

## TOWARDS SELF-ASSESSMENT IN MACHINE LEARNING MODELS

Jakob Drachmann Havtorn



# UNCERTAINTY AND THE MEDICAL INTERVIEW

# UNCERTAINTY AND THE MEDICAL INTERVIEW

2024-03-05

- Welcome to my PhD defense.
- Thank you to the moderator and the assessment committee for taking part today.
- I will present my work on uncertainty estimation in AI systems for medical domains.
- I will start with an overview of the thesis followed by a brief introduction.
- Then I will present a selection of the research chapters.
- Finally, I will discuss the broader implications of the work.

# OVERVIEW Thesis



- CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND
- 
- CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
- CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION  
USING COMBINED STATISTICAL TESTS
- CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH  
REPRESENTATION LEARNING
- CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH
- CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND  
MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY
- CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-  
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
- 
- CHAPTER 10 DISCUSSION AND CONCLUSION



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

overview

Thesis

introduction, research questions, and background

hierarchical vaes know what they don't know

model-agnostic out-of-distribution detection using combined statistical tests

benchmarking latent variable models for speech representation learning

automated medical coding on mimic-iii and mimic-iv: a critical review and replicability study

a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

discussion and conclusion

1. The thesis is structured into 10 chapters.
2. The first three chapters are introductory.
3. The next six chapters are research chapters.
4. The final chapter is a discussion and conclusion.

# OVERVIEW Thesis

## CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION  
USING COMBINED STATISTICAL TESTS

CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH  
REPRESENTATION LEARNING

CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH

CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND  
MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-  
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

overview  
└ Thesis

1. The thesis is structured into 10 chapters.
2. The first three chapters are introductory.
3. The next six chapters are research chapters.
4. The final chapter is a discussion and conclusion.

CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND	
CHAPTER 4	HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
CHAPTER 5	MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS
CHAPTER 6	A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING
CHAPTER 7	BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH
CHAPTER 8	AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY
CHAPTER 9	A RETROSPECTIVE STUDY ON MACHINE LEARNING- ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
CHAPTER 10	DISCUSSION AND CONCLUSION

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION  
USING COMBINED STATISTICAL TESTS

CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH  
REPRESENTATION LEARNING

CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH

CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND  
MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-  
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION

## PROJECT Background

Industrial PhD project with Corti AI and DTU Compute.

- **2020-2023**
- **Collaboration** between academia and industry partially funded by InnovationFund Denmark.
- **Corti:** Using machine learning to augment communication in the healthcare sector.
- **Project goal:** Pursue research in machine learning at the interface between academic and company interests.



## UNCERTAINTY AND THE MEDICAL INTERVIEW

project

Background

2024-03-05

PROJECT  
Background

Industrial PhD project with Corti AI and DTU Compute.

- **2020-2023**
- **Collaboration** between academia and industry partially funded by InnovationFund Denmark.
- **Corti:** Using machine learning to augment communication in the healthcare sector.
- **Project goal:** Pursue research in machine learning at the interface between academic and company interests.

## INTRODUCTION Medical dialogue

Central to an **interaction** within a healthcare system is the **medical dialogue**:

- General practitioner
- Nurse
- Midwife
- Emergency medical dispatcher
- Paramedic
- Emergency room
- Health insurance



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

## └ introduction

### └ Medical dialogue

1. We focus on medical communication.
2. Involves many different parties.
3. Different contexts and purposes.
4. Busy emergency room, calls to emergency medical dispatchers, visits at general practitioners, etc.

INTRODUCTION  
Medical dialogue

Central to an interaction within a healthcare system is the medical dialogue:

- General practitioner
- Nurse
- Midwife
- Emergency medical dispatcher
- Paramedic
- Emergency room
- Health insurance



## Errors in medical dialogue

- Communication is everywhere in healthcare.
- It is complex, involving multiple participants, different contexts, and different purposes.



# UNCERTAINTY AND THE MEDICAL INTERVIEW

## └ introduction

### └ Errors in medical dialogue

1. So yes communication can be complex and noisy, but what are the consequences?
2. Failure of communication is a leading cause of adverse events.
3. Adverse events are episodes of medical error that result in harm to the patient.
4. Luckily, many of these are preventable, although exact numbers vary.
5. Better communication could help reduce these numbers.

- Communication is everywhere in healthcare.
- It is complex, involving multiple participants, different contexts, and different purposes.



## Errors in medical dialogue

- Communication is everywhere in healthcare.
- It is complex, involving multiple participants, different contexts, and different purposes.
- **Adverse events:** Failure of communication contributes to two out of three adverse events [54].
- **Preventability:** Many adverse outcomes are preventable [10].



└ introduction

└ Errors in medical dialogue

1. So yes communication can be complex and noisy, but what are the consequences?
2. Failure of communication is a leading cause of adverse events.
3. Adverse events are episodes of medical error that result in harm to the patient.
4. Luckily, many of these are preventable, although exact numbers vary.
5. Better communication could help reduce these numbers.

- Communication is everywhere in healthcare.
- It is complex, involving multiple participants, different contexts, and different purposes.
- **Adverse events:** Failure of communication contributes to two out of three adverse events [54].
- **Preventability:** Many adverse outcomes are preventable [10].



## Documenting medical encounters

- Almost every patient interaction has to be documented.
- Patient records, insurance claims, billing, research, training, legal purposes.



# UNCERTAINTY AND THE MEDICAL INTERVIEW

## └ introduction

### └ Documenting medical encounters

1. Another central aspect of medical communication is documentation.
2. Essential for a number of purposes
3. But, it is time-consuming and of varying quality.
4. (Ambulatory ≡ outpatient care, Tertiary ≡ specialized care)

- Almost every patient interaction has to be documented.
- Patient records, insurance claims, billing, research, training, legal purposes.



## Documenting medical encounters

- Almost every patient interaction has to be documented.
- Patient records, insurance claims, billing, research, training, legal purposes.
- Time-consuming: Physicians spend 34-37% of their time on writing documentation [29, 52, 55]<sup>a</sup>.
- Varying quality: Discharge summaries almost never meet *all* timeline, transmission, and content criteria. [24]<sup>b</sup>

<sup>a</sup>Ambulatory care across four specialties in four states and tertiary care at an academic medical center.

<sup>b</sup>Outpatient visits, Yale-New Haven Hospital.



1. Another central aspect of medical communication is documentation.
2. Essential for a number of purposes
3. But, it is time-consuming and of varying quality.
4. (Ambulatory ≡ outpatient care, Tertiary ≡ specialized care)



# How might machine learning help?

- Assist with documentation.
- Augment communication.
- Improve decision-making.



