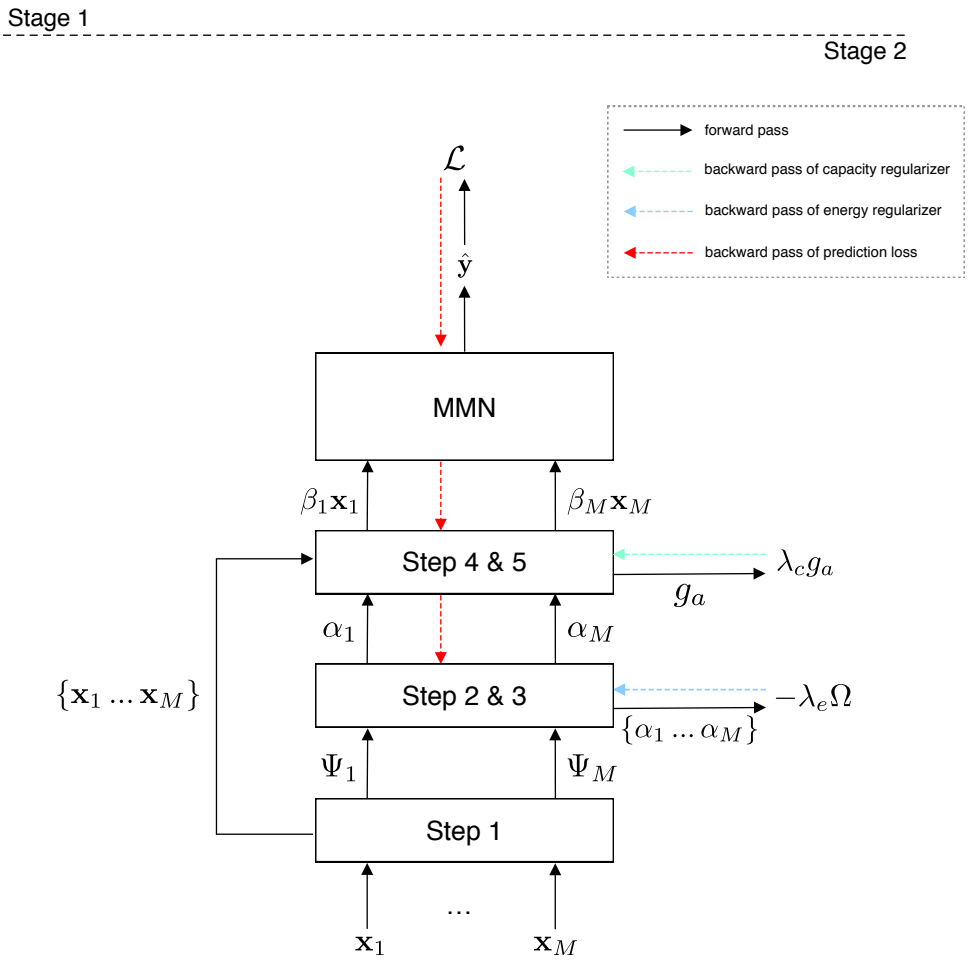


FIGURE 5.5: Summary of end-to-end training

Chapter 6

Experiments & Results

In this chapter the attention module outlined in Chapter 5 will be evaluated in two main experiments. The first experiment evaluates the potential energy[1] as a differentiatior of failing modes on a real-case dataset. In the second experiment, the robustness and interpretability is analysed on an MMN, with and without EMMA.

6.1 Pulsar Stars

For the following experiments, our models will be trained on the HTRU2 dataset[2], 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 emission detectable here on Earth. They 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[3] and skewness of the integrated profile.
- *dispersion measure* (DM): the amount of dispersive smearing 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 samples, 16.259 negative samples). This issue was handled by imposing a penalty in the loss function to fight the class imbalance.

[1] Discussed Chapter 4

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

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

	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

In the experiments the dataset will be split into a training, validation and test set (see Table 6.1). The validation set is used for tuning purposes, and to implement an Early Stopping[4] algorithm. Commonly, the model is then retrained for a few iterations on both the combined training and validation set, however, for the sake of simplicity, we do not retrain the models on the combined sets. Prior to the training, the data is standardized for the purpose of having the same signal-to-noise ratio corruption effect on all the features. To avoid information leakage, the statics (i.e., mean and variance) for the standardization are computed from the training set instead of the whole dataset, and applied to the three splits.

6.2 Experiment II

To verify whether the potential 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 measure a clear difference in the values of the potential enery on the the failing samples. Notably, missing values are implicitly solved provided that they are replaced by zeros[5]. With this in mind, we did not evaluate missing modes. Further details about the experimental setup are described in Appendix B.1.

Noisy values

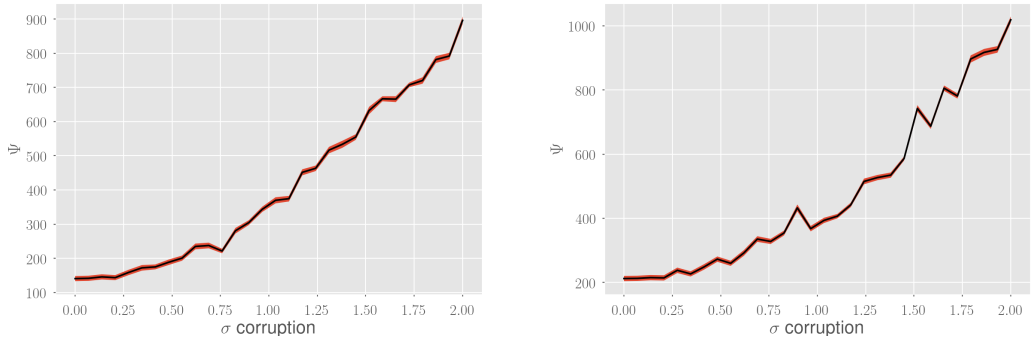
To simulate noisiness we add Gaussian white noise, $\mathcal{N} \sim (0, \sigma^2_{\text{corruption}})$, to the test set. We compute the mean and variance of the potential energies of all the samples. This process is then repeated for different noise intensities. The result for each mode is show in Figure 6.1. It is confirmed that potential energy can capture the noisiness of data, relative to the training samples.

