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

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

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

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

Acknowledgements

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

Thank contributions opens

Contents

Abstract iii

Acknowledgements v

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Proposed solution . . . 2
- 1.3 Contributions . . . 3
- 1.4 Thesis Outline . . . 3

2 Background 5
- 2.1 Machine Learning . . . 5
- 2.2 Deep Learning . . . 5
- 2.3 Deep Learning meets Physics . . . 5

3 Literature Review 7
- 3.1 Multi-Modality . . . 7
- 3.2 Attention mechanisms . . . 7
- 3.3 Conclusion . . . 9

4 Energy Estimation 11
- 4.1 Autoencoder . . . 11
- 4.2 Negative log-likelihood Estimators . . . 12
- 4.3 Experiment I . . . 14
- 4.4 Limitations . . . 16

5 Energy-based Multi-Modal Attention 17
- 5.1 Problem Statement . . . 17
- 5.2 General Framework . . . 18
- 5.3 From Potential to Modal energies (step 2) . . . 19
- 5.4 From Modal energies to Importance scores (step 3) . . . 20
- 5.5 From Importance to Attention scores (step 4) . . . 20
- 5.6 Training & Regularization . . . 22
- 5.7 Advantages . . . 24
- 5.8 Research questions . . . 25

6 Experiments & Results 27
- 6.1 Pulsar detection . . . 27
- 6.2 Corruption . . . 27
- 6.3 Experiment II . . . 27
- 6.4 Experiment III . . . 28

7 A Unified Model for Multi-Modal Attention 31

8 Conclusion **33**
8.1 Contributions 33
8.2 Research questions 33
8.3 Future work 33

A Dataset **35**

B Miscellaneous **39**
B.1 Integrability criterion 39
B.2 Gradient with respect to gamma 39

Bibliography **41**

List of Figures

1.1 Lidar & Camera view in self-driving cars 2
1.2 Multi-Modal model without EMMA (left) to with EMMA (right) 3

4.1 Two families of autoencoder architectures 11
4.2 Vectorial representation of undercomplete AE 12
4.3 Vectorial representation of overcomplete AE 13
4.4 Vector field circle manifold 13
4.5 Manifold generation of 200 samples 15
4.6 Vector fields on wave and circle manifold 15
4.7 Heatmap of estimators on wave and circle manifold 16

5.1 High-level view of a Multi-Modal Network 17
5.2 High-level view of a Multi-Modal Network & EMMA 18
5.3 Summary of main steps in EMMA 19
5.4 Input-output of Gibbs distribution for two different temperatures 21
5.5 Attention function 22
5.6 Summary of end-to-end training 26

7.1 A general model for Multi-Modal Attention 31

A.1 Evolutionary endpoints for main sequence stars 36
A.2 Lighthouse model of a radio pulsar 37
A.3 Pulse profiles of two separate pulsars 37
A.4 Signal Dispersion 38

Notation and Abbreviations

$\triangleq$	Is defined as
N	Number of samples
M	Number of modes
$\mathcal{L}$	Loss function
λ_c	Weight of capacity penalty
λ_e	Weight of energy penalty
Ω	Energy regularizer
e	Euler's number, base of the natural logarithm (2.71828)
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
ρ	Coldness parameter in Gibbs distribution

AE	**A**uto**e**conder
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
ML	**M**achine **L**earning
MLP	**M**ulti **L**ayer **P**erceptron
MMDL	**M**ulti **M**odal **D**eep **L**earning
MMN	**M**ulti **M**odal **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork

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 multiple amazing industrial application such as autonomous driving, early cancer detection, better machine translation, etc. However, industrials often face a difficult problem during the deployment of AI in the real-world: how to make sure these systems can handle unexpected situations correctly? New deep learning models are usually tested in conditions relatively similar to those in which the models where trained, leading to an overestimation of their robustness. Indeed, inputs of the model in real-world scenarios could be missing, subject to variable noises or unseen (e.g. a dog-cat image classifier receiving an image of a fish). These kind of inputs generally degrade the quality of predictions because the inputs do not contain enough information (missing) or the information cannot be processed (too noisy or unseen).

A possible remedy to the above problem is to use multiple modalities. Our experience of the world is multi-modal, i.e., we see objects, hear sound, feel the texture, smell odours, and taste flavours. A modality, also called mode, refers to a representation with very different statistics than the other modalities in the setting (Baltrušaitis, Ahuja, and Morency, 2019). Multi-Modal Deep Learning (MMDL) is used in the hope is that the information carried by each mode is additive, as a result the model has more information to learn to make more accurate and more robust predictions. Impressive works has been done in this domain. Authors in (Afouras et al., 2018) created an AI performing speech-to-text using not only audio of the speaker but also video of the lip movements, achieving state-of-the art results on unconstrained natural language sentences from British television. They also investigated to what extent lip reading is complementary to noisy audio signal, showing robustness was indeed improved. In a like manner, research in self-driving cars started exploring ways to combine sensorial inputs from wide angle cameras and LIDAR[1] sensors for road detection (Caltagirone et al., 2018). An example of the representation obtained by these sensors is shown in Figure 1.1. 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 and provide accurate distance measurements but have a limited range. Fusing sensorial information with deep learning is done to get the best of both sensors. Despite the progress made, noisy and unseen modes are still fed into the mode and thus potentially perturbing the predictions. Not to mention that combining modalities is

[1] Laser Detection and Ranging

challenging due to modalities having different quantitative influence over the prediction output and having different levels of noise.

On the other hand, humans seem to handle these situations robustly on a daily basis. A famous example showing this 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. The visual stimuli is then processed by the brain, now able to more easily extract the speaker from the audio. Similarly, our attention is shifted from vision to touch when we are wandering in a room where the lights suddenly switch off. These examples demonstrate that humans are able to handle noisy, missing and unseen modes by focusing its attention smoothly on the most relevant mode(s). In this work I introduce a novel attention mechanism, named *Energy-based Multi-Modal Attention* (EMMA), mimicking human's multi-modal attention. More specifically, EMMA decides how much attention to devote to each mode, such that the relevant information is kept while masking out the perturbations.

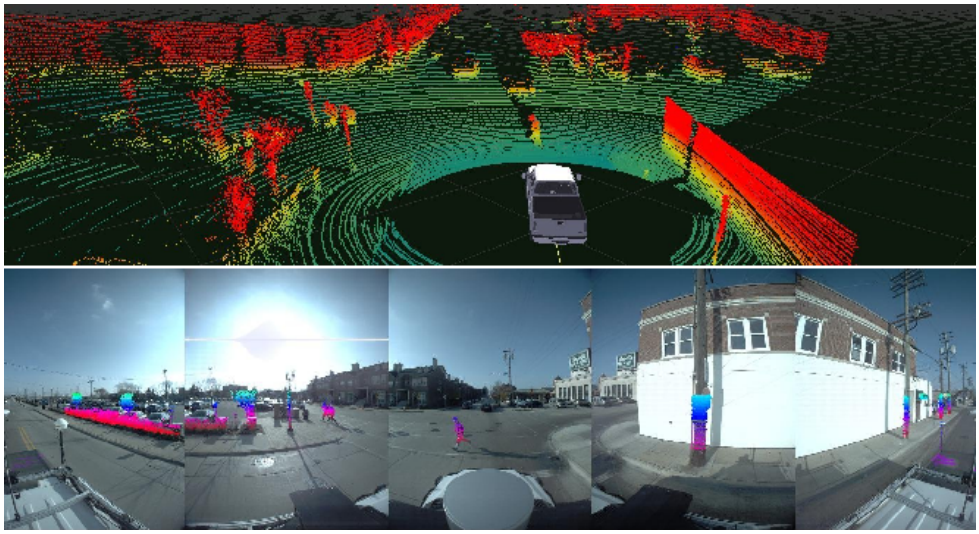


FIGURE 1.1: Same environment, different modes (top: LIDAR view, bottom: camera view)

1.2 Proposed solution

An example of how EMMA is used is illustrated in Figure 1.2. The attention module is inserted in front of the model, focusing attention on the best modes of the sample. The less outlying a mode, the more information of the mode EMMA will let pass by. We denote outlyingness as a general notion for noisiness, missing and unseen values. Additionally, EMMA takes into account how each mode contributes differently to the accuracy, and the kind of information between a pair of modes (redundant, complementary, conflicting).

Hypothesis

EMMA is constructed under the assumption that if a mode is outlying, at least one other mode can be found that is not outlying, which is thus able to "take over".