## UNCERTAINTY AND THE MEDICAL INTERVIEW

### └ introduction

#### └ How might machine learning help?

1. Machine learning can help in many ways.
2. The main goal of augmenting communication is:
  - Reduce the impact.
  - Free up time.

- Assist with documentation.
- Augment communication.
- Improve decision-making.



# How might machine learning help?

- **Assist** with documentation.
- **Augment** communication.
- **Improve** decision-making.
- **Reduce** the impact of medical errors and adverse events.
- **Free up** time spent on documentation for patient care.



## └ introduction

### └ How might machine learning help?

1. Machine learning can help in many ways.
2. The main goal of augmenting communication is:
  - Reduce the impact.
  - Free up time.

## Building a decision-support system



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

INTRODUCTION  
Building a decision-support system

└ introduction

└ Building a decision-support system

1. We will take a modular approach to building a decision-support system.
2. First we need source data
3. Then we need to convert it into representations useful for downstream tasks.
4. Then we can perform the downstream tasks.
5. Finally, we need to estimate the reliability of our data, representations, and predictions.

# Building a decision-support system

- **Source data:** Speech or text (potentially images, video, electronic health records, etc.).



└ introduction

└ Building a decision-support system

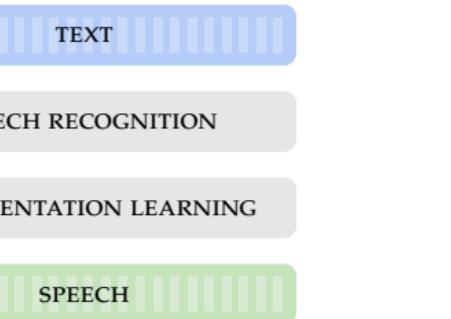
- **Source data:** Speech or text (potentially images, video, electronic health records, etc.).

more

speech

# Building a decision-support system

- **Source data:** Speech or text (potentially images, video, electronic health records, etc.).
- **Foundation modelling:** Converting the input into representations useful for downstream tasks.



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

## └ introduction

### └ Building a decision-support system

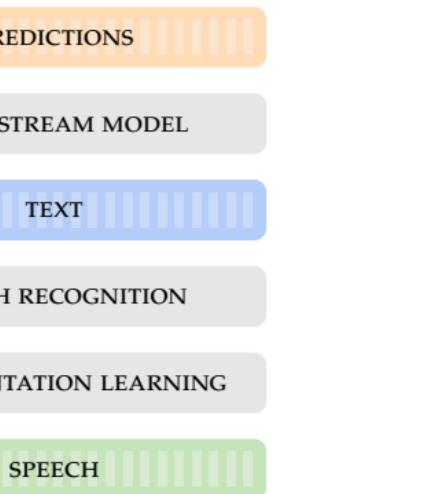
1. We will take a modular approach to building a decision-support system.
2. First we need source data
3. Then we need to convert it into representations useful for downstream tasks.
4. Then we can perform the downstream tasks.
5. Finally, we need to estimate the reliability of our data, representations, and predictions.

- **Source data:** Speech or text (potentially images, video, electronic health records, etc.).
- **Foundation modelling:** Converting the input into representations useful for downstream tasks.



# Building a decision-support system

- **Source data:** Speech or text (potentially images, video, electronic health records, etc.).
- **Foundation modelling:** Converting the input into representations useful for downstream tasks.
- **Downstream tasks:** Suggesting questions, summarizing conversations, classifying illnesses, translating, etc.



└ introduction

└ Building a decision-support system

2024-03-05

- **Source data:** Speech or text (potentially images, video, electronic health records, etc.).

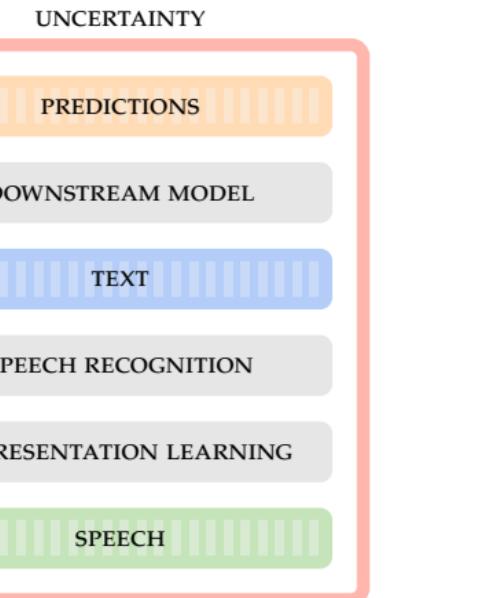
- **Foundation modelling:** Converting the input into representations useful for downstream tasks.

- **Downstream tasks:** Suggesting questions, summarizing conversations, classifying illnesses, translating, etc.



# Building a decision-support system

- **Source data:** Speech or text (potentially images, video, electronic health records, etc.).
- **Foundation modelling:** Converting the input into representations useful for downstream tasks.
- **Downstream tasks:** Suggesting questions, summarizing conversations, classifying illnesses, translating, etc.
- **Uncertainty:** Estimating the reliability of data, representations, predictions.



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

## └ introduction

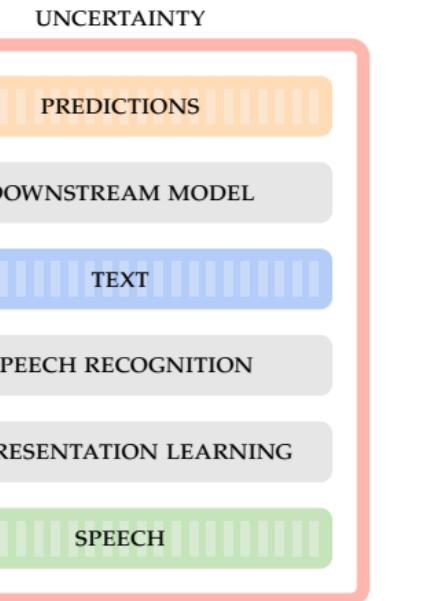
### └ Building a decision-support system

1. We will take a modular approach to building a decision-support system.
2. First we need source data
3. Then we need to convert it into representations useful for downstream tasks.
4. Then we can perform the downstream tasks.
5. Finally, we need to estimate the reliability of our data, representations, and predictions.