Out-of-distribution samples

Since we only have two classes, we modify the conventional procedure of training the autoencoders on all the samples, and instead train it only on positive samples. In this case, the negative samples can be considered as being out-of-distribution. The potential energies computed on the test set (containing positive and negative samples) were as expected (see Figure 6.2): potentials of likely data (positive) are lower and distinguishable from unlikely data (negative).

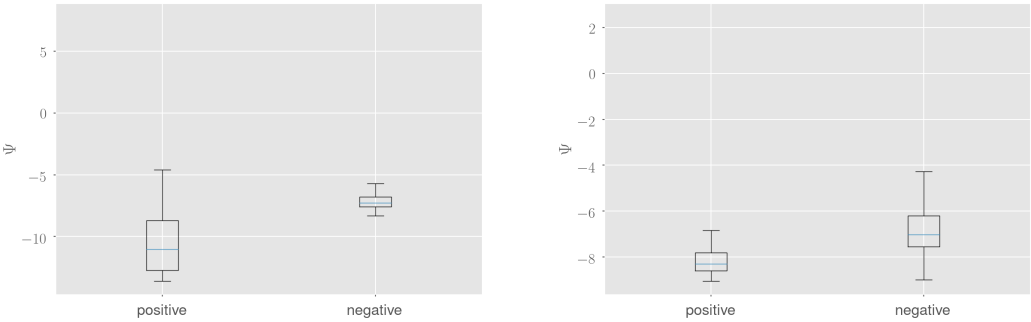
[4]Early stopping (Prechelt, 1998) is a form of regularization to avoid overfitting, where the error on the validation set is used in determining when overfitting has begun i.e., when the validation error starts increasing.

[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)



(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 III

To asses the robustness gain from using EMMA along the MMN, we compare it with a standard data augmentation technique consisting in adding noisy samples to the training set in the hope that the MMN learns to suppress the noise by itself. Furthermore, three types of models will be 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). Our hope is that the attention module will learn how to handle failing modes and be able to better generalize than `model-without`. The MMN being released of this task, only has to learn to make predictions with available highlighted information.

The corruption process is applied in the following pattern: 50% stays uncorrupted, 25% noise on IP mode only, 25% noise on DM mode only. The noisiness will be simulated 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. The three types of models are trained for 30 epochs with Early Stopping, where the variety implemented is as follows: 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 minima. Moreover, on the validation set the optimal classification threshold is determined by choosing the threshold on the ROC curve which is the nearest to the upper left corner.

In Table 6.2, the F1-scores are showed of the three types of models on uncorrupted and corrupted samples[6] of the test set. Several combinations (125) of hyperparameters were tested, only the fifteen best ones are displayed in the table below[7]. Obviously, the `base-model` performs badly on samples with a noisy mode, since it was only trained on uncorrupted data. As can be seen, using data augmentation improves the performance on noisy data while being slightly worsened on uncorrupted samples. In contrast, models with our attention module outperform `model-without` and `base-model` by a significant margin. In the further experiment we will demonstrate how this difference is increased for more intensive failing modes. Surprisingly, some `model-with` slightly improve the F1-score on uncorrupted data of the `base-model`. No explanation could be found with certainty, however, it could be explained that EMMA is able to capture a certain spectrum of quality about the data, this information could then be used by the MMN to improve its predictions. Lastly, we can notice that the models are generally more robust when the failing mode is DM then when it is mode IP. This could mean that the IP mode carries more relevant information, since the SNR is the same on the two modes, we can rule out the possibility of a less sensitive mode.

Attention-shift

In order to study the attention shift, we display the importance and attention scores for different levels of noises. Figure 6.3 shows a clear attention shift. Interestingly, for a lower level of noise, more attention will be given to the IP-mode even when it is noisy, however, if the noise intensity is high enough the attention shifts towards the other mode (see Figure 6.3d). Notice that these results are consistent with Table 6.2 where we suggested that the IP mode seems to be more relevant. A more complete view of what the attention module actually learns is illustrated in Figure 6.4.

