Energy-based Multi-Modal Attention

AURELIEN WERENNE

Master Thesis
2018-2019

Energy-based Multi-Modal Attention

Author:
Aurélien Werenne

Supervisor:
Dr. Raphaël Marée

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

Montefiore Institute
Faculty of Applied Sciences
Liège, Belgium

Academic Year 2018 - 2019

"*Artificial Intelligence has the same relation to intelligence as artificial flowers have to flowers.*"

David Parnas

Abstract

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

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

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

Acknowledgements

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

Contents

List of Figures

Notation and Abbreviations

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

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

Chapter 1

Introduction

1.1 Motivation

- AI – DL – lots of applications/advances/succes-stories in recent years perception (not reasoning yet). Cite important papers/reviews and exciting applications (autonomous driving, . . .)
- Cite/explain two multimodal examples: heart disease from multiple sensors and lip/audio speech recognition. This is multimodal (more general than multisensorial, more on this in Section Background)
- Common good practice in engineering is if one sensor/process fails, other will take over (name for this?). Humans cocktail-party effect, cross-modal attention.
- Most AI systems are tested in ideal situations, not realistic. Imagine when one sensor fails, or noisy or unknown (outlying).. AI systems are not explicitly prepared for this.
- I introduce EMMA that solves previous paragraph problem
- What this thesis is about

crossmodal attention

1.2 Proposed solution

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

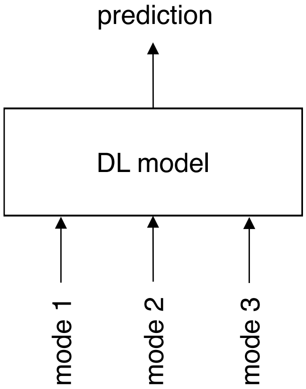


FIGURE 1.1: Main idea

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

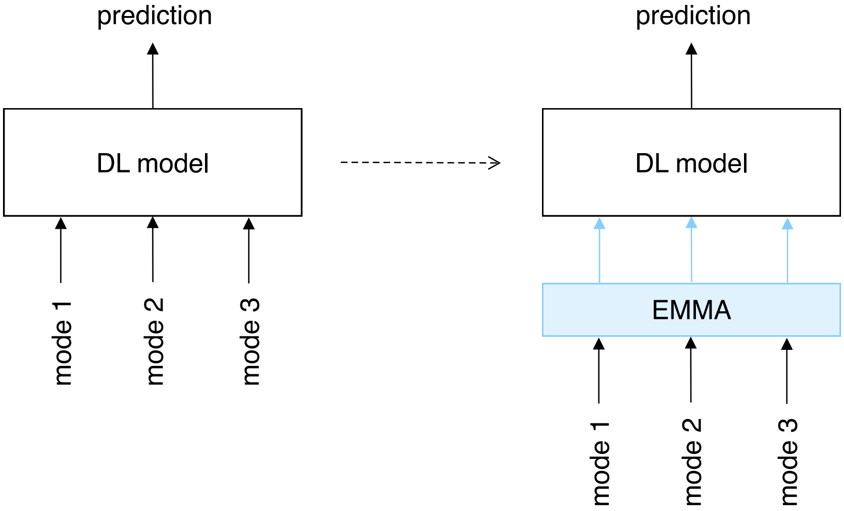


FIGURE 1.2: Main idea

1.3 Thesis Outline

High level overview of this work. Contributions.

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: classification, regression most common tasks T

2.1.1 Deep Learning

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

2.2 Modalities

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

2.3 Energy-models

- Negative log likelihood. Ising model (thermodynamics, stat mechanics). Gibbs distribution. Explain how relate to outlying (high E, low p)
- Boltzmann. MC training. Contrastive divergence. Explain why difficult. Thus we will explore approximators.
- To avoid confusion when we speak about neg likeli, we name it true energy and write it $E^{(T)}$

Boltzmann distrib Energy-based models

2.4 Attention mechanisms

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

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

2.5 Physics

Motivation why, otherwise it will be weird. With a fun example go through model differential equations – condition field – potential - force field - gradient, divergence

Chapter 3

Literature Review

two three pages review current crossmodal attention, multimodal, single mode, weakness, strength, citations. And how they differ from this work. Build up. speak about showing uncertainty in predictions. about attention between modes and its interpretation. interpreta deep lip reading comparisons other cocktail-party solution multi input cross attention transformer example cross attention1 example cross attention2