# OVERVIEW Thesis

- CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND
- CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
- CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS
- CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING
- CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH
- CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY
- CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
- CHAPTER 10 DISCUSSION AND CONCLUSION



UNCERTAINTY AND THE MEDICAL INTERVIEW

2024-03-05

overview

Thesis

INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH

AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

DISCUSSION AND CONCLUSION

INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH

AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

DISCUSSION AND CONCLUSION

PREDICTIONS

DOWNTSTREAM MODEL

TEXT

SPEECH RECOGNITION

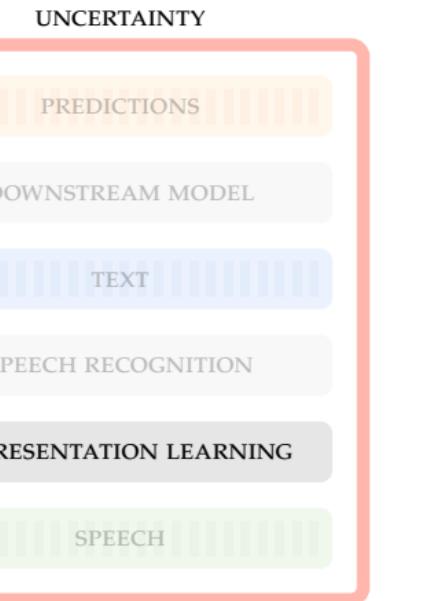
REPRESENTATION LEARNING

SPEECH

1. Examine how a certain class of generative models can be used to detect data that is different from training data.
  2. Generalize this to be model-agnostic and phrased as a statistical test.
  3. Step back and provide an overview of modern methods for unsupervised speech representation learning.
  4. Benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.
  5. Move to a different domain, medical coding, and provide a critical review of recent works.
  6. Provide a retrospective study on a machine learning-assisted stroke recognition system for medical helpline calls.
1. So how does the thesis tackle this problem of building such a system?

# OVERVIEW Thesis

- CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND
- CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
- CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS
- CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING
- CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH
- CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY
- CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
- CHAPTER 10 DISCUSSION AND CONCLUSION



UNCERTAINTY AND THE MEDICAL INTERVIEW

Thesis

2024-03-05

overview

Thesis

UNCERTAINTY

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION

USING COMBINED STATISTICAL TESTS

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH

REPRESENTATION LEARNING

BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH

AUTOMATED MEDICAL CODING ON MIMIC-III AND

MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY

A RETROSPECTIVE STUDY ON MACHINE LEARNING-

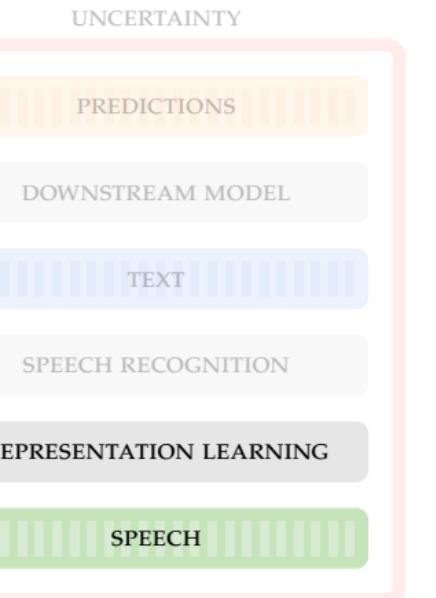
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

DISCUSSION AND CONCLUSION

1. Examine how a certain class of generative models can be used to detect data that is different from training data.
  2. Generalize this to be model-agnostic and phrased as a statistical test.
  3. Step back and provide an overview of modern methods for unsupervised speech representation learning.
  4. Benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.
  5. Move to a different domain, medical coding, and provide a critical review of recent works.
  6. Provide a retrospective study on a machine learning-assisted stroke recognition system for medical helpline calls.
1. So how does the thesis tackle this problem of building such a system?

# OVERVIEW Thesis

- CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND
- CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
- CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS
- CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING**
- CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH
- CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY
- CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
- CHAPTER 10 DISCUSSION AND CONCLUSION



UNIVERSITY Thesis

2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

overview

Thesis

CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS

**CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING**

CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH

CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION

UNIVERSITY Thesis

2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

overview

Thesis

CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS

**CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING**

CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH

CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY

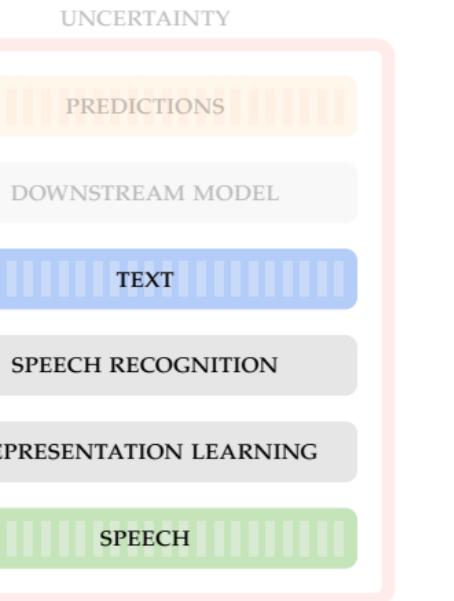
CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION

1. Examine how a certain class of generative models can be used to detect data that is different from training data.
  2. Generalize this to be model-agnostic and phrased as a statistical test.
  3. Step back and provide an overview of modern methods for unsupervised speech representation learning.
  4. Benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.
  5. Move to a different domain, medical coding, and provide a critical review of recent works.
  6. Provide a retrospective study on a machine learning-assisted stroke recognition system for medical helpline calls.
1. So how does the thesis tackle this problem of building such a system?