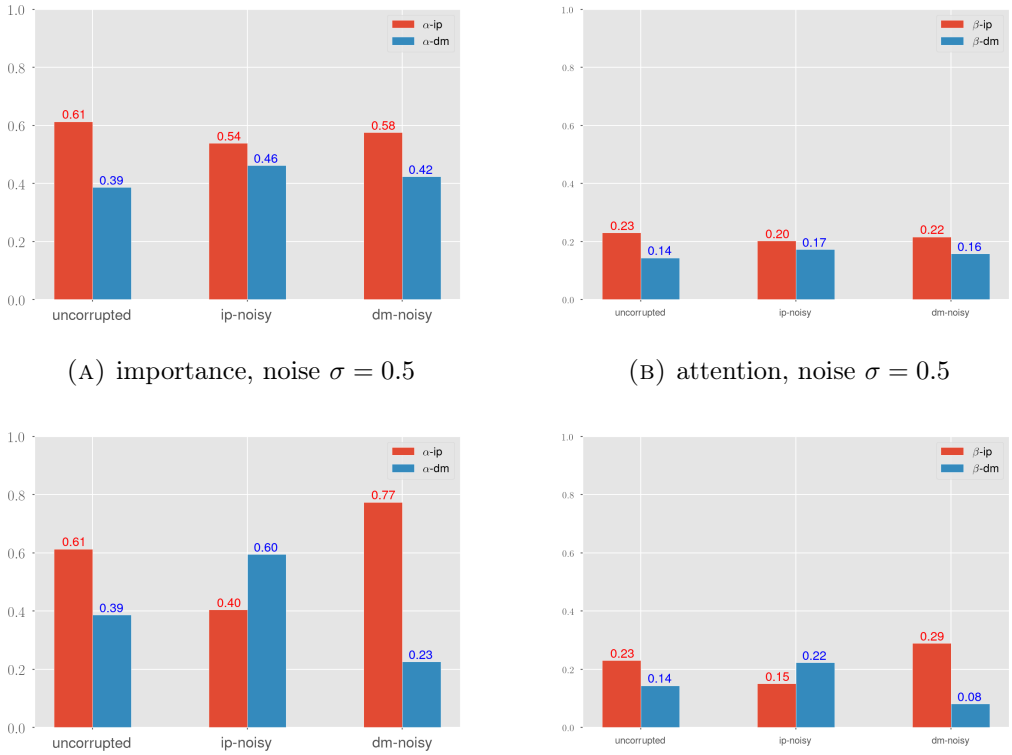
In Figure 6.5, the importance scores of two models trained with different temperatures are shown. As expected, the entropy of the importance scores is higher for the one with a low coldness ρ (Figure 6.5a) enabling it to stay more stable for higher level of noise (Figure 6.5b). On the contrary, the model trained with a lower temperature has more pronounced attention shifts. To better visualise this, the attention scores are plotted on the continuum of noise intensities in Figure 6.6. The crossing point in Figure 6.6a demonstrates that this model learns the trade-off between relevance and failure intensity. The model in Figure 6.6c is not able to learn it since its temperature and capacity are too high. Moreover, in Figure 6.7 it can be seen that regularizing the capacity too strongly can lead to a collapse of the attention scores i.e., all modes are masked.

[6] Same corruption process as the training set, $\sigma_{\text{corruption}} = 0.5$

[7] Ranked by average F1-score on corrupted and uncorrupted samples

	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 models (the models are ranked by their weighted average F1-score)



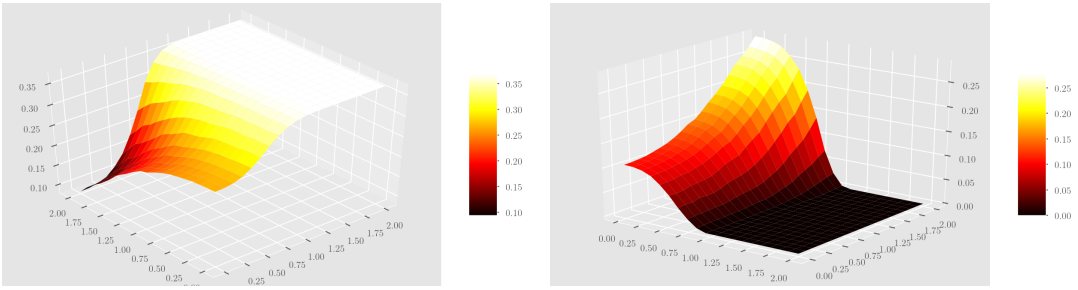
(A) importance, noise $\sigma = 0.5$

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

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

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

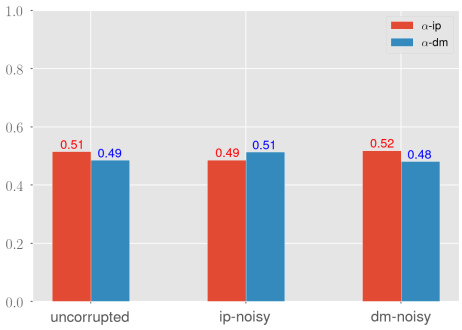
FIGURE 6.3: Importance and attention scores for the 1st ranked `model-with` ($\rho = 10^{-4}$, $\lambda_e = 10^{-3}$, $\lambda_c = 10^{-2}$), on two different levels of noises



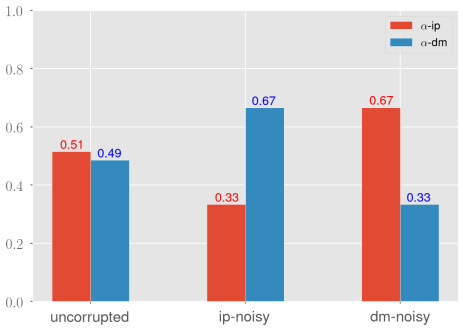
(A) β-ip

(B) β-dm

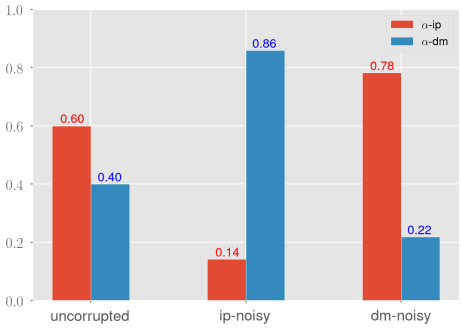
FIGURE 6.4: 3D visualization of attention scores on varying levels of noises between the modes (of the same model as Figure 6.3