Chapter 4

Energy Approximators

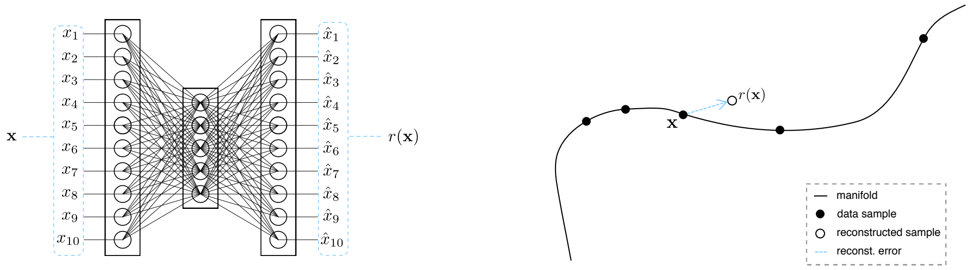
- As we will sse in next chapter, EMMA uses a measure of outlyingness, ideally true energy
- However difficult to train, explore two simple alternative measures: ... those are approximations
- Obtained training a DAE (see section below) explains the model and then next section explain where approximations come from.
- Finally, we conclude with experiment to choose the best

4.1 Autoencoder

Principle AE. unsupervised training. MSE, x, r(x), recon, encoder, decoder. Why useful to do input. Let us distinguish two cases: under/overcomplete

4.1.1 Undercomplete

- undercomplete, hidden representation, forces to keep only useful information, latent space...
- explain diagrams. when we minimize MSE, we minimize norm of vector r-x



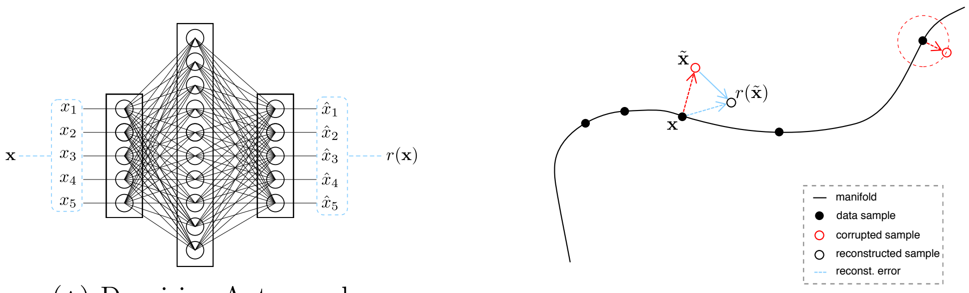
(A) Autoencoder

(B) Data space

FIGURE 4.1: Reconstruction

4.1.2 Overcomplete

overcomplete to avoid copying - corruption becomes denoising. Can be used for denoising of signals. explain diagrams



(A) Denoising Autoencoder

(B) Data space

FIGURE 4.2: Reconstruction with corruption

4.2 Reconstruction norm & Potential

Small introduction. Explain article of Bengio intuitively and how it leads to score is reconstruction.

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

Explain how this relates to vector field. Show circle and vector field.
Then show idea of other paper intuitively. Potential. Conditions. Force.

$$\Psi_i = -\int (r(\tilde{\mathbf{m}}_i) - \mathbf{m}_i) d\mathbf{m}_i$$

Develop not show, just mention it is proofed in paper and not difficult.

$$\Psi_i = -\int f(\mathbf{m}_i) d\mathbf{m}_i - \frac{1}{2}\|\mathbf{m}_i + \mathbf{b}_r\|_2^2 + \text{const}$$

in this work sigmoid is only activation used but other can be easliy used (see paper).

$$\Psi_i = -\sum_k \log(1 + \exp(W_{.k}^T \mathbf{m}_i + b_k^h)) + \frac{1}{2}\|\mathbf{m}_i - \mathbf{b}_r\|_2^2 + \text{const}$$

4.3 Experiment I

We expect potential be better, because better grounded. Describe generation manifolds + parametric functions. Train AE with noise. experimental setup in appendix. Evaluate on grid. Two manifolds formulas. Mention manifold and their particular form were not chosen for a particular reason.