# OVERVIEW Thesis

- CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND
- CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
- CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS
- CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING
- CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH
- CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY
- CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
- CHAPTER 10 DISCUSSION AND CONCLUSION



UNCERTAINTY AND THE MEDICAL INTERVIEW

overview

Thesis

2024-03-05

UNCERTAINTY

HIERARCHICAL VAE'S KNOW WHAT THEY DON'T KNOW

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS

CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH

CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION

UNCERTAINTY AND THE MEDICAL INTERVIEW

overview

Thesis

2024-03-05

UNCERTAINTY

HIERARCHICAL VAE'S KNOW WHAT THEY DON'T KNOW

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS

CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH

CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY

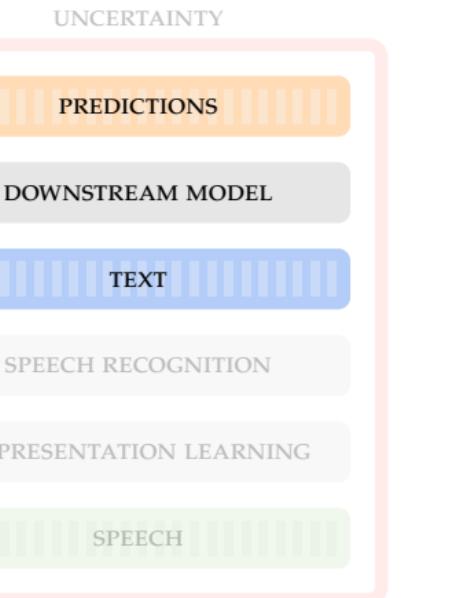
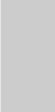
CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION

1. Examine how a certain class of generative models can be used to detect data that is different from training data.
  2. Generalize this to be model-agnostic and phrased as a statistical test.
  3. Step back and provide an overview of modern methods for unsupervised speech representation learning.
  4. Benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.
  5. Move to a different domain, medical coding, and provide a critical review of recent works.
  6. Provide a retrospective study on a machine learning-assisted stroke recognition system for medical helpline calls.
1. So how does the thesis tackle this problem of building such a system?

# OVERVIEW Thesis

- CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND
- CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
- CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS
- CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING
- CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH
- CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY
- CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
- CHAPTER 10 DISCUSSION AND CONCLUSION



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

## overview

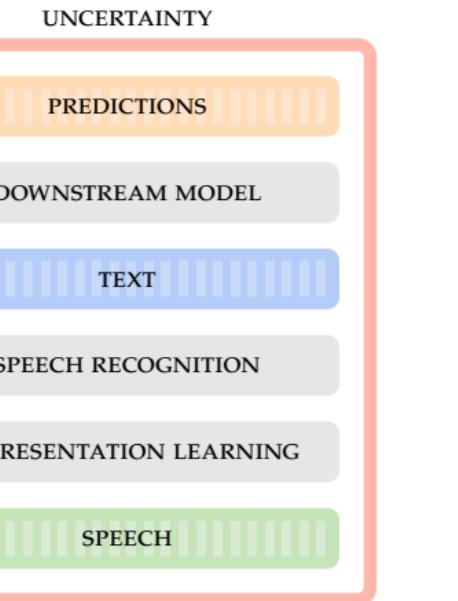
## Thesis

1. Examine how a certain class of generative models can be used to detect data that is different from training data.
  2. Generalize this to be model-agnostic and phrased as a statistical test.
  3. Step back and provide an overview of modern methods for unsupervised speech representation learning.
  4. Benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.
  5. Move to a different domain, medical coding, and provide a critical review of recent works.
  6. Provide a retrospective study on a machine learning-assisted stroke recognition system for medical helpline calls.
- 
1. So how does the thesis tackle this problem of building such a system?



# OVERVIEW Thesis

- CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND
- CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
- CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS
- CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING
- CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH
- CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY
- CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
- CHAPTER 10 DISCUSSION AND CONCLUSION



UNCERTAINTY AND THE MEDICAL INTERVIEW

2024-03-05

overview

Thesis

UNCERTAINTY

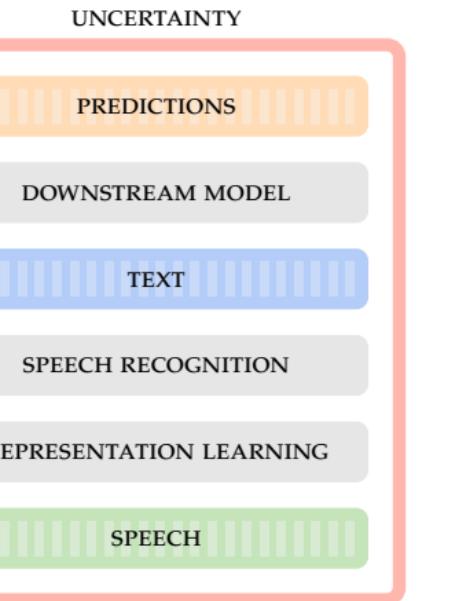
The table of contents lists the following chapters:

- HIERARCHICAL VAE'S KNOW WHAT THEY DON'T KNOW
- MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS
- A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING
- BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH
- AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY
- A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
- DISCUSSION AND CONCLUSION

1. Examine how a certain class of generative models can be used to detect data that is different from training data.
  2. Generalize this to be model-agnostic and phrased as a statistical test.
  3. Step back and provide an overview of modern methods for unsupervised speech representation learning.
  4. Benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.
  5. Move to a different domain, medical coding, and provide a critical review of recent works.
  6. Provide a retrospective study on a machine learning-assisted stroke recognition system for medical helpline calls.
1. So how does the thesis tackle this problem of building such a system?

# OVERVIEW Presentation

- CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND**
- CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW**
- CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS**
- CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING**
- CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH**
- CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY**
- CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS**
- CHAPTER 10 DISCUSSION AND CONCLUSION**



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

overview

↳ Presentation

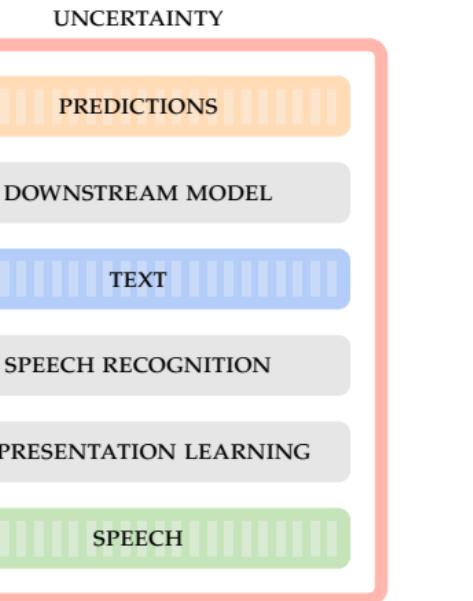
1. We want to keep papers that represent each of the three main themes of the thesis: (Uncertainty estimation, representation learning, and clinical applications)

UNCERTAINTY  
Presentation



# OVERVIEW Presentation

- CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND
- CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
- CHAPTER 5 MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS
- CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING
- CHAPTER 7 BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH
- CHAPTER 8 AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY
- CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
- CHAPTER 10 DISCUSSION AND CONCLUSION