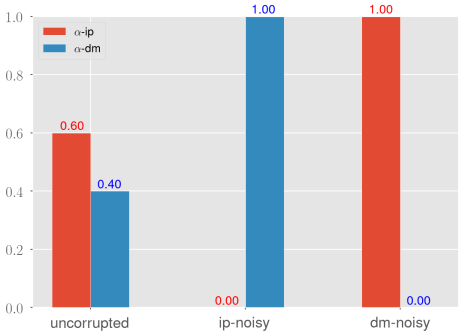
(A) importance, $\rho = 10^{-4}$, noise $\sigma = 0.5$



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

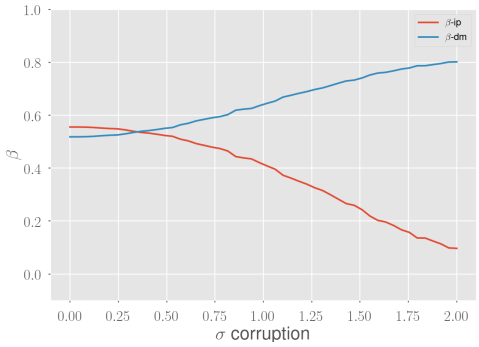


(C) importance, $\rho = 10^{-1}$, noise $\sigma = 0.5$

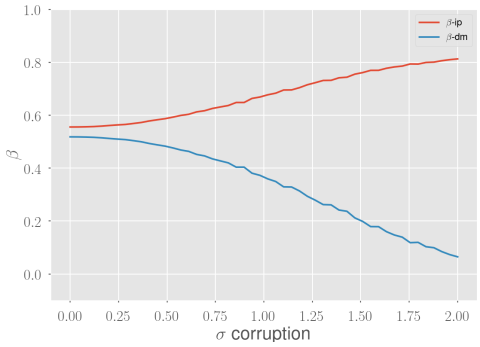


(D) importance, $\rho = 10^{-1}$, noise $\sigma = 2$

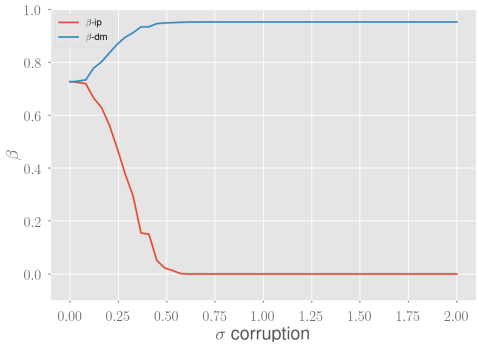
FIGURE 6.5: Importance scores comparison of models with different temperatures: `model-with` ($\rho = 10^{-4}$, $\lambda_e = 10^{-2}$, $\lambda_c = 10^{-3}$) and `model-with` ($\rho = 10^{-1}$, $\lambda_e = 10^{-2}$, $\lambda_c = 0$)



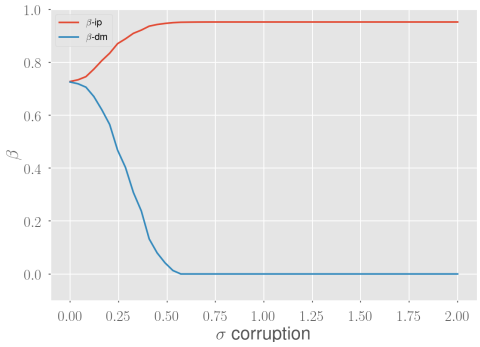
(A) attention, $\rho = 10^{-4}$, noise on IP



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

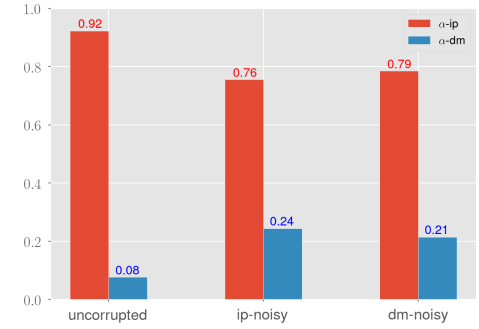


(C) attention, $\rho = 10^{-1}$, noise on IP

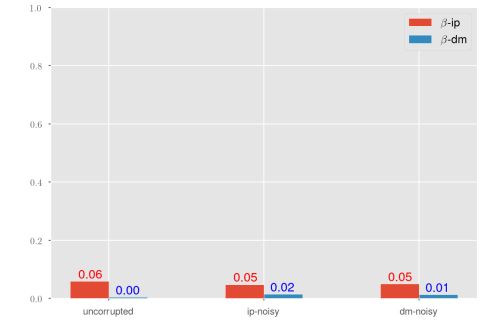


(D) attention, $\rho = 10^{-1}$, noise on DM

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` ($\rho = 10^{-1}$, $\lambda_e = 10^{-2}$, $\lambda_c = 0$). The learned capacity for these models are respectively 0.49 and 0.63



(A) Importance



(B) Attention

FIGURE 6.7: Importance and attention scores for `model-with` ($\rho = 10^{-3}$, $\lambda_e = 10^{-4}$, $\lambda_c = 10^{-1}$). The learned capacity is 0.027

Total Energy

As mentionned previously, one of the advantages of the attention module is the interpretable clues it provides. One of those is the total energy, corresponding to the sum of all modal energies of the sample. To analyse this quantity, we add noise on both modes of the whole test set. Next, we compute the total energy and F1-score for each sample. This process is repeated for increasing noise intensity, as a result we show on Figure 6.8 the mean of the total energy and F1-score for each level of noise. A significant correlation can be observed between the two, hence, the total energy could be used as a proxy for uncertainty on the predictions. Remarkably, this pattern is observed in all the modules for whatever set of hyperparameters. The usefulness of the energy regularizer could thus be discused. However, in more complex datasets with more modes the coupling effect could cancel this correlation, in those cases the regularizer can help to find the optimal trade-off.