(A) Wave manifold

(B) Circle manifold

FIGURE 4.3: Two manifolds

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

4.4 Limitations

Explain easy but not easily extensible to images/sounds. Mention alternatives. Our goal is not to analyze best energy approximator bt to show that roughly any approx can be used to solve our problem. alternative

Chapter 5

EMMA

Intro to goal of chapter. Make sure they understand it is my contribution. Use active voice.

5.1 Problem Statement

Hypothesis. What and why.

- no sequence
- no image
- MLP dimension D_i
- ...

Explain notation. $\{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ M modes and $N \times D_i$ matrices, $\mathbf{y}$, $\hat{\mathbf{y}}$ loss ... Only makes sense if MI lower or equal MI. Explain with diagram.

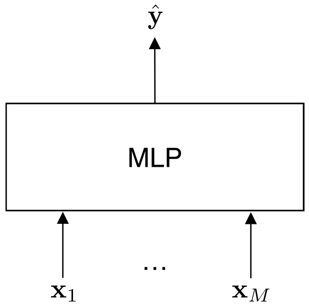


Figure 5.1: blabl

If outlying data (often the case in real world to a certain degree) then MI drops, we cannot do anything to that however in DL it is worse the accuracy drops even more than it loses information (high activations if noise or unknown). Another problem is the fact that DL makes prediction even when big loss of information and does not indicate it is less certain. We can solve this with EMMA it basically determines a β_i for each mode between zero and one. The problem about uncertainty will be explained later. explain with diagram.

How compute/determine β, see next section

5.2 General Framework

What is importance, what do we mean by this. Motivate why determine importance. Three things influence importance:

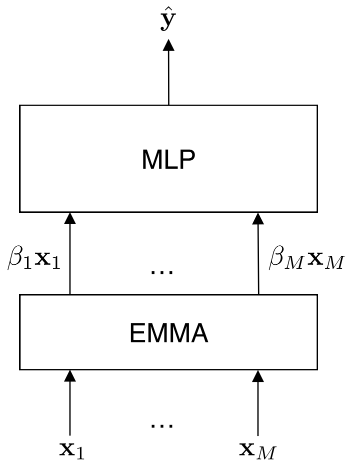


FIGURE 5.2: Summary

- Out-of-distribution: the more unknown, the less accuracy, the less important
- Relevance: at the same, some carry more information for prediction or less
- Independence: how independ with others, are two modes very coupled and carrying the same information and thus redundant (one noisy, other takes over) or completely independent (not easy to completely dismiss info of a noisy mode in that case)

Start with potential satisfies first one (out of distrib). Imagine if we create a new energy taking potential but also relevance and independence. We call this modal energies. And has the following form. Loosely inspired from Boltzmann machines. Explain f is relevance (influence by optim algorithm, minimize loss) and g (how to explain???)

$$f(\Psi) + \sum g(\Psi)$$

Link E_i to importance scores via gibbs distribution. Interpretability α_i. (How to explain??)
Idea of attention is selection but at the same time capactiy. Model learns an optimal capactiy for the task. How much information it needs to led pass by. (How to explain?)

In summary, quick review from start to end. Details explain in each section.

$$\underbrace{\Psi}_{\text{potential energy}} \xrightarrow[\text{Step 1}]{} \underbrace{E}_{\text{modal energy}} \xrightarrow[\text{Step 2}]{} \underbrace{\alpha}_{\text{importance scores}} \xrightarrow[\text{Step 3}]{} \underbrace{\beta}_{\text{attention scores}}$$

5.3 Step 1

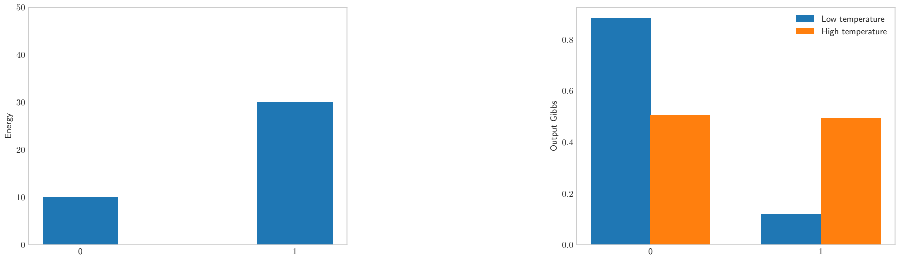
One paragraph per item.