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

## overview

### Presentation

UNCERTAINTY

CHAPTER 1-3	INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND
CHAPTER 4	HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
CHAPTER 5	MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION USING COMBINED STATISTICAL TESTS
CHAPTER 6	A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING
CHAPTER 7	BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH
CHAPTER 8	AUTOMATED MEDICAL CODING ON MIMIC-III AND MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY
CHAPTER 9	A RETROSPECTIVE STUDY ON MACHINE LEARNING- ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
CHAPTER 10	DISCUSSION AND CONCLUSION

PRESENTATION

INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION  
USING COMBINED STATISTICAL TESTS

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH  
REPRESENTATION LEARNING

BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH

AUTOMATED MEDICAL CODING ON MIMIC-III AND  
MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY

A RETROSPECTIVE STUDY ON MACHINE LEARNING-  
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

DISCUSSION AND CONCLUSION

# OVERVIEW

# Presentation

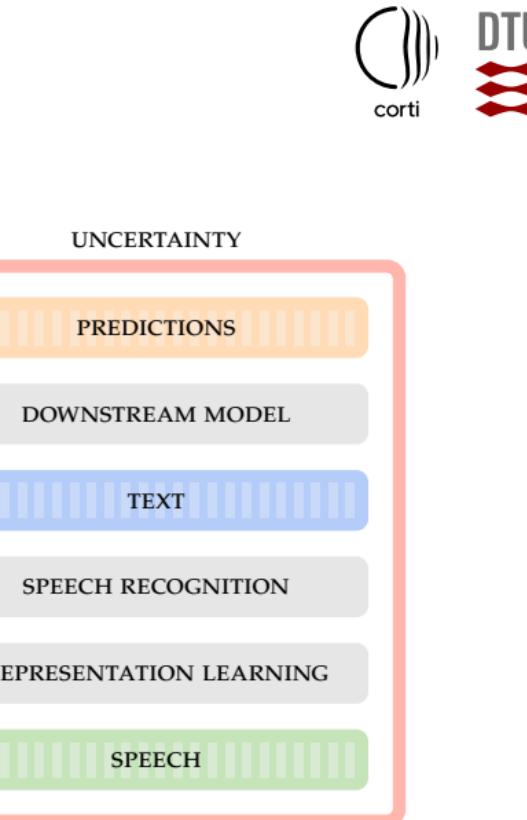
## **CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND**

## CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

## CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

# CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

## CHAPTER 10 DISCUSSION AND CONCLUSION



# UNCERTAINTY AND THE MEDICAL INTERVIEW

## └ overview

└ Presentation

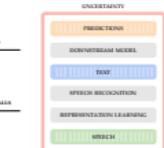
```
graph TD; A[UNCERTAINTY AND THE MEDICAL INTERVIEW] --> B[overview]; A --> C[Presentation]; C --> D[interview 1.2]; C --> E[interview 1.3];
```

UNCERTAINTY AND THE MEDICAL INTERVIEW

↳ overview

↳ Presentation

- ↳ interview 1.2
- ↳ interview 1.3



# OVERVIEW Presentation

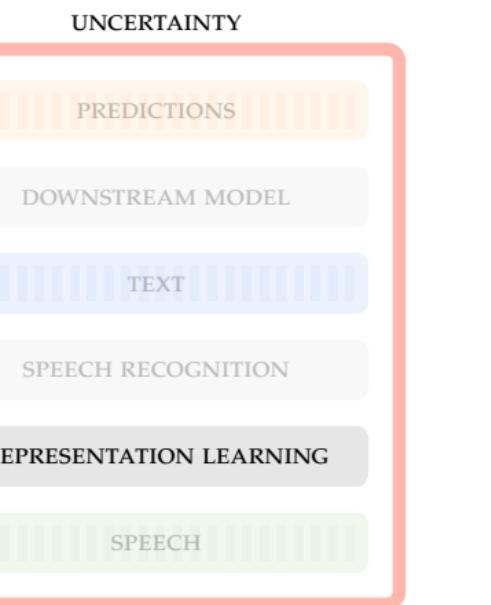
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH  
REPRESENTATION LEARNING

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-  
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION



UNCERTAINTY AND THE MEDICAL INTERVIEW

2024-03-05

overview

Presentation

DTU Presentation

UNCERTAINTY

PREDICTIONS

HIERARCHICAL VAE'S KNOW WHAT THEY DON'T KNOW

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

DISCUSSION AND CONCLUSION

UNCERTAINTY

PREDICTIONS

HIERARCHICAL MODELS

TEXT

SPEECHrecognition

REPRESENTATION LEARNING

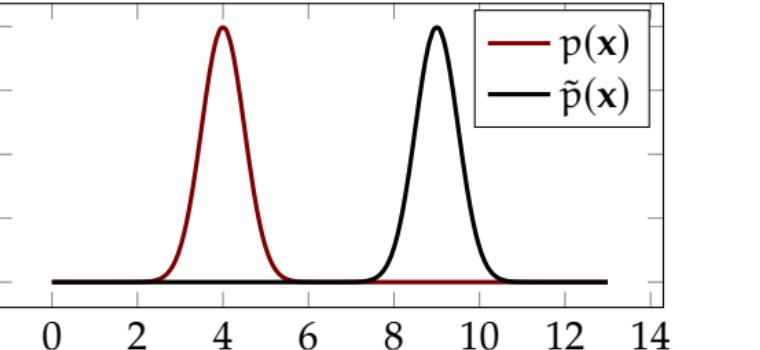
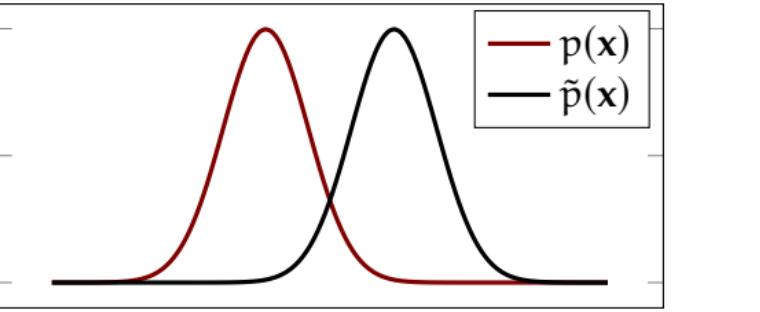
SPEECH

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW  
Defining OOD detection

Enable models to distinguish the training data distribution  $p(x)$  from any other distribution  $\tilde{p}(x)$ .

Do this for any given single observation, i.e. answer the question:

"Was  $x$  sampled from  $p(x)$  or not?"