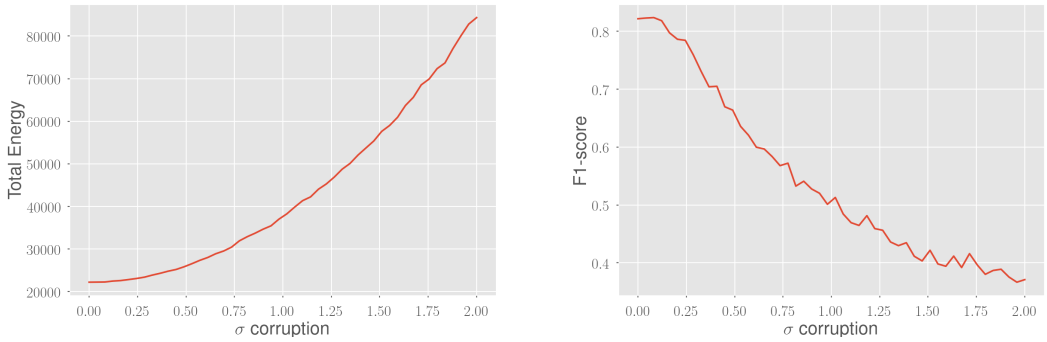
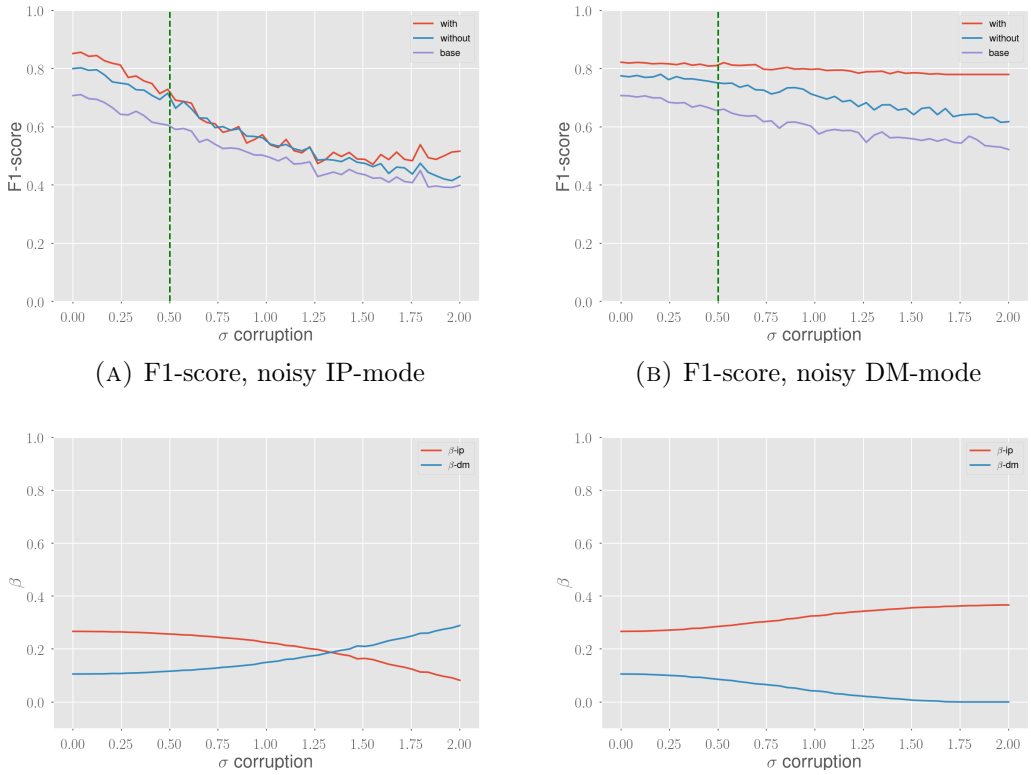


FIGURE 6.8: Total energy for `model-with` ($\rho = 10^{-4}$, $\lambda_e = 10^{-3}$, $\lambda_c = 10^{-2}$)

Robustness Generalisation

As was discussed in Chapter 3, robustness is often tested in the same conditions as the training. This does not seem to be a realistic assumption. To assess the robustness, we are going to check the F1-score on test set corrupted with varying levels of noises. What we would like to have is stability: it does not matter how high the perturbations are, if the other mode takes over the F1-score must keep stable. The 1st ranked model in Table 6.2, is evaluated in this manner and the results are shown in Figure 6.9. When the DM mode is noisy (see Figure 6.9b), the module performs as we wish and masks the perturbations out thanks to EMMA which the other models are not able to do. There is still a small decrease in the F1-score of the attention module on the noisy DM mode, but this can be explained by the fact that we lose information form the DM-mode. For the IP-mode, the `model-with` does not perform better than `model-without` (Figure 6.9a). However, it can be observed that the performance seem to stabilize from the point the attention-shift between the mode occurs ($\sigma \approx 1.35$). To compare, we show another model (see Figure 6.10) which learned a higher capacity (less regularization) and was thus able to learn the attention shift more optimally. As a consequence, we see that the F1-score remains drastically more constant (see Figure 6.10a). There is still a descrease, which is due to the loss of information of the IP-mode. Another example is shown in Figure 6.11, were the the extreme case is shown for a too low capacity.