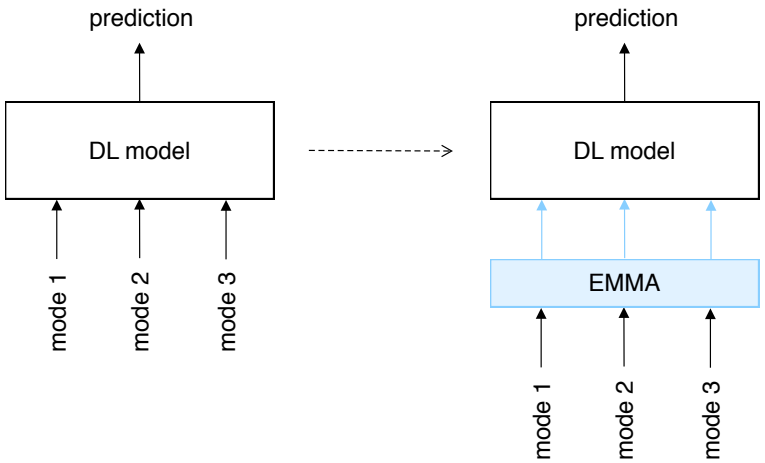


FIGURE 1.2: Multi-Modal model without EMMA (left) to with EMMA (right)

Software Implementation

All the implemented models and experiments are available at this repository[2], with a wiki explaining how to run the experiments. Moreover, the main framework used regarding the Machine Learning part is PyTorch[3].

1.3 Contributions

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

Contribution 1: a module mimicking human's endogenous attention. A new Deep Learning attention mechanism, EMMA, has been developed. It is based on Energy models (Osogami, 2017) to measure the outlyingness of modes and thereupon decide where to focus attention. TODO: present main results obtained.

Contribution 2: a proposal for a unified model for multi-modal attention. In Chapter 2, a review of the literature from a Psychological standpoint on attention reveals current multi-modal attention mechanisms in DL are incomplete. In Chapter 7, an architecture for a general multi-modal attention is proposed combining EMMA and the current multi-modal attentions.

Contribution 3: a constraint permitting to link capacity as in (Kahneman, 1975) **and deep learning attention functions.** A common parametric function for Attention in DL is slightly transformed such that it resembles more human attention. The transformation can easily be generalized to most attention functions.

1.4 Thesis Outline

The remainder of this work is organised as follows.

Chapter 2 explains the background this work is based upon.

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

Chapter 3 reviews the existing literature of Multi-Modal Deep Learning, putting forward their limitations.

Chapter 4 describes two ways of estimating the outlyingness of data and how these are derived.

Chapter 5 presents the ideas and architecture of the Energy-based Multi-Modal Attention module (Contribution 1 & 3).

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

Chapter 7 from current attention mechanisms and the developed EMMA module, a more general multi-modal attention architecture is discussed (Contribution 2).

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

Chapter 2

Background

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

2.1 Machine Learning

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

2.2 Deep Learning

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

2.3 Deep Learning meets Physics

- Negative log likelihood. Ising model (thermodynamics, stat mechanics). Gibbs distribution. Explain how relate to outlying (high E, low p). EBM = high everywhere except on data manifold.
- Boltzmann. MC training. Contrastive divergence. Explain why difficult (see paper of potential energy, good explanation!!). Thus we will explore approximators.

The potential energy article has some paragraphs. Also explain what the score is.

Boltzmann distrib Energy-based models Alleen Mara!!! Deep Structured Energy Based Models for Anomaly Detection (Zhai et al., 2016): ....

Chapter 3

Literature Review

Arguments: crossmodal most of them, specific (not easily inserted to any model without architecural modifications) and implicit (not interpretable and does not burden the model) and nothing against unseen/missing. Multimodal, single mode, weakness, strength, citations. And how they differ from this work. Build up. others attend only a part of the mode. Show the gap, no current attention is specifally designed to in general add real robustness against very outlying data. https://www.scribbr.com/dissertation/literature-review/

In this chapter we review the recent works in attention and more specifically multi-modal attention...

3.1 Multi-Modality

More general than sensorial, tasks, general representations, gates of informations, robot sensors. Give examples.

- Multimodal Machine Learning: A Survey and Taxonomy (Baltrušaitis, Ahuja, and Morency, 2019): ....
- Learn to Combine Modalities in Multimodal Deep Learning (Liu et al., 2018): ....

3.2 Attention mechanisms

Psychological perspective

selection stimuli multi-modal endo,exo,cross then viewed selection, capacity, different models are not exclusive as we will see. in chapter 7 we show one model combining them all.

Attention Is Amplification, Not Selection Peter Fazekas and Bence Nanay

Kahnemann book

Attention refers to those processes that allow for the selective processing of incoming sensory stimuli. Mechanisms of attention help us to prioritize those stimuli that are most relevant to achieving our current goals and/or to performing the task at hand. The term 'attention' is used to describe those processes that give rise to a temporary change (often enhancement) in signal processing. This change will often be manifest in only a subset of the stimuli presented at any time. Researchers have attempted to distinguish attentional effects from other temporary changes in the efficiency of information processing, such as those induced by changes in arousal and/or alerting. These latter processes can be contrasted with attention both on the grounds of their non-selectivity (i.e., increased arousal or alertness tends to influence the processing of all incoming stimuli; that is, its effects are stimulus non-specific), and behaviourally, by the fact that while alerting, for example, can speed up a person's response it also tends to result in increased errors

(i.e., perceptual sensitivity is not enhanced). (More recently, however, it has become somewhat more difficult to distinguish between attention, alerting, and arousal. This is both because researchers have started to argue that certain kinds of attention effect may only lead to a speeding of participants' responses (i.e., without any concomitant change in perceptual sensitivity; see Prinzmetal et al., 2005a, b, 2009) and also because there is some evidence of the selectivity of alerting effects (e.g., in terms of the modality affected; e.g., Posner, 1978).) Attention can either be oriented endogenously or exogenously: 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, say, 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 (i.e., involuntary) by the sudden onset of an unexpected event, such as when someone calls our name at a noisy cocktail party, or when a mosquito suddenly lands on our arm. However, our attention can also be captured by intrinsically salient or biological significant stimuli. Attended stimuli tend to be processed more thoroughly and more rapidly than other potentially distracting ('unattended') stimuli (Posner, 1978; Spence & Parise, 2010). Although attention research has traditionally considered selection among the competing inputs within just a single sensory modality at a time (most often vision; see Driver, 2001, for a review), the last couple of decades have seen a burgeoning of interest in the existence and nature of any crossmodal constraints on our ability to selectively attend to a particular sensory modality, spatial location, event, or object (Spence & Driver, 2004). In fact, crossmodal interactions in attention have now been demonstrated between most combinations of visual, auditory, tactile, olfactory, gustatory, and even painful stimuli (Calvert et al., 2004).

Deep Learning perspective

subsec dl, intra-modal speech recog, different types Normal attention Multi-Modal attention

Tutorial: tutorial on attention soft vs hard attention other tutorial attention

- Attention Is All You Need (Vaswani et al., 2017): ...
- Attention, please! A Critical Review of Neural Attention Models in Natural Language Processing (Galassi, Lippi, and Torroni, 2019): ....
- Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation (Ephrat et al., 2018): ....
- Input Combination Strategies for Multi-Source Transformer Decoder (Libovický, Helcl, and Mareček, 2018): ....
- Watch, Listen, and Describe: Globally and Locally Aligned Cross-Modal Attentions for Video Captioning (Wang, Wang, and Wang, 2018): ....
- Cross-Modal Attentional Context Learning for RGB-D Object Detection (Li et al., 2019): ....
- Stacked Cross Attention for Image-Text Matching (Lee et al., 2018): ....
- Modeling Latent Attention Within Neural Networks (Grimm et al., 2017): ....
- Noise-tolerant Audio-visual Online Person Verification using an Attention-based Neural Network Fusion (Shon, Oh, and Glass, 2018): ....
- deep audio-visual speech recognition (Afouras et al., 2018): ....

3.3 Conclusion

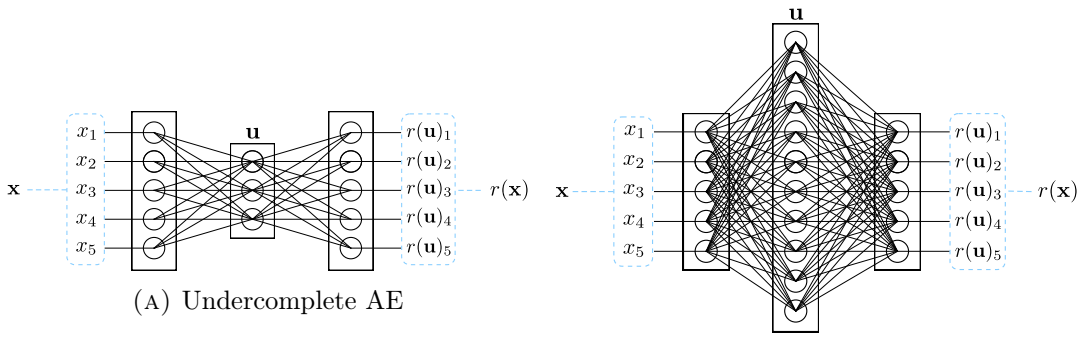
Chapter 4

Energy Estimation

As mentioned earlier, the module EMMA takes the outlyingness of each mode into account. This helps the module determine which modes are important and which are not. The negative log-likelihood (NLL) gives us a measure of outlyingness, having values close to zero for likely data and high values everywhere else. Energy-based models are trained to approximate the NLL but are difficult to train[1]. This chapter discusses two alternative estimators of NLL obtained training an autoencoder. We will start by explaining the purpose and intuition behind autoencoders. Next, we describe the estimators and how they are derived. And to conclude, a simple experiment is performed to compare the estimators on generated toy data.