UNCERTAINTY AND THE MEDICAL INTERVIEW

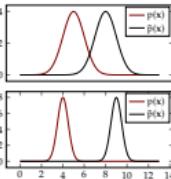
hierarchical vae's know what they don't know

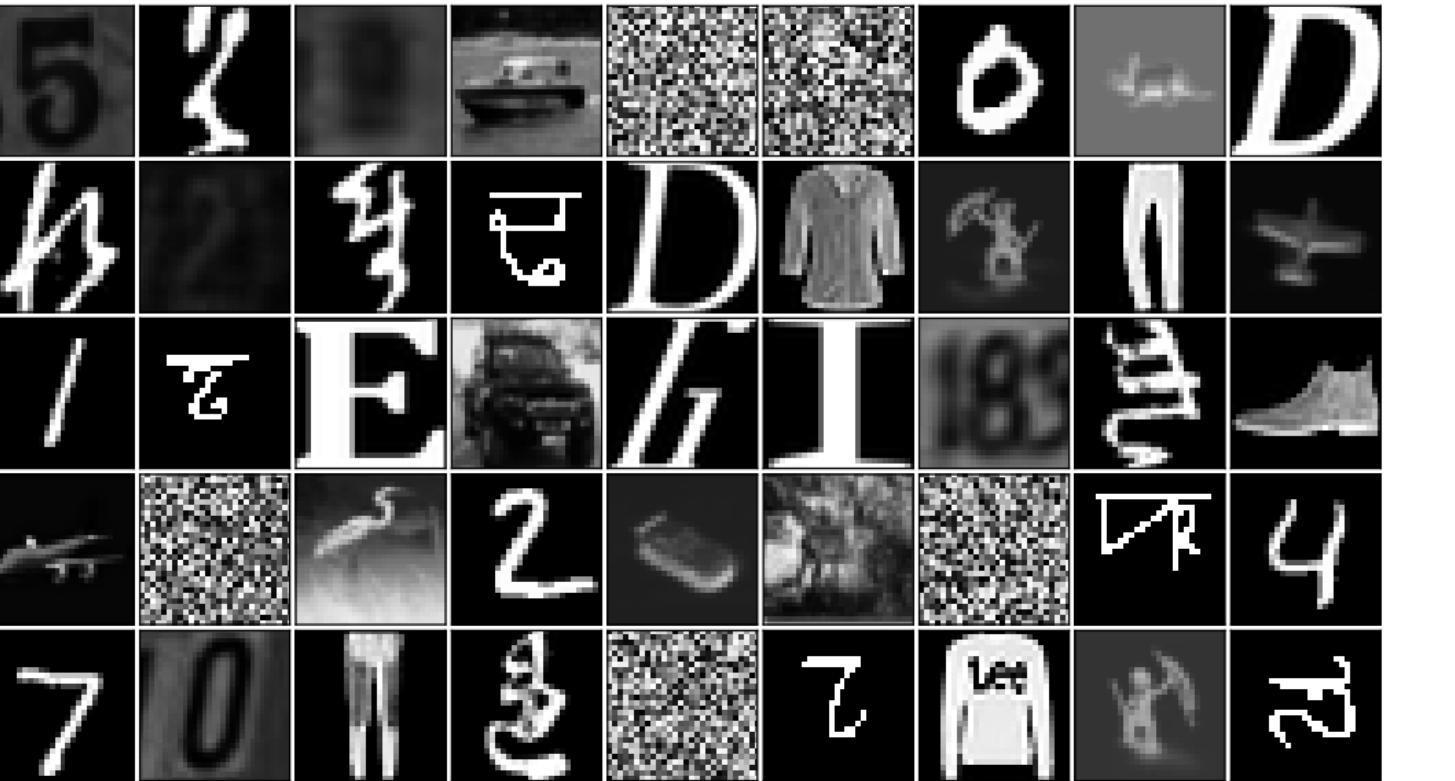
Defining OOD detection

2024-03-05

Enable models to distinguish the training data distribution  $p(x)$  from any other distribution  $\tilde{p}(x)$ .  
Do this for any given single observation, i.e. answer the question:

"Was  $x$  sampled from  $p(x)$  or not?"



**In distribution?****UNCERTAINTY AND THE MEDICAL INTERVIEW**

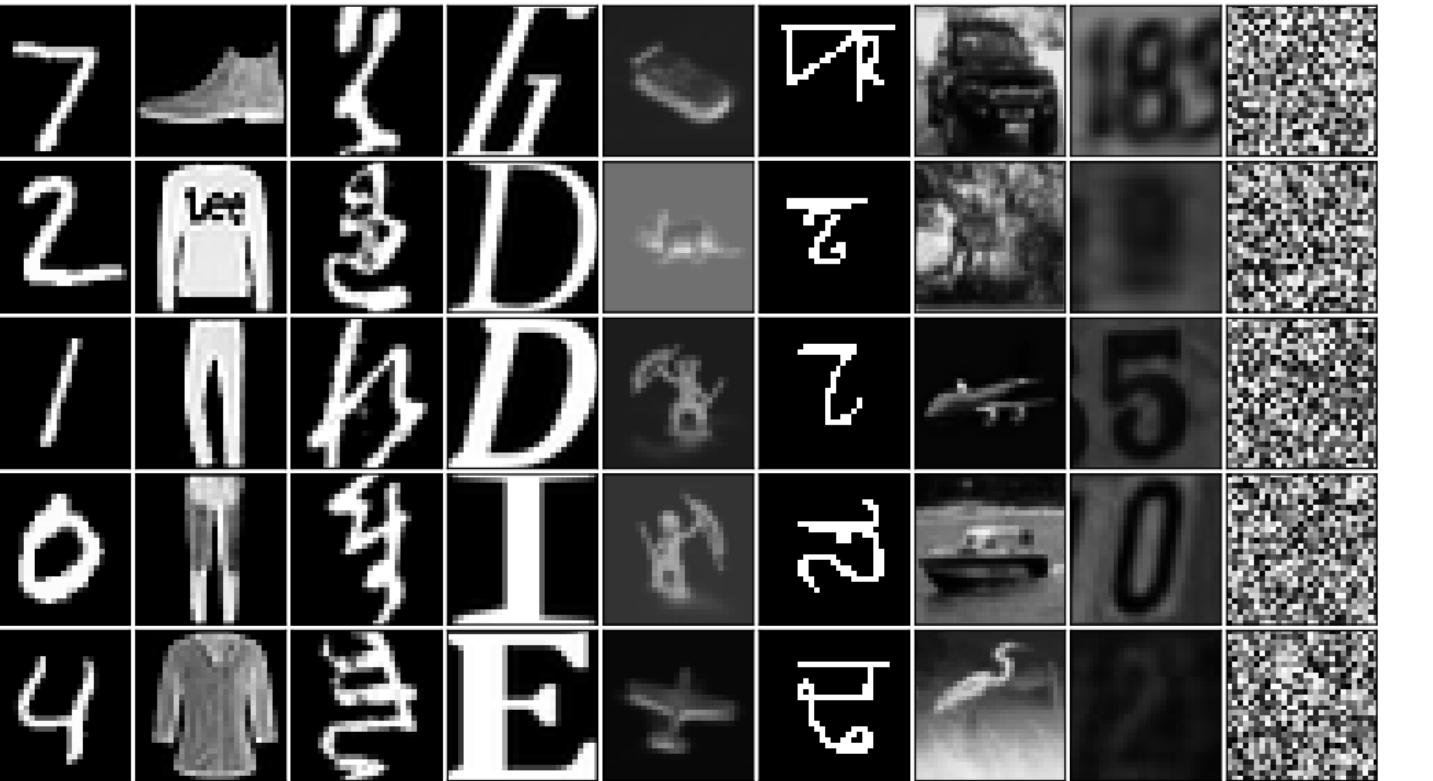
└ hierarchical vaes know what they don't know