(A) F1-score, noisy IP-mode

(B) F1-score, noisy DM-mode

(C) attention score, noisy IP-mode

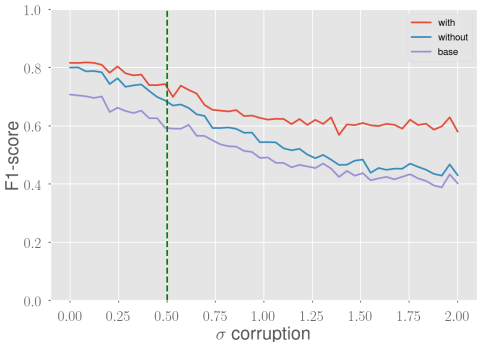
(D) attention score, noisy DM-mode

FIGURE 6.9: Noise generalisation of `model-with` ($\rho = 10^{-4}$, $\lambda_e = 10^{-3}$, $\lambda_c = 10^{-2}$). The learned capacity is 0.19

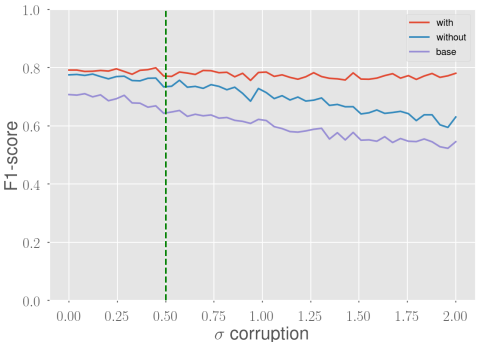
6.4 Limitation

Several limitations of our experiments can be noticed:

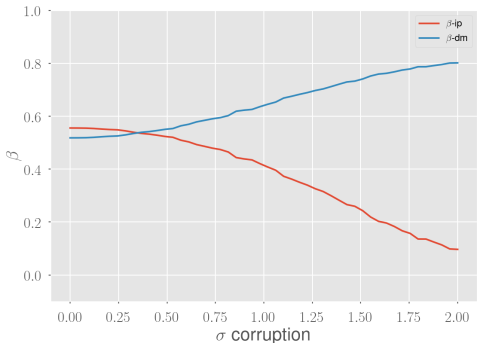
- The main drawback is that the dataset only contained two modes, making it not possible to study asymmetric dependencies
- The data were fairly simple. Not tested on high-dimensional complex data such as images. However, we think the module would perform even better since for such data suppressing noise is more difficult
- Lastly no out-of-distribution class were in the dataset. As was shown in experiment II the potential energies seems able to distinguish it. The challenge would be to see if there is a conflict between the scales. Out-of-distribution samples with lower energies than noisy modes need to be more masked