4.1 Autoencoder

Autoencoders (AE) are models trained to reproduce the input to their output. They are 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, $\mathcal{L}_{\text{MSE}} = \|r(\mathbf{x}) - \mathbf{x}\|_2^2$. But why is it useful to train a model to copy its input? Well to answer this question we need to distinguish two families of autoencoders: the undercomplete and overcomplete autoencoders.



(A) Undercomplete AE

(B) Overcomplete AE

FIGURE 4.1: Two families of autoencoder architectures

[1] See Section 2.3

Undercomplete

An autoencoder is undercomplete if 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 keep 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 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 $r(\mathbf{x}) - \mathbf{x}$.

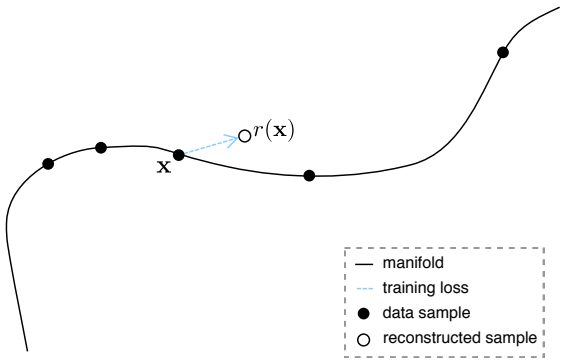


FIGURE 4.2: Vectorial representation of an undercomplete reconstruction process

Overcomplete

Conversely, an overcomplete AE has more hidden units than its input/output layer (see Figure 4.1b). Straightforwardly, the model can learn to copy the 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 $\tilde{\mathbf{x}} = \mathbf{x} + \epsilon$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, 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, $r(\tilde{\mathbf{x}}) - \tilde{\mathbf{x}} \to -(\tilde{\mathbf{x}} - \mathbf{x})$.

4.2 Negative log-likelihood Estimators

The norm of the reconstruction error, $\|r(\tilde{\mathbf{x}}) - \tilde{\mathbf{x}}\|_2^2$, is sometimes used in the Machine Learning community as a way to detect oultiers. The motivation is that samples far away of the data manifold will have larger reconstruction error than samples close to the manifold. However, as we will see later on in the experiments, the reconstruction error is not a good estimator of the NLL and can lead to false positives.

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

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

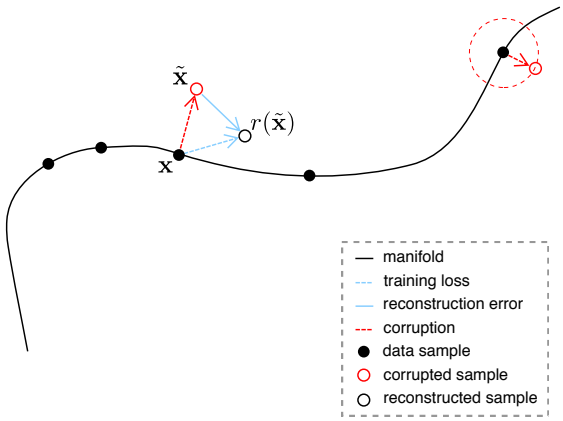


FIGURE 4.3: Vectorial representation of an overcomplete reconstruction process

To put it differently, the reconstruction error points in the same direction as the corresponding most likely datapoint. To illustrate thisn a DAE is trained on generated circle manifold (more details about the experiment in Section 4.3). As we can see below, the vector field of the reconstruction error does indeed point towards the data manifold.

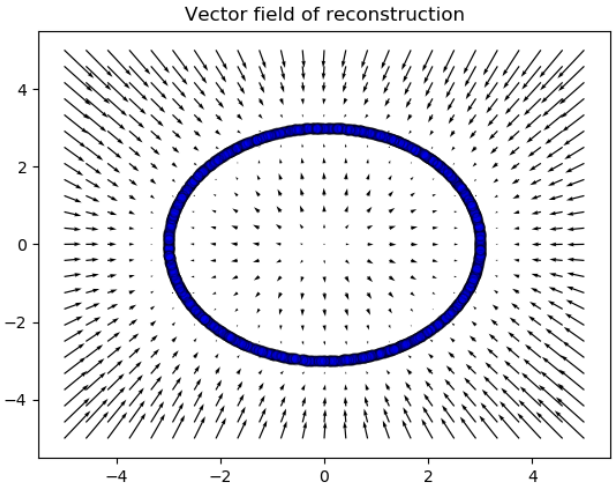


FIGURE 4.4: Vector field of reconstruction error on circle manifold. No corruption is applied at test time, the reconstruction error vector is simply the output minus the input.

In (Kamyshanska and Memisevic, 2014), authors observed that using tied weights ($W_f = W_g^T$), makes the integrability criterion[2] 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 and $W = W_f$. The vector field under those circumstances can be expressed as a gradient of a scalar field $-\Psi$, such that $r(\tilde{\mathbf{x}}) - \tilde{\mathbf{x}} =$

[2] See Section **??**

$-\partial\Psi(\tilde{\mathbf{x}})/\partial\tilde{\mathbf{x}}$. In analogy to physics, we can interpret the vector field 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 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 $g = W^T\mathbf{x} + \mathbf{b}_g$

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

The intermediate steps between (4.5) and (4.7) are detailed in (Kamyshanska and Memisevic, 2014). In this work we will only use the sigmoid activation function, we solve

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

where $W_{\cdot k}^T$ is the k^{th} column of W^T, and b_k^f the k^{th} element of $\mathbf{b}_f$. The consequences of neglecting the constant will be discussed in the next chapter. To sum up, we now have two estimators of the negative log-likelihood being derived from a trained denoising autoencoder: the reconstruction error and the potential energy. The latter being more theoretically grounded.

4.3 Experiment I

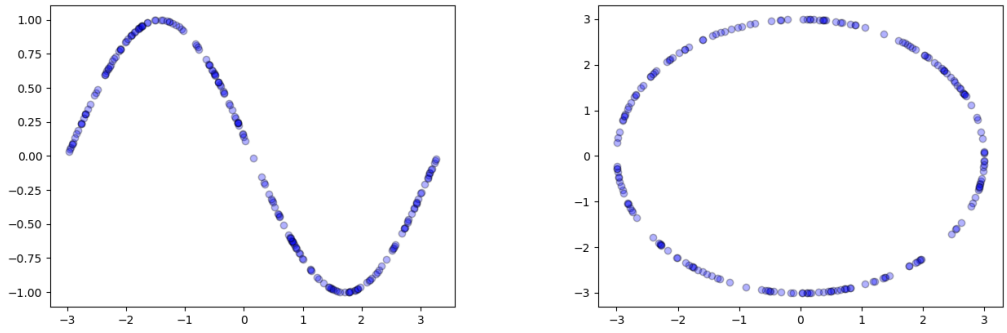
In this experiment we are going to generate two simple data manifolds. A denoising autoencoder will then be trained on each manifold. From those autoencoders we will compute the estimators and compare them.

Manifolds

For no particular reason, we decide to generate a set of N samples $\mathbf{x} \in \mathbb{R}^2$ in the form of a wave and a circle. Both manifolds are generated in two steps. First, N samples are randomly selected in an interval $[0, 2\pi]$. The N samples are written $\mathbf{t}$, 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.



(A) Wave (B) Circle

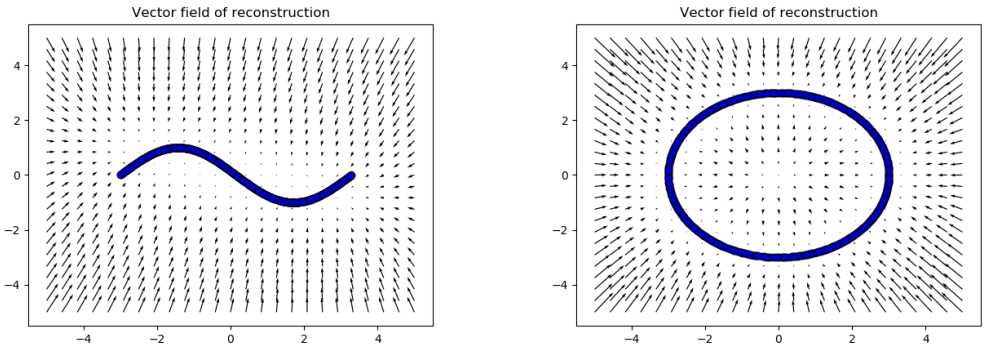
FIGURE 4.5: Manifold generation of 200 samples