- Mention room from improvement, multiple other ways
- Correction. Minimum potential. Max because log gradient. Ψ'.

- Self-energy e_i. Shape.
- Shared energy $e_{ij} = e_i^{\gamma_{ij}} e_j^{1-\gamma_{ij}}$. Symmetry, coupling.
- Modal energy (Boltzmann) self + shared.
- Notice system energy. Interpretation.

5.4 Step 2

- Gibbs distribution short with $\mathcal{C} = \frac{1}{T}$, high temperature (low coldness) then ... and low temperature (high coldness) ... (explain intuitive influence with diagrams.
- Hyperparameter tuned because can have influence on accuracy. Explain what temperature really is (a way of tuning the strictness, better explain??)
- Mention annealing



(A) Autoencoder

(B) Data space

FIGURE 5.3: Reconstruction

5.5 Step 3

Every item is one subsection. First one without header title.

- Show formula + explain main points design
  - tanh
  - relu
  - $g_a > 0$ and $b_a \in [0, 1]$
  - common (explain later in subsection)
- Energy threshold
- Capacity. What is it intuitively. Math derivation (uninteresting part in appendix, complete version). Small paragraph on intuitive difference with temperature with analogy (fun), notice one is tuned and other is learned, trade-off between the two (crossing).
- Asymmetric effect. most of time attention is $\tanh(\mathbf{Wx}+\mathbf{b})$, in this case gain and bias are not common. We have a problem if this is not common – assymetry.

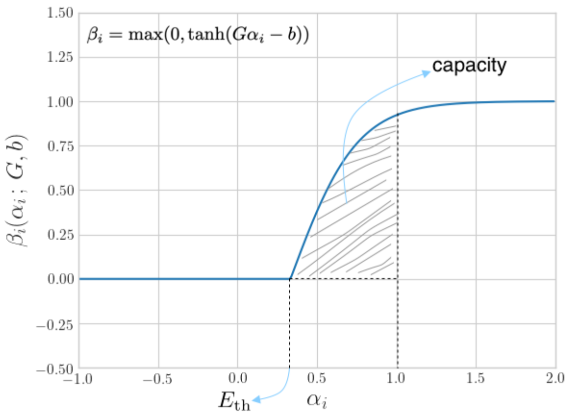


FIGURE 5.4: Summary

5.6 Training & Regularization

- Explain training, 2 phases, train with corrupted/unsee data let us say 50/50 but depend on application.
- The problem is - loose of interpretation because (how to explain??). This is why we add a regularizer. Explain ξ_i (replace $\eta(m_k)$ by ξ_k). Math expression. Explain intuitively and lambda. Quite intuitive, however as we will see some care needs to be taken. Phenomenon in particular case.

Backpropagation of regularizer

- param step update $\theta - \epsilon\lambda\nabla_\theta\Omega$
- develop $\nabla\Omega$, stop before M' introduction
- explain what happens for bad choice of M' and for good choice M'

Contrastive divergence

Explain two cases. No itemize. Illustrations.

5.7 Summary

Describe diagram summarizing end-to-end in two phases with gradients.

5.8 Research questions

- does EMMA improve robustness of model, not only on corrupted data. But on more corrupted data, and unseen data. In other words does it robustness generalize.
- does EMMA help with interpretability (E, alpha, gamma)
- what is the tradeoff between interpretability (lambda) and accuracy
- what is the tradeoff between temperature and capacity? Do we need to minimize capacity to generalize robustness?