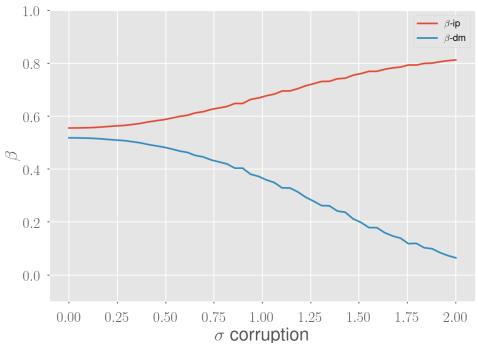
(A) F1-score, noisy IP-mode



(B) F1-score, noisy DM-mode

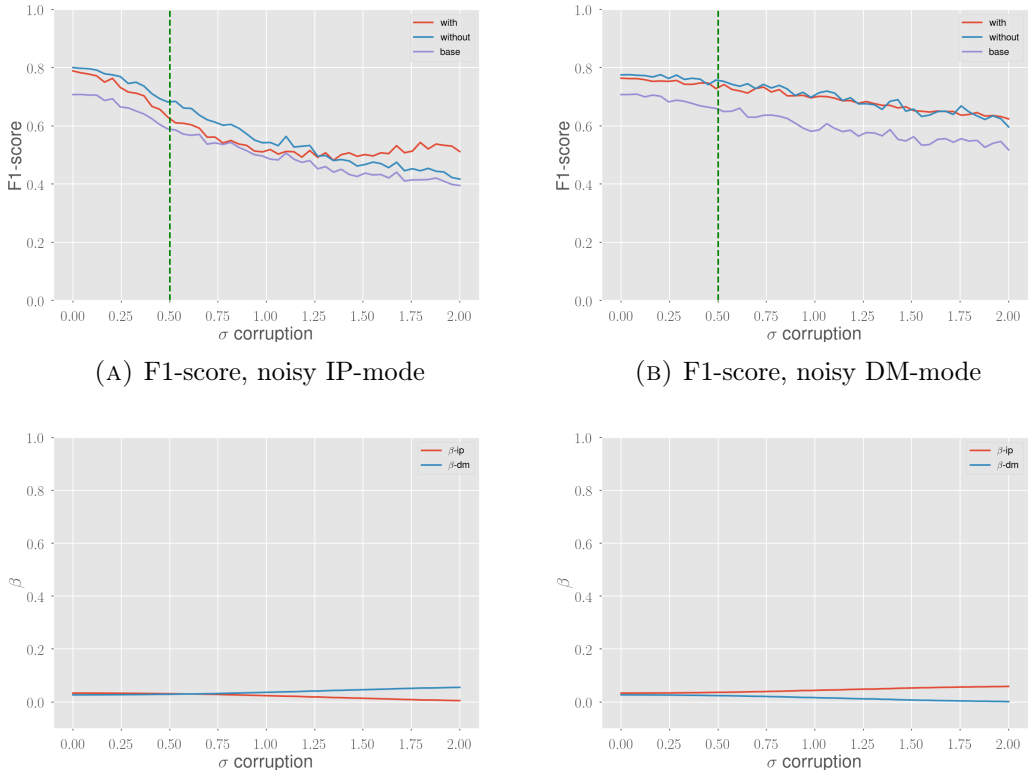


(C) attention score, noisy IP-mode



(D) attention score, noisy DM-mode

FIGURE 6.10: Noise generalisation of `model-with` ($\rho = 10^{-4}$, $\lambda_e = 10^{-2}$, $\lambda_c = 10^{-3}$). 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 6.11: 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 modalities depending on their general contributions. The literature review[1] discussed self-attention and crossmodal 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 crossmodal attention uses information from all the available modes. In combination with EMMA, we claim to have all the ingredients to construct a complete multi-modal network like humans. As a reminder, human's multi-modal attention consists of three different components: exogenous, endogenous and crossmodal attention. Attention is endogenous when we voluntary choose to attend to something whereas exogenous attention is triggered by the sudden onset of an unexpected event (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 considered as an equivalence to 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 crossmodal 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] Chapter 3

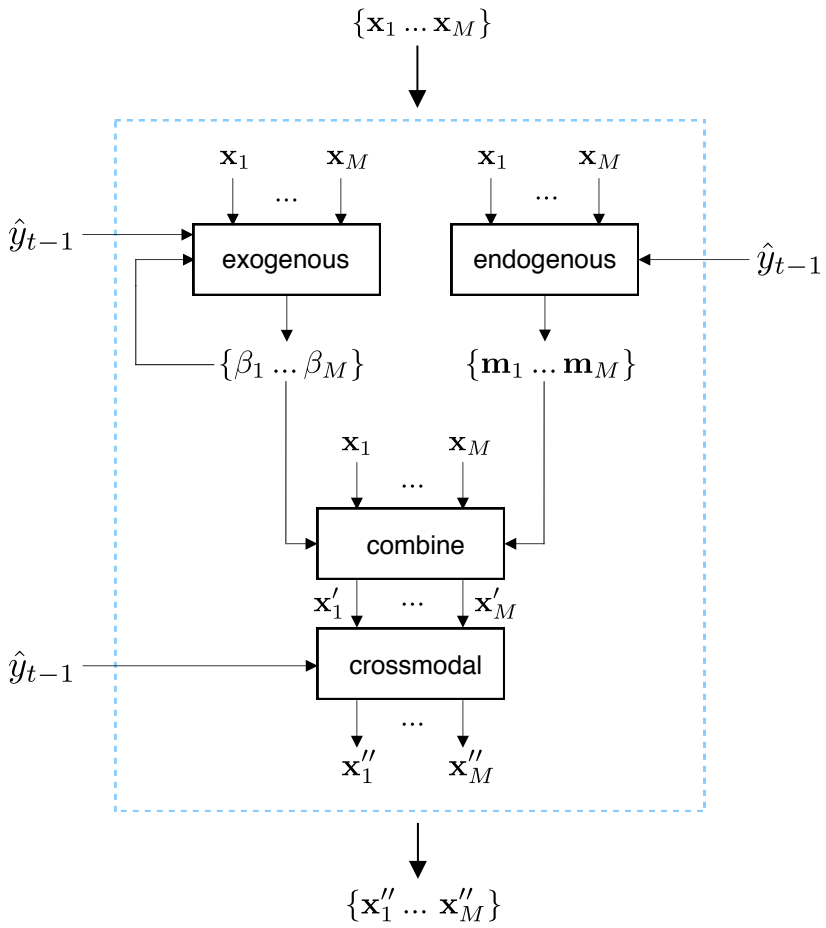


FIGURE 7.1: A possible architecture for a unified multi-modal attention

Chapter 8

Conclusion

Throughout this thesis, a new attention module was described and analysed. The module was designed to help multi-modal networks handle situations with failing modes. Our definition of failing modes included modes with out-of-distribution values, which to the best of our knowledge was never considered in previous research. Experiments showed that the shift of attention produced by our module was indeed beneficial for the prediction model in failing modes situations. Especially in cases where a mode contains more perturbations than in the training set (up to 20% improvement compared to standard data augmentation, on noise levels it was not trained on).

Our second contribution was to translate the concept of capacity from psychology to deep learning. This new insight led to the idea of constructing a simple regularizer which forces the module to limit the amount of information that can pass through it. This could potentially be generalized to any attention-based model, in order to control the degree of selectiveness of the information to put to the forefront.

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 adressed 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 is 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 dynamical rescaled and masked 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 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

[1] Several alternatives were proposed in Chapter 5

[2] Explanation from here

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.