Setup

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

Results

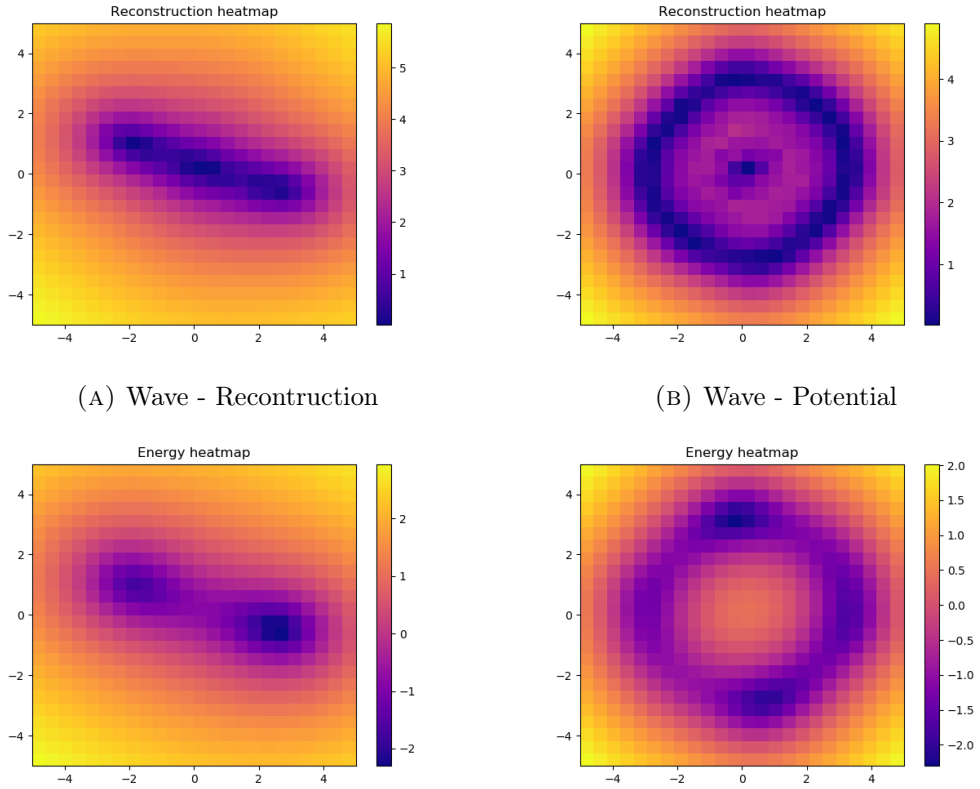
The autoencoders for both manifolds converged quickly. We observe on Figure 4.6 that the vector fields of the reconstruction error are directed towards the manifolds. We can think of the manifold as sinks in the vector field. Notably, observe the presence of a source in the center of the manifold (see Figure 4.6b).



(A) Wave (B) Circle

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

Next, the estimators are computed and plotted onto heatmaps (see Figure 4.7). Not surprisingly, the estimators have low values in the neighbourhood of the manifold and are high everywhere else. More interestingly, a clear difference is observed between the estimators on the circle manifold. This can be explained by the fact that the norm of the reconstruction error at the source is small just as on the manifold itself. A small additional detail is that some potential energies are negative, which is due to the neglected integration constant in Equation (4.7).



(A) Wave - Recontruction

(B) Wave - Potential

(C) Circle - Recontruction

(D) Circle - Potential

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

4.4 Limitations

TODO. The intergrability criterion is guaranteed to be satisfied when using tied weights, this assumption is valid only because there are no more than one hidden layer. Adding more non-linear layers breaks the criterion. Potential energy as it is currently defined cannot directly be applied on convolutional AE, LSTM AE, stacked AE.

Alternative bengio, and others. Future work could investigate if possible to find other potential energy or add constraint on CNN, .. to make it work. Or simply test with current definition and see if it is that bad.

Scope is to make module that uses any measure of outlyingness. Not to find the best one, somebody else work.

Chapter 5

Energy-based Multi-Modal Attention

In the Literature review (Chapter 3) we concluded that previous work in multi-modal deep learning has been mostly focused on trying to increase the accuracy of the prediction. Few research has explicitly addressed the question of how using multiple modalities can improve the robustness. Inspired by how humans handle multiple senses robustly, I created a novel generic module that can easily be inserted into every trained multi-modal architecture. This chapter describes the ideas and architecture of the Energy-based Multi-Modal Attention module.

5.1 Problem Statement

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 (e.g. images and sounds). The multi-modal network will be abbreviated as MMN. The model tries to make predictions $\hat{y}$ as close as possible to the groundtruth y (see Figure 5.1). The internal architecture of the MMN is often structured as a many-to-one encoder-decoder, where each encoder extracts features and the decoder is in charge of merging those together. Nonetheless, the EMMA module is not constrained to any specific MMN architecture.

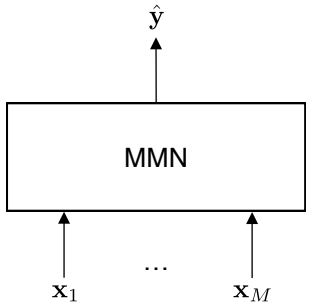


FIGURE 5.1: High-level view of a Multi-Modal Network

Current research in MMDL is mostly motivated by how to make use of the information gain of adding an extra mode to make better predictions. In this work, we want to leverage this same information gain to improve the robustness. We start from the assumption that for a sample where one mode i is outlying, it is likely to find at least another mode j who is not outlying. This permits us to think that if we find a way in that case to shift the attention from mode i to mode j, the performance would be better off. This is done by computing an importance score α_i for each mode i between zero and one, giving us the relative importance. The importance scores α_i are a measure of how valuable each mode i is taking into account the outlyingness of all the modes. From

those, we determine the attention scores β_i, representing the quantity of information that can pass through. Each mode is then multiplied by its respective attention score (see Figure 5.2). We justify later on why we do not directly multiply be the importance score instead. TODO: explain how missing values problem is implicitly solved

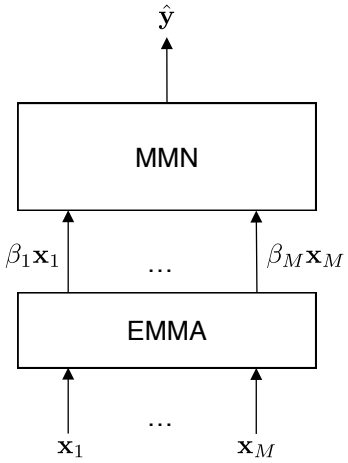


FIGURE 5.2: High-level view of a Multi-Modal Network with the EMMA module

There are two ways of interpreting this solution. First, EMMA can be seen as a sort of gate filtering perturbations out. Indeed, outlying 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. The model then learns to make more robust predictions based on the extra inputs β_i.

In Section 5.2 we describe the different steps along with their main objectives. Hereafter, each section will correspond to a specific step and will detail how it works. Next, the training of the module is explained along with some regularizers. The chapter ends with an enumeration of the main research questions that needs to be addressed in the experiments.

5.2 General Framework

As we just saw, the model needs to compute how important each mode is. To do this let us start by introducing three intrinsically tied properties describing the importance of a mode i:

- *relevance*: how much does mode i help improve the the accuracy?
- *outlyingness*: is the current sample much different from the training set? How much will it import the predictions?
- *coupling*: does mode j has a strong influence on mode i? If so in which way? Does mode i and j carry complementary or redundant information?

We define the modal energy E_i for a mode i as an embedding of the three properties. Modal energies are constructed as learnable parametric functions of potential energies,

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

The function f is able to capture the relationship between relevance and outlyingness. Because f is optimized with a loss on the predictions and is also a function of the potential Ψ_i. The role of the function g is to learn the optimal coupling between modes. Modal energies are then normalized to the importance scores. Which is in some kind equivalent of going from a measure of absolute importance to one of relative importance.

Attention is viewed in psychology as a selection process between senses or more generally, modes. Deep Learning research regarding attention is in majority based on this view. In contrast, the famous economist and psychologist Daniel Kahneman sees attention as a shared resource with a limited capacity being allocated between the modes (Kahneman, 1975). We mimic the latter by slightly modifying a common attention function. This is done with the intention of improving the interpretability of EMMA (see Section 5.5). In this work, we use attention to decide how much information of a certain mode will pass.

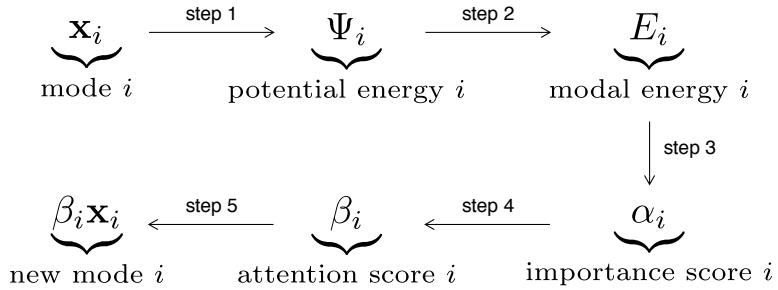


FIGURE 5.3: Summary of main steps in EMMA (step 2, 3 and 4 are detailed in the following sections, step 1 was explained in Chapter 4)

5.3 From Potential to Modal energies (step 2)

A problem overlooked so far is that we neglected the integration constant in the computation of the potential energy[1]. The potential can therefore take negative values. This is a problem because evaluating the gradient during the backpropagation involves taking a logarithm of Ψ_i[2], which is undefined for negative values. This problem is corrected by lowing the potential Ψ_I to Euler's number e as

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

With $\Psi_i^{(\min)}$ the lowest value of Ψ_i in the training set. This correction avoids undefined values ($\Psi_i' \geq 0$) but also exploding gradient ($\Psi_i' \geq e$). The reason a max-operator is used is because lower energy values than $\Psi_i^{(\min)}$ can occur during inference.