└ In distribution?



1. Datasets can overlap quite a bit in their raw data space.
2. What we usually care about is a more semantic notion of similarity.

# Out of distribution?



## UNCERTAINTY AND THE MEDICAL INTERVIEW

-hierarchical vaes know what they don't know

## └ Out of distribution?

- Datasets can overlap quite a bit in their raw data space.  
What we usually care about is a more semantic notion of similarity.

## Out-of-distribution detection with generative models

- Generative models learn to approximate the **data distribution**  $p(x)$ .
- The likelihood of the model given a sample  $x$  is a measure of how well the model **explains the data**.
- **Model likelihood** has long been thought of as useful for OOD detection [5].



## UNCERTAINTY AND THE MEDICAL INTERVIEW

↳ hierarchical vae's know what they don't know

↳ Out-of-distribution detection with generative models

2024-03-05

- Generative models learn to approximate the **data distribution**  $p(x)$ .

- The likelihood of the model given a sample  $x$  is a measure of how well the model **explains the data**.

- Model likelihood has long been thought of as useful for OOD detection [5].

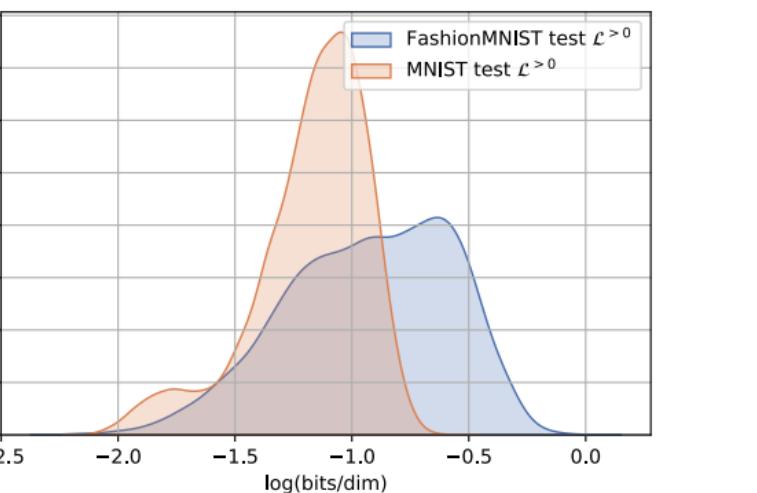
## Out-of-distribution detection with generative models

- Generative models learn to approximate the **data distribution**  $p(x)$ .
- The likelihood of the model given a sample  $x$  is a measure of how well the model **explains the data**.
- **Model likelihood** has long been thought of as useful for OOD detection [5].

## UNCERTAINTY AND THE MEDICAL INTERVIEW

↳ hierarchical vae's know what they don't know

↳ Out-of-distribution detection with generative models

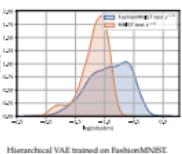


Hierarchical VAE trained on FashionMNIST.

- Generative models learn to approximate the data distribution  $p(x)$ .

- The likelihood of the model given a sample  $x$  is a measure of how well the model explains the data.

- Model likelihood has long been thought of as useful for OOD detection [5].



HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW  
**Hierarchical VAE**

We choose the hierarchical VAE as our model [33, 48].

$$p_{\theta}(x) = \int p_{\theta}(x, z) dz = \int p_{\theta}(x|z)p_{\theta}(z) dz$$

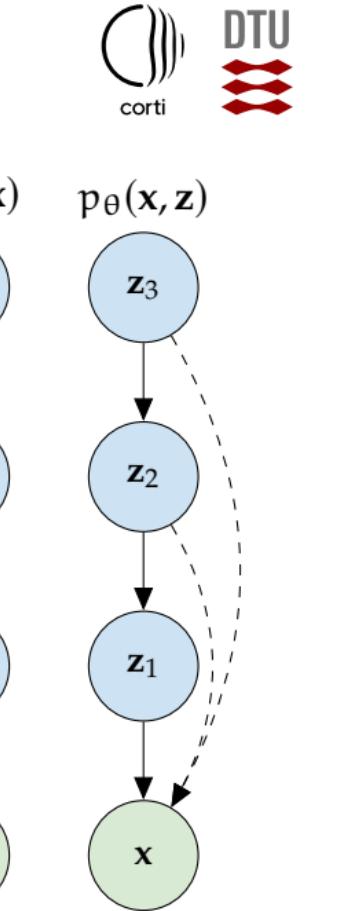
Specifically we use

- ① a three-layered hierarchical VAE with bottom-up inference and deterministic skip-connections for both inference and generation.

Generative model:  $p_{\theta}(x|z) = p_{\theta}(x|z_1)p_{\theta}(z_1|z_2)p(z_2)$ ,

Inference model:  $q_{\phi}(z|x) = q_{\phi}(z_1|x)q_{\phi}(z_2|z_1)q_{\phi}(z_3|z_2)$ .

- ② a ten-layered layered Bidirectional-Inference Variational Autoencoder (BIVA) [40].



2024-03-05

**UNCERTAINTY AND THE MEDICAL INTERVIEW**

hierarchical vae know what they don't know

Hierarchical VAE

We choose the hierarchical VAE as our model [33, 48].