[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

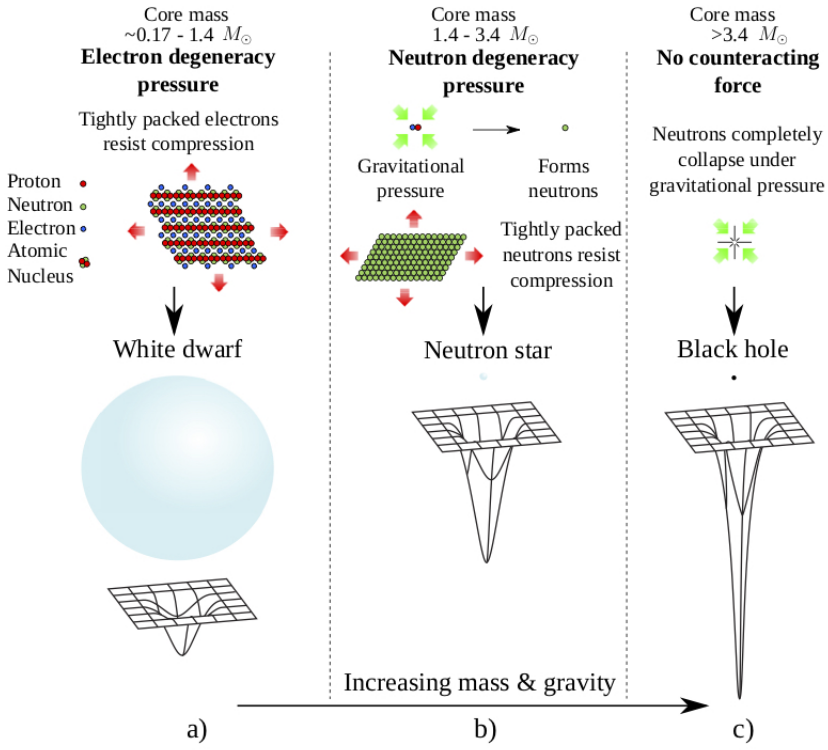


FIGURE A.1: Common evolutionary endpoints for main sequence stars. In a) electron degeneracy pressure prevents gravitational collapse, leading to the formation of a white dwarf star. In b) electron degeneracy pressure is no longer enough to counteract the inward force of gravity, however the gravitational 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.

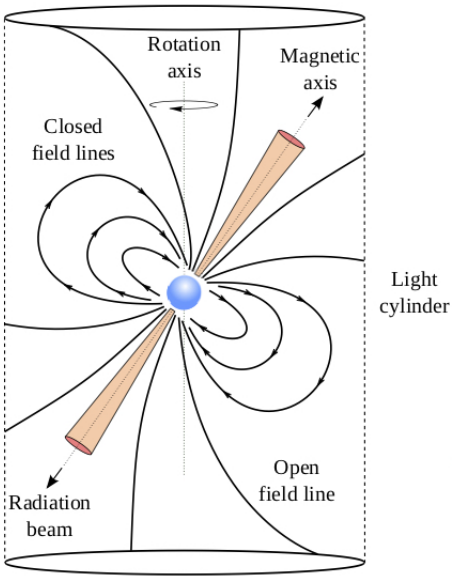


FIGURE A.2: Simplification of the lighthouse model of a radio pulsar, the pulsar is surrounded by a strong magnetic field comprising of open and closed field lines unable to close at the light cylinder. The light cylinder is an imaginary cylinder aligned with the pulsars rotational axis, that synchronously 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.

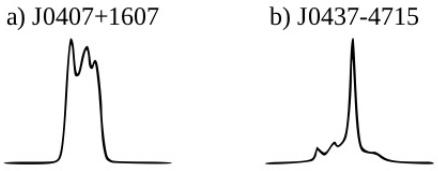


FIGURE A.3: Example pulse profiles of two separate pulsars. These profiles were adapted from those originally presented in (D.R. and M., 2005). *Image and caption copied from* (Lyon, 2016).

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.

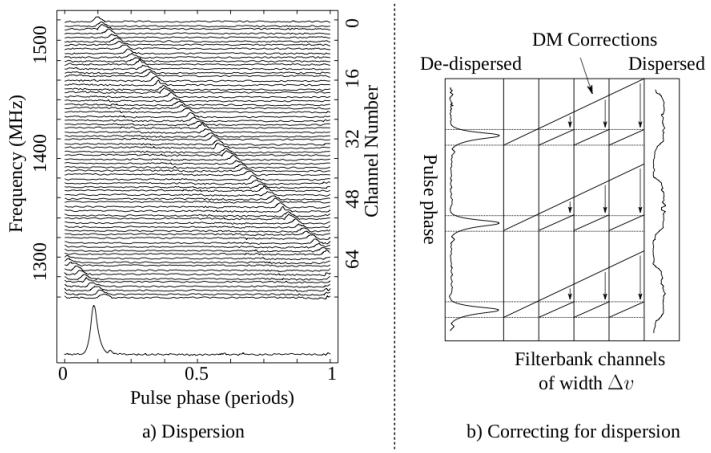


FIGURE A.4: An example of signal dispersion. Based upon diagrams originally presented in (D.R. and M., 2005). Plot a) shows how a signal is dispersed in time. Dispersion hides the true pulse shape and causes a lowering of the detected signal-to-noise. Plot b) shows the application of DM corrections to 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 II

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

B.2 Experiment III

detail intialization, describe architecture MMN

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:

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.

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].

Li, Yuxi (2017). "Deep Reinforcement Learning: An Overview". In: *arXiv e-prints*, arXiv:1701.07274, arXiv:1701.07274. arXiv: 1701.07274 [cs.LG].

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.

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 Hamiltonian 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].

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

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].