The *self-energy* of mode i is defined as an energy capturing the outlyingness and the relevance of the mode. We write,

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

The parameters w_i, b_i are trained via a loss function on the predictions, thus adding the influence of the relevance. It also enables the model to face potentials on possibly very different scale, caused by the proportionality in Equation (4.7).

[1] Equation (4.7)
[2] See Appendix B

Furthermore, the *shared energy* of mode j on i is constructed from the self-energies as follows

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

Adding the constraint $\gamma_{ij} = \gamma_{ji}$, the model can now discover the optimal coupling between the modes. It if learns a γ_{ij} close to zero, mode i and j will influence each other much more than a γ_{ij} near to one. The parameter γ_{ij} learns the degree of coupling in the spectrum from strongly coupled ($\gamma_{ij} = 0$) to independent ($\gamma_{ij} = 1$). The direction of coupling between mode i and j are learned by the weights w_{ij} and w_{ji}. For a positive w_{ij}, an increase in self-energy e_j causes an increase in e_{ij}. Whereas if w_{ij} is negative, an increase in e_j leads to a decrease in e_{ij}. The weights w_{ij} are not imposed to be equal to w_{ji}, such that modes can influence each other asymmetrically. This assymetry is justified by the following example: a multi-modal problem with three modes A, B and C. Imagine the case where if mode A is missing, it is optimal that mode B takes over. But if B is missing it is optimal for C to take over. This can only be modelled with asymmetry.

Finally, the *modal energy* of a mode is the sum of its self-energy and the shared energies with all the other modes:

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

We can recognize Equation 5.1. This gives us also the *total energy*, $E_{\text{total}} = \sum_i E_i$, offering an intuitive way to measure how uncertain the model is about its predictions.

5.4 From Modal energies to Importance scores (step 3)

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

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

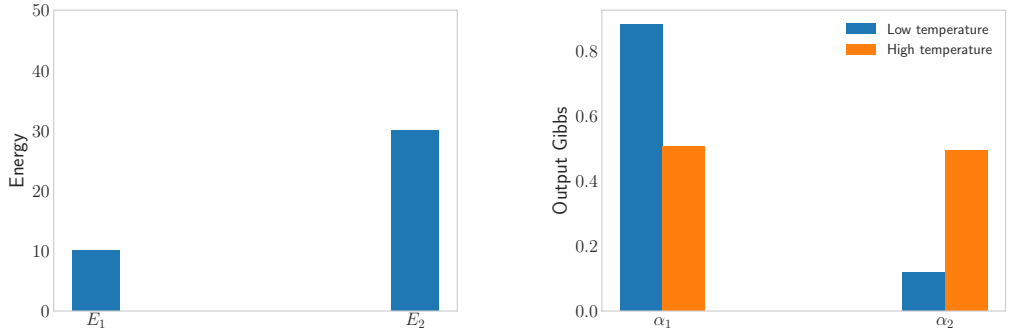
This guarantess the scores to be normalized and summing up to one. A mode i will be said to be important if its score is close to one (low modal energy E_i). The hyperparameter ρ represents the coldness, the inverse of the temperature. It controls the entropy of the importance scores distribution. At 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 (see Figure 5.4). Careful tuning is thus necessary.

5.5 From Importance to Attention scores (step 4)

The attention scores are given by

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

The hyperbolic tangent adds non-linearity while the gain g_a and bias b_a permits the model to control the threshold and capacity (see Figure 5.5). The latter two concepts are detailed below.



(A) Energies (B) Importance scores

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

Energy threshold

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

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

where $E_{\text{threshold}}$ represents the maximal amount of energy a mode i is allowed to have in order to pass. As be seen the gain and bias control the threshold. Nevertheless, the value of the partition function Z makes the threshold dynamic. The partition function will be higher if the modes are more outlying, thus diminishing the thresholds. In other words, EMMA adapts the selectiveness with respect to the quality of the data. TODO: This can be linked more arousal link. Notice that the influence of the temperature (ρ^{-1}) is non-trivial to analyse, because Z also depends on ρ.

Capacity

There is nothing new about using an hyperbolic tangent as an attention mechanism. The difference however resides in how the linear combination is restricted. A very common attention function is written as $\tanh(\mathbf{W}\alpha + \mathbf{b})$ whereas we have $\tanh(g_a\mathbf{I}\alpha - b_a\mathbf{u})$ with $\mathbf{u}$ the unit vector $(1 \dots 1)^T$. We argue the latter mimics better human's crossmodal attention. The capacity in psychology is viewed as the amount of resource that can be allocated. This can be translated in our case as,

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

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

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

we can write

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

If the capacity is too low, there is no sufficient information passed to the MMN to make predictions and thus the accuracy drops. On the other hand, if the capacity is too high, too much perturbations will pass leading to bad performances. The module learns the optimal trade-off. Observe that the concept of capacity can also be applied to $\tanh(\mathbf{W}\alpha + \mathbf{b})$. In that case, each mode would have a different capacity. This could make EMMA more expressive, but the importance scores would be less meaningful. Another advantage of having only one capacity is that it is easier to control. In the next section, we show a simple regularizer giving us some control on the capacity. The interest of doing this, is the idea that minimizing the capacity would allow to gain more robustness againts unseen situations at the cost of some accuracy. This regularizer is played along with in the experiments (see Chapter 6), giving us some interesting insight.

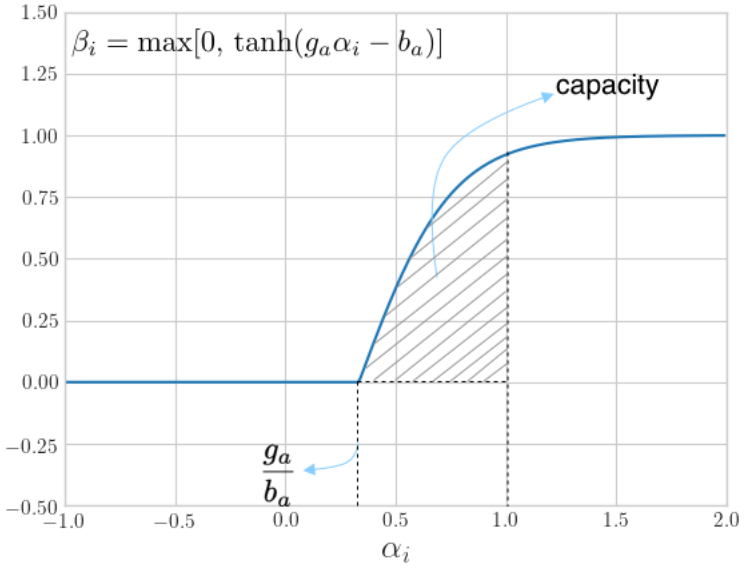


FIGURE 5.5: Attention function, the max-operator generalizes the attention function to cases where $\alpha \in \mathbb{R}$

5.6 Training & Regularization

The end-to-end model (MMN & EMMA) is trained in two phases (see Figure 5.6). First, all the autoencoders are trained, one per mode. Once trained, 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.

The motivation of EMMA is not only to improve predictions on the test-set. But also to see if it is able to handle new situations (robustness generalization) as expected.

We also want EMMA to keep a certain level of interpretability. For those reasons we introduce two regularizers. The first one controls the capacity, where λ_c can be positive/negative depending on if we want to maximize/minimize the capacity.

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

Additionally, regularizing the energy (λ_e) is done for interpretability purposes. If the parameters of modal energies are optimized only regarding the predictions ($\mathcal{L}$), we could have 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.