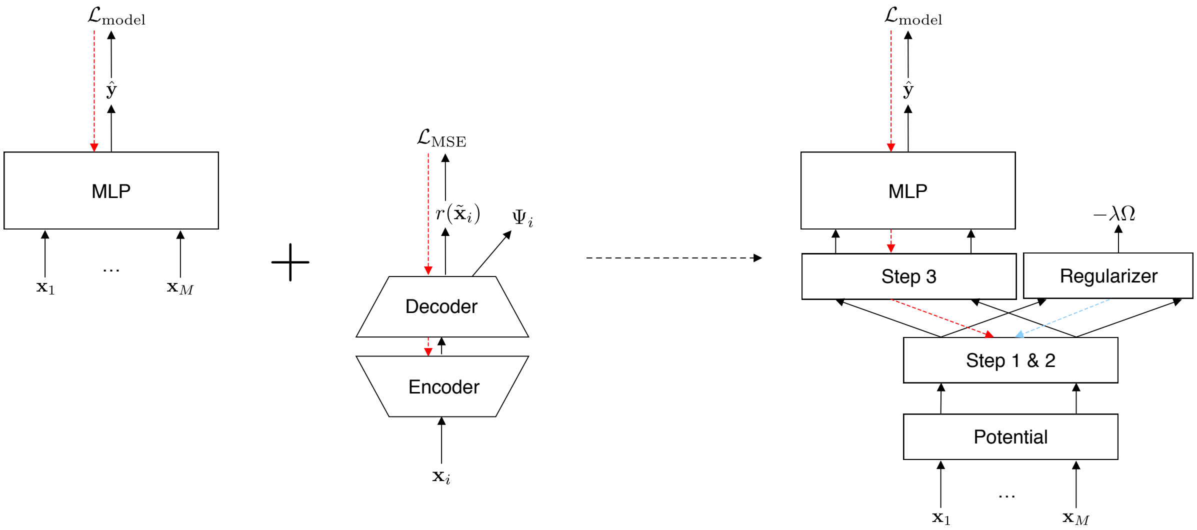


FIGURE 5.5: Summary

Chapter 6

Dataset

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

Chapter 7

Experiments & Results

Explain ideo of chapter, experiment 2 and 3.

7.1 Dataset

Explain pulsar dataset in details (2-3 pages) with cool images from his thesis.

7.2 Evaluation

Standardize: explain important because of noise snr, explain only on train-set and apply same on valid test. Binary crossentropy. Only signal is corrupted. Train-valid-test, valid choose best threshold, temperature and lambda. F1-score (not AUC) explain why, recall. standardize on different, apply from train to test valid. error standardize F1vsAUC1 F1vsAUC2 F1vsAUC3

7.3 Experiment II

Describe experiment + external noise (note the diff with corruption DAE) in dB SNR. Setup in appendix.

trained on signal-train. Eval on signal - val+test, background - val + test. Observations.

(A) IP

(B) DM-SNR

FIGURE 7.1: Potential seen vs unseen

Eval on noisy signal - val+test . Observations.

(A) IP

(B) DM-SNR

FIGURE 7.2: Potential vs noise

7.4 Experiment III

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pellentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

7.4.1 F1-score

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

7.4.2 Interpretation

Morbi rutrum odio eget arcu adipiscing sodales. Aenean et purus a est pulvinar pellentesque. Cras in elit neque, quis varius elit. Phasellus fringilla, nibh eu tempus venenatis, dolor elit posuere quam, quis adipiscing urna leo nec orci. Sed nec nulla auctor odio aliquet consequat. Ut nec nulla in ante ullamcorper aliquam at sed dolor. Phasellus fermentum magna in augue gravida cursus. Cras sed pretium lorem. Pellentesque eget ornare odio. Proin accumsan, massa viverra cursus pharetra, ipsum nisi lobortis velit, a malesuada dolor lorem eu neque.

7.4.3 Regularizer

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

7.4.4 Coldness

Morbi rutrum odio eget arcu adipiscing sodales. Aenean et purus a est pulvinar pellentesque. Cras in elit neque, quis varius elit. Phasellus fringilla, nibh eu tempus venenatis, dolor elit posuere quam, quis adipiscing urna leo nec orci. Sed nec nulla auctor odio aliquet consequat. Ut nec nulla in ante ullamcorper aliquam at sed dolor. Phasellus fermentum magna in augue gravida cursus. Cras sed pretium lorem. Pellentesque eget ornare odio. Proin accumsan, massa viverra cursus pharetra, ipsum nisi lobortis velit, a malesuada dolor lorem eu neque.

Chapter 8

Conclusion

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

8.1 Contributions

Summary of contributions

8.2 Research questions

Each research question + answer

8.3 Future work

- 2-level attention
- Annealing + init, end temperature + explain init of next layer Linke annealing
- different shared energies design
- Images/sound sequences
- multiple modes then blows up. Image 100 modes (altough not realistic in real-world problems) then tend to zero. Solution add a common gain? Analyze influence of multiple modes etc..

Appendix A

Code

A.1 Structure

A.2 Pytorch

A.3 Experiments

Appendix B

Experimental Setup

Appendix C

Miscellaneous