Energy regularization

Let $\theta = \{\mathbf{\Gamma}, \mathbf{W}, \mathbf{b}\}$ be the set of all the parameters of step 2. The effect of the regularizer on this set of parameters with Gradient Descent is written

$$\theta' \leftarrow \theta + \epsilon \lambda_e \nabla_\theta \Omega \tag{5.13}$$

Remember the objective, we want this update to cause lower/higher modal energies E_i for low/high potential energies. To verify this let us compute $\nabla_\theta \Omega$,

$$\nabla_\theta \Omega = \sum_{k=1}^{M} \xi_k \nabla_\theta \log(\alpha_k) \tag{5.14}$$

The gradient of the logarithm can be developed as

$$\begin{aligned} \nabla_\theta \log(\alpha_k) &= \nabla_\theta \log \left[\frac{e^{-\rho E_k}}{Z} \right] \\ &= \nabla_\theta(-\rho E_k) - \nabla_\theta \log \sum_{l=1}^{M} e^{-\rho E_l} \\ &= -\rho \nabla_\theta E_k - \frac{\sum_{l=1}^{M} \nabla_\theta e^{-\rho E_l}}{\sum_{l=1}^{M} e^{-\rho E_l}} \\ &= -\rho \nabla_\theta E_k + \rho \frac{\sum_{l=1}^{M} e^{-\rho E_l} \nabla_\theta E_l}{\sum_{l=1}^{M} e^{-\rho E_l}} \\ &= \rho \left[-\big(1 - \frac{e^{-\rho E_k}}{Z}\big) \nabla_\theta E_k + \sum_{l \neq k}^{M} \frac{e^{-\rho E_l}}{Z} \nabla_\theta E_l \right] \\ &= \rho \left[-\big(1 - \alpha_k\big) \nabla_\theta E_k + \sum_{l \neq k}^{M} \alpha_l \nabla_\theta E_l \right] \end{aligned} \tag{5.15}$$

We go further by expressing the equation above with respect to the subset of parameters $\theta_i = \{[\gamma_{ik}, w_{ik}]_{k=1}^{M}, w_i, b_i\}$:

$$\nabla_{\theta_i} \log(\alpha_k) = \begin{cases} -\rho(1 - \alpha_i) \nabla_{\theta_i} E_i, & \text{if } i = k \\ \rho \alpha_i \nabla_{\theta_i} E_i, & \text{if } i \neq k \end{cases} \tag{5.16}$$

The gradient of the regularizer can now be computed by plugging Equation (5.16) into the summation (5.14). Let M' be the number of uncorrupted modes. We obtain for an uncorrupted mode i,

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

and for a corrupted mode i,

$$\nabla_{\theta_i}\Omega = \xi_-\big[-\rho(1-\alpha_i)\nabla_{\theta_i}E_i\big] + \big[M'\xi_+ + (M-M'-1)\xi_-\big]\alpha_i\rho\nabla_{\theta_i}E_i \tag{5.18}$$

Substituting ξ_k, we can summarize Equations (5.17) and (5.18) as

$$\boxed{\nabla_{\theta_i}\Omega = -\big[(M-2M')\alpha_i + \xi_i\big]\rho\nabla_{\theta_i}E_i} \tag{5.19}$$

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

$$\theta_i' \leftarrow \theta_i - \epsilon\lambda_e\rho\xi_i\nabla_{\theta_i}E_i \quad \text{with} \quad \lambda_e \in \mathbb{R}^+ \tag{5.20}$$

Ignoring the second-order effects of the Taylor expansion, we can conclude from the equation above that the regularizer will update the parameters such that modal energies E_i increase/decrease for corrupted/uncorrupted modes i.

In analogy, if M is uneven we have

$$\theta_i' \leftarrow \begin{cases} \theta_i - \epsilon\lambda_e\rho(1-\alpha_i)\nabla_{\theta_i}E_i, & \text{if } i \text{ is uncorrupted} \\ \theta_i + \epsilon\lambda_e\rho(1+\alpha_i)\nabla_{\theta_i}E_i & \text{otherwise} \end{cases} \tag{5.21}$$

The principle is the same as in the even case with an additional effect: the correction will be proportional to the error. High energies that have to 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 (5.13) becomes

$$\theta_i' \leftarrow \theta_i - \epsilon\lambda_e\rho(\alpha_i + \xi_i)\nabla_{\theta_i}E_i \tag{5.22}$$

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

5.7 Advantages

Listed below are the advantages of using EMMA instead of standard data-augmentation techniques.

- The generic design of EMMA permits it to be easily added to any architecture of multi-modal network
- The burden on the multi-modal network is reduced, it only has to learn to make good predictions from the received information
- Interpretability is increased, notably the uncertainty on the predictions as we will see in Section 6

5.8 Research questions

- Does EMMA increase the robustness compared to data augmentation techniques?
- Is the use of the two regularizers experimentally validated?
- Is the end-to-end model more interpretable with EMMA?

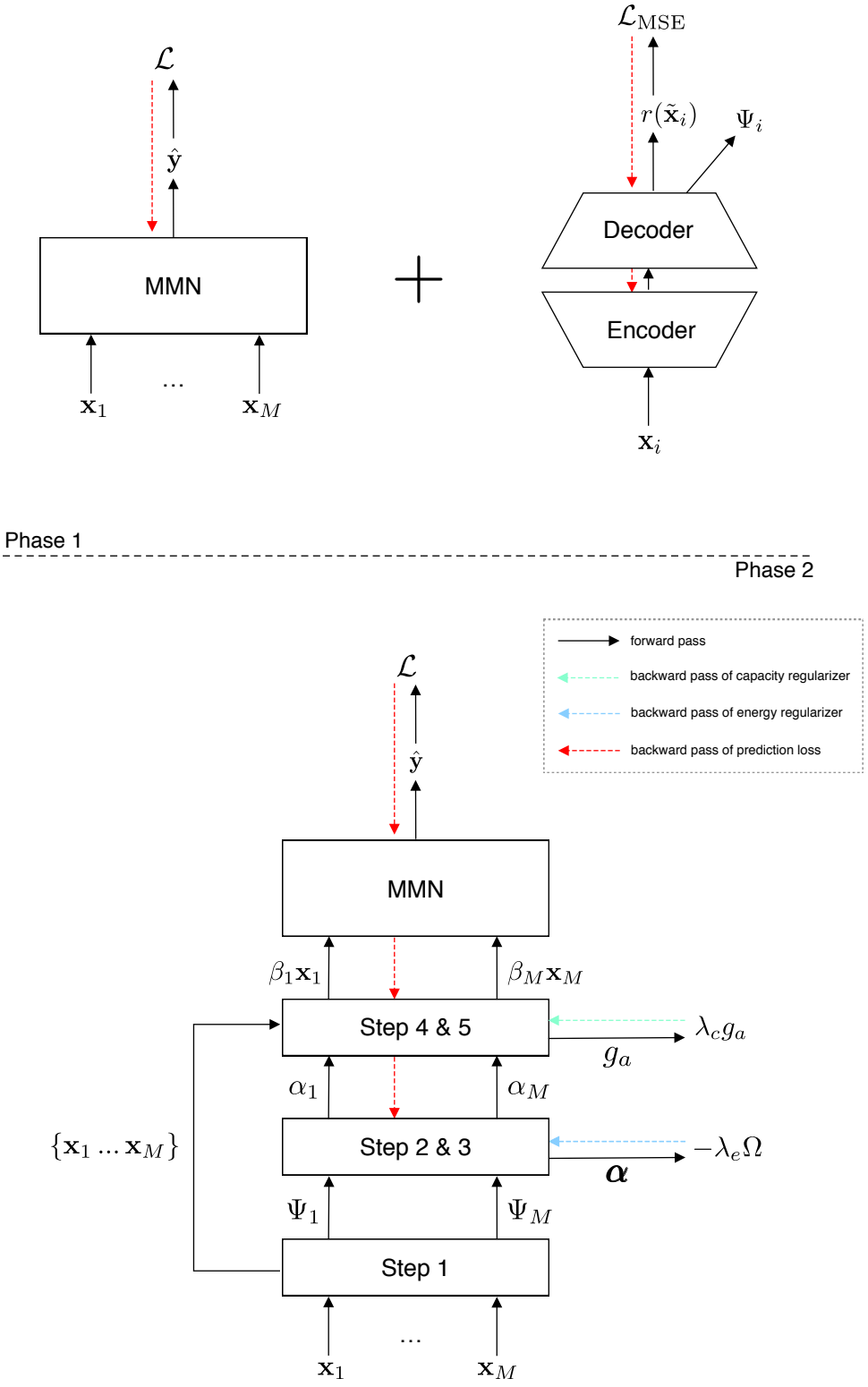


FIGURE 5.6: Summary of end-to-end training

Chapter 6

Experiments & Results

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

6.1 Pulsar detection

The models will be trained to detect pulsar stars. In (very) short, pulsar stars are neutron stars emitting radio waves on a periodic time-frame. A summary of the seminal work of (Lyon, 2016) can be found in Appendix A. The thesis can be accessed here[1] and the dataset here[2].

Short description of two modes (ip and dm) and internal features (mean...). Explain why detection is difficult. Classification signal/background. Skewed dataset, give numbers.

6.2 Corruption

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

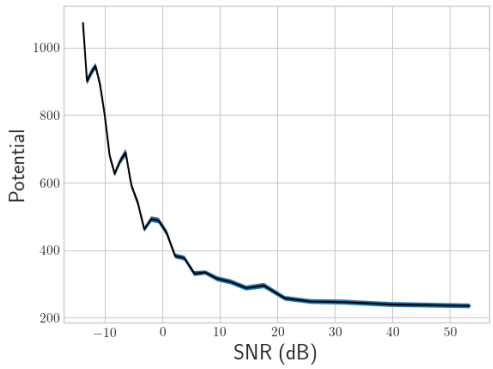
6.3 Experiment II

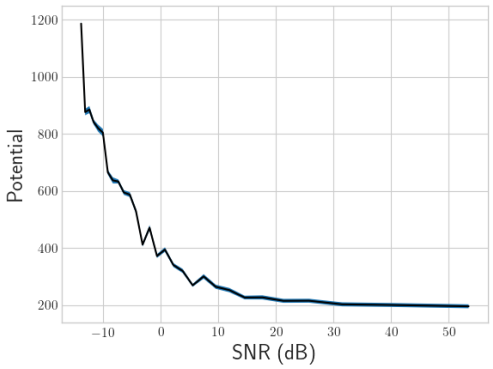
- train-test split
- AE trained on train-set and then test on test-set
- matrix with number of signals, ... (eda)

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

[1] http://www.scienceguyrob.com/wp-content/uploads/2016/12/WhyArePulsarsHardToFind_Lyon_2016.pdf

[2] https://archive.ics.uci.edu/ml/datasets/HTRU2

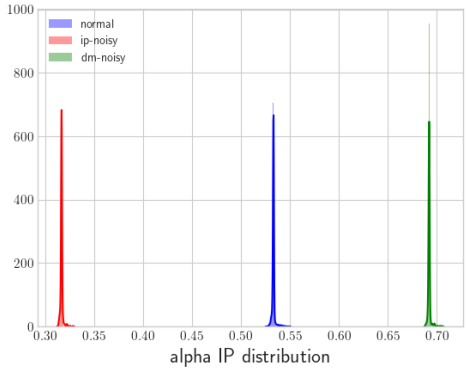


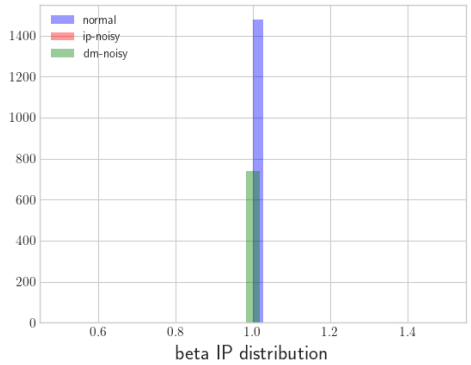


6.4 Experiment III

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

Attention-shift



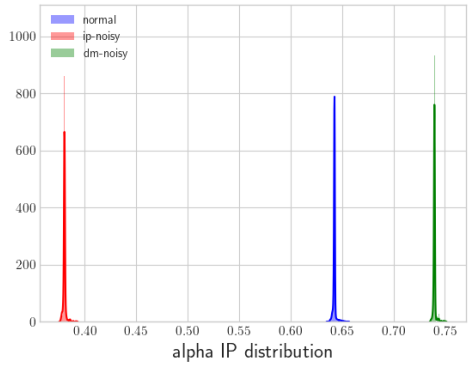


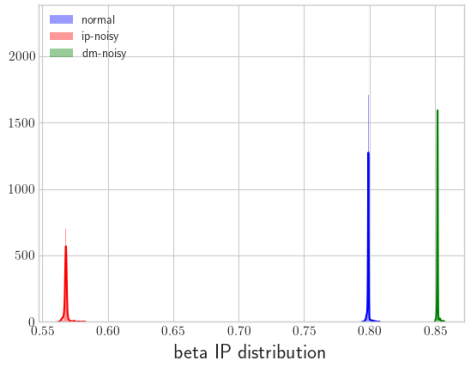
Robustness generalisation

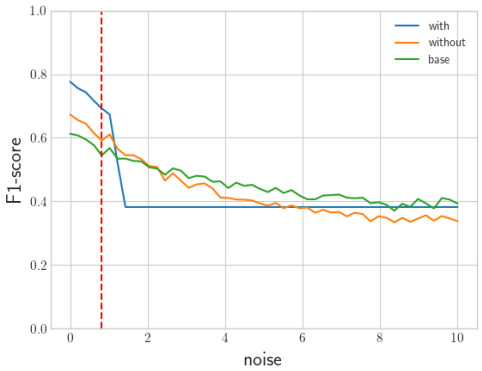
Yerkes-Dodson curve

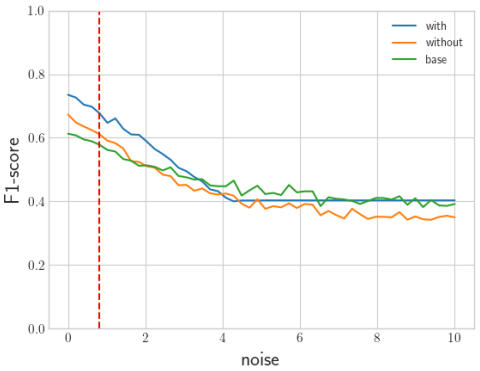
over-under arousal. do on larger range.

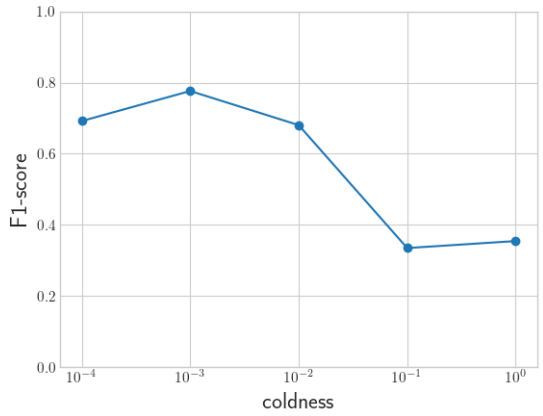
Energy generalisation

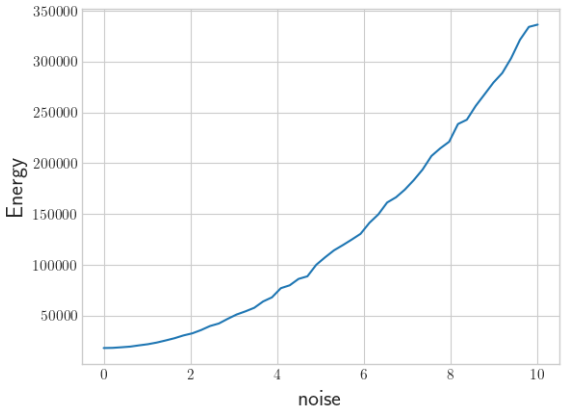

normal
ip-noisy
dm-noisy
alpha IP distribution


normal
ip-noisy
dm-noisy
beta IP distribution


with
without
base
F1-score
noise


with
without
base
F1-score
noise


F1-score
coldness


Energy
noise

Chapter 7

A Unified Model for Multi-Modal Attention

In Section 3.2, three types of multi-modal attention in humans were discussed i.e., endogenous, exogenous and crossmodal attention (Driver and Spence, 1998). As a reminder, attention is endogenous when we *voluntary* choose to attend to something whereas exogenous orients occurs when a person's attention is captured *reflexively* by the sudden onset of an unexpected event (Driver and Spence, 1998). Many exogenous and crossmodal attention mechanisms have been developed in MMDL in recent years (Vaswani et al., 2017; Libovický, Helcl, and Mareček, 2018). By contrast, at the best of our knowledge no previous study has investigated exogenous attention mechanism in deep learning. The EMMA module was created to close this gap. Indeed, EMMA handles unexpected situations by focusing instead on the most relevant modes.

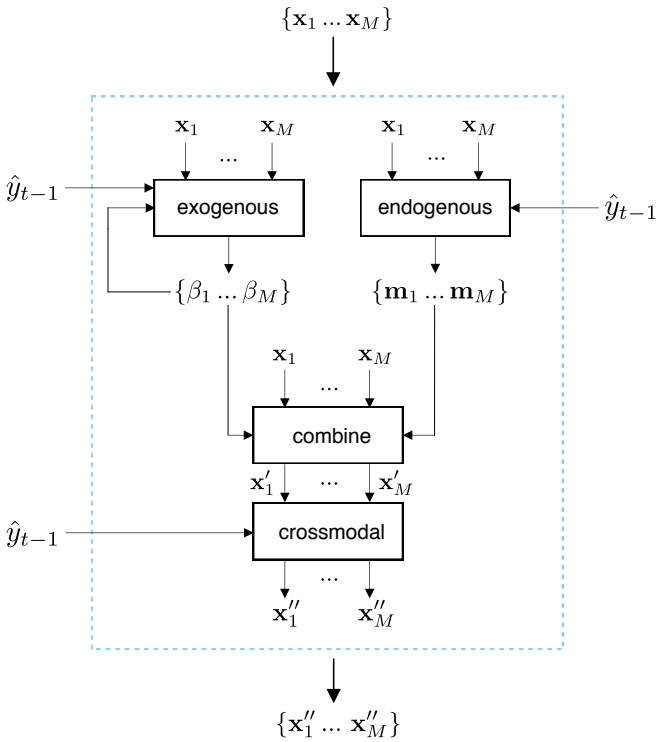


FIGURE 7.1: A general model for Multi-Modal Attention

Three attention mechanisms at our disposal could be used to construct a complete Multi-Modal attention (see Figure 7.1). First, the multi-modal input is forwarded in

parallel through an exo- and endogenous module. The exogenous module outputs attention scores as in Chapter 5. In a similar way, the endogenous module outputs attention masks, where each mask is structured as the corresponding mode and indicates where to focus attention inside the mode. The scores and masks are multiplied and applied to the input. To refine the result is processed by the crossmodal component, leveraging the intrinsic links between the modes to focus on the most important information. In many real-world scenarios, inputs appear in the form of sequences (e.g. time, text, ...). Under those circumstances, the unified model can be further refined by sending the predicting made for the previous input to the attention modules.

In summary, the exogenous module masks perturbations, the endogenous module learns to attend to important information inside each mode, and the crossmodal extracts knowledge by looking at the whole picture.

Chapter 8

Conclusion

Summary of what was seen/done during the master thesis from start to end. Scales up quadratically with number of modes, explain.

8.1 Contributions

Summarize contributions

8.2 Research questions

Each research question + answer

8.3 Future work

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

Appendix A

Dataset

Most of this chapter is directly a summary of the Background chapter of the doctoral thesis (Lyon, 2016). Many details have been voluntary ignored for the sake of simplicity as it is not the focus of this work. Some parts of the text are barely changed from the original.

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 star. Gravitational forces are by consequence directed inwards. During the majority of its life, the start will fuse hydrogen to helium, this generates outwards pressure, balancing the gravity (Ghosh, 2007). When there is no hydrogen to fuse anymore, it starts to use other elements in the surrounding layers 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 undergo a rapid and violent collapse (Ghosh, 2007). The collapse can lead to a number of potential evolutionary outcomes for the leftover core. These depends 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 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 againts pressure. 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.

Pulsars are unique form of neutron stars that retained most of their 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 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 with the velocity c, is known as the light cylinder (see Figure A.2). The

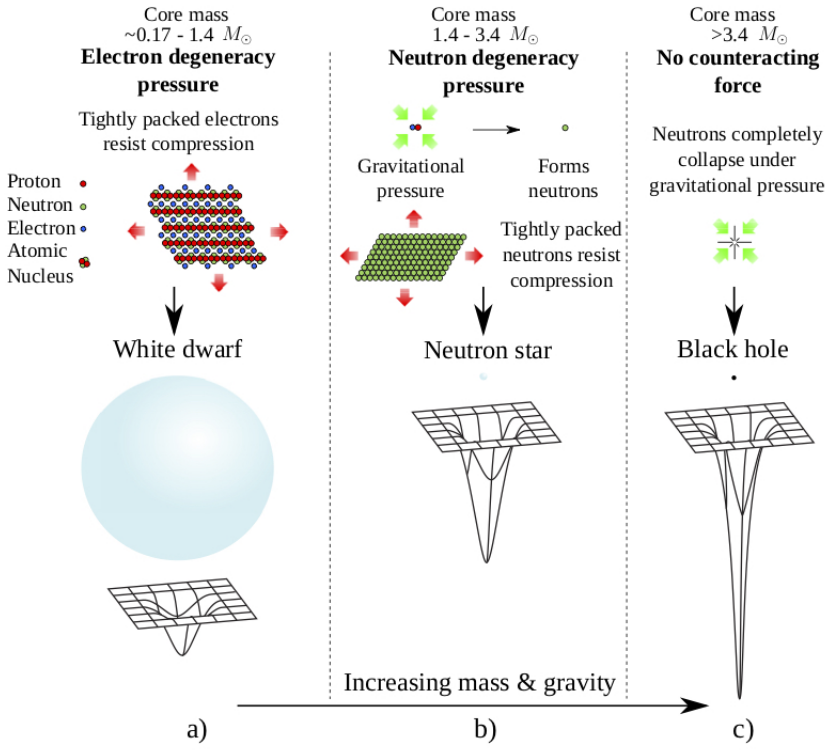


FIGURE A.1: Common evolutionary endpoints for main sequence stars. In a) electron degeneracy pressure prevents gravitational collapse, leading to the formation of a white dwarf star. In b) electron degeneracy pressure is no longer enough to counteract the inward force of gravity, however the gravitational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

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). This is because the beam of radiation is analogous to a lighthouse warning light rotating very quickly.

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.

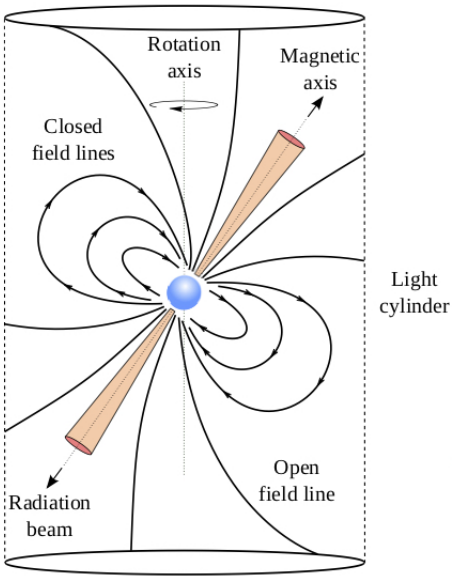


FIGURE A.2: Simplification of the lighthouse model of a radio pulsar, the pulsar is surrounded by a strong magnetic field comprising of open and closed field lines unable to close at the light cylinder. The light cylinder is an imaginary cylinder aligned with the pulsars rotational axis, that synchronously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

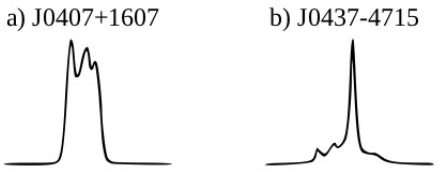


FIGURE A.3: Example pulse profiles of two separate pulsars. These profiles were adapted from those originally presented in (D.R. and M., 2005). *Image and caption copied from* (Lyon, 2016).

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 (row in M') 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.

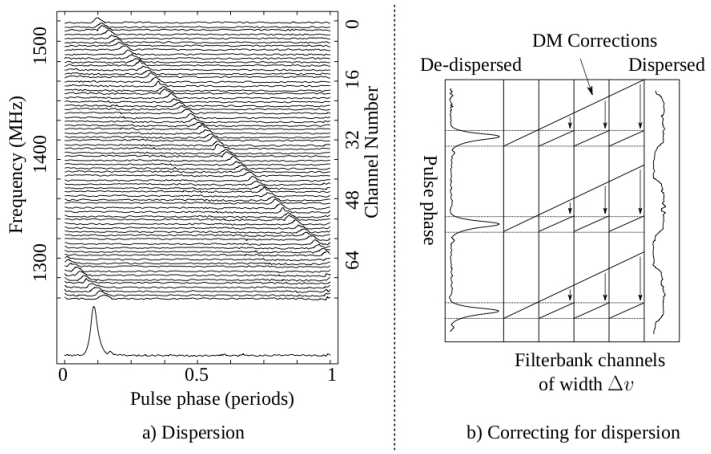


FIGURE A.4: An example of signal dispersion. Based upon diagrams originally presented in (D.R. and M., 2005). Plot a) shows how a signal is dispersed in time. Dispersion hides the true pulse shape and causes a lowering of the detected signal-to-noise. Plot b) shows the application of DM corrections to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

Miscellaneous

B.1 Integrability criterion

B.2 Gradient with respect to gamma

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

$$\frac{\partial \mathcal{L}}{\partial \gamma_{ij}} = \frac{\partial \mathcal{L}}{\partial E_i} \cdot \frac{\partial E_i}{\partial \gamma_{ij}} + \frac{\partial \mathcal{L}}{\partial E_j} \cdot \frac{\partial E_j}{\partial \gamma_{ij}} \tag{B.1}$$

Furthermore,

$$\begin{aligned}
\frac{\partial E_i}{\partial \gamma_{ij}} &= \frac{\partial}{\partial \gamma_{ij}} \sum_{k=1}^{M} E_{ik} \\
&= \frac{\partial E_{ij}}{\partial \gamma_{ij}} \\
&= \frac{\partial}{\partial \gamma_{ij}} (w_{ij} e_i^{\gamma_{ij}} e_j^{1-\gamma_{ij}}) \\
&= w_{ij} e_i^{\gamma_{ij}} \frac{\partial}{\partial \gamma_{ij}} e_j^{1-\gamma_{ij}} + e_j^{1-\gamma_{ij}} \frac{\partial}{\partial \gamma_{ij}} e_i^{\gamma_{ij}} \\
&= w_{ij} e_i^{\gamma_{ij}} e_j^{1-\gamma_{ij}} (\log e_i + \log e_j)
\end{aligned} \tag{B.2}$$

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].

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].

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.

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.

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].

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].

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

Grimm, Christopher et al. (2017). "Modeling Latent Attention Within Neural Networks". In: *arXiv e-prints*, arXiv:1706.00536, arXiv:1706.00536. arXiv: 1706.00536 [cs.AI].

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.

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)*.

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].

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.

Lee, Kuang-Huei et al. (2018). "Stacked Cross Attention for Image-Text Matching". In: *arXiv e-prints*, arXiv:1803.08024, arXiv:1803.08024. arXiv: 1803.08024 [cs.CV].

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

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

Liu, Kuan et al. (2018). "Learn to Combine Modalities in Multimodal Deep Learning". In: *arXiv e-prints*, arXiv:1805.11730, arXiv:1805.11730. arXiv: 1805.11730 [stat.ML].

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.

Osogami, Takayuki (2017). "Boltzmann machines and energy-based models". In: *arXiv e-prints*, arXiv:1708.06008, arXiv:1708.06008. arXiv: 1708.06008 [cs.NE].

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].

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].

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].