$p_{\theta}(x) = \int p_{\theta}(x, z) dz = \int p_{\theta}(x|z)p_{\theta}(z) dz$

Specifically we use

- ➊ a three-layered hierarchical VAE with bottom-up inference and deterministic skip-connections for both inference and generation.
- ➋ Generative model:  $p_{\theta}(x|z) = p_{\theta}(x|z_1)p_{\theta}(z_1|z_2)p(z_2)$ ,
- ⌾ Inference model:  $q_{\phi}(z|x) = q_{\phi}(z_1|x)q_{\phi}(z_2|z_1)q_{\phi}(z_3|z_2)$ .
- ➌ a ten-layered layered Bidirectional-Inference Variational Autoencoder (BIVA) [40].

q<sub>φ</sub>(z|x)      p<sub>θ</sub>(x, z)

z<sub>3</sub>            z<sub>3</sub>  
z<sub>2</sub>            z<sub>2</sub>  
z<sub>1</sub>            z<sub>1</sub>  
x                x

corti      DTU

## What is wrong with the ELBO for OOD detection?

We can split the ELBO into two terms

$$\mathcal{L}(\mathbf{x}; \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction likelihood}} - \underbrace{D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))}_{\text{regularization penalty}}. \quad (1)$$

The first term is high if the data is well-explained by  $\mathbf{z}$ . The second term we can rewrite as,

$$D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \sum_{i=1}^{L-1} \log \frac{p_\theta(\mathbf{z}_i|\mathbf{z}_{i+1})}{q_\phi(\mathbf{z}_i|\mathbf{z}_{i-1})} + \log \frac{p_\theta(\mathbf{z}_L)}{q_\phi(\mathbf{z}_L|\mathbf{z}_{L-1})} \right]. \quad (2)$$

Since the individual terms are computed by summing over the dimensionality of  $\mathbf{z}_i$ , the absolute log-ratios grow with  $\text{dim}(\mathbf{z}_i)$ .

Since the lower-most latent variables are usually higher dimensional than top ones, these are weighted higher in the ELBO.



2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

↳ hierarchical vae's know what they don't know

↳ What is wrong with the ELBO for OOD detection?

We can split the ELBO into two terms

$$\mathcal{L}(\mathbf{x}; \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})). \quad (1)$$

The first term is high if the data is well-explained by  $\mathbf{z}$ . The second term we can rewrite as,

$$D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \sum_{i=1}^{L-1} \log \frac{p_\theta(\mathbf{z}_i|\mathbf{z}_{i+1})}{q_\phi(\mathbf{z}_i|\mathbf{z}_{i-1})} + \log \frac{p_\theta(\mathbf{z}_L)}{q_\phi(\mathbf{z}_L|\mathbf{z}_{L-1})} \right]. \quad (2)$$

Since the individual terms are computed by summing over the dimensionality of  $\mathbf{z}_i$ , the absolute log-ratios grow with  $\text{dim}(\mathbf{z}_i)$ .

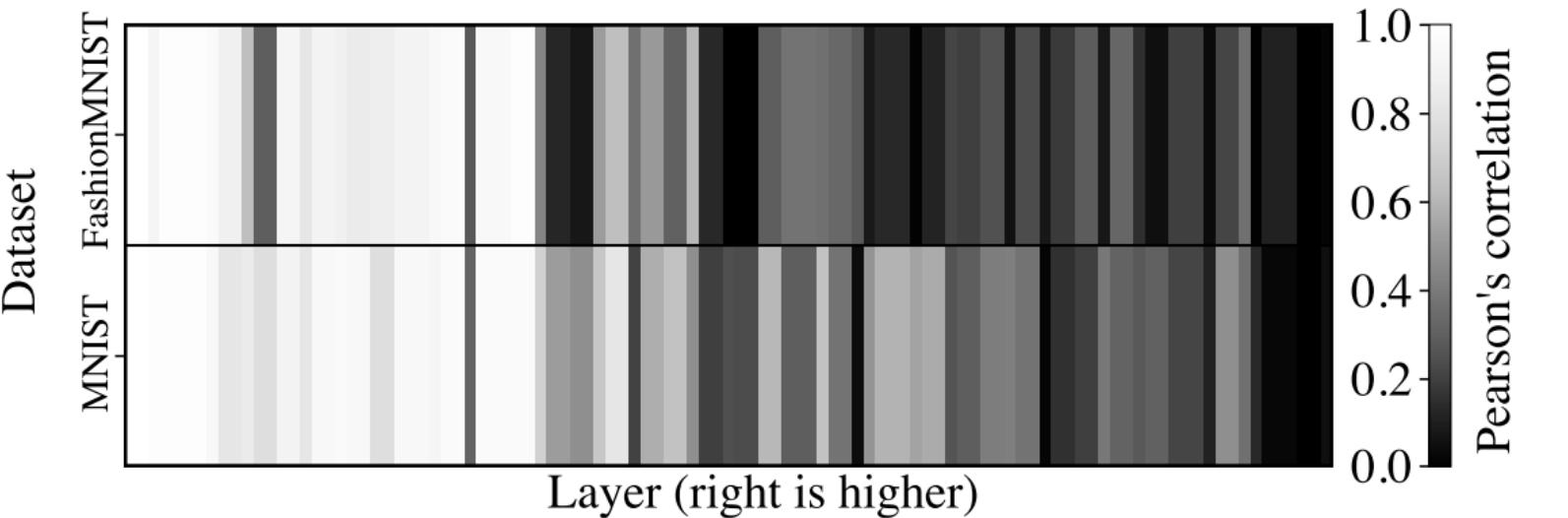
Since the lower-most latent variables are usually higher dimensional than top ones, these are weighted higher in the ELBO.

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

## What do the lowest latent variables represent?

Absolute Pearson correlations between data representations in all layers of the inference network of a hierarchical VAE trained on FashionMNIST and of another trained on MNIST.

Correlation computed between the representations of the two different models given the same data, FashionMNIST (top) and MNIST (bottom).

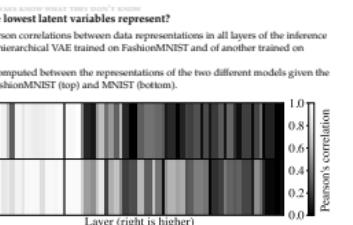


2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

↳ hierarchical vae's know what they don't know

↳ What do the lowest latent variables represent?



1. Strong evidence that the lowest latent variables are generalizing across datasets.

## Likelihood ratios

We suggest to define a likelihood ratio score [9] using the ELBO  $\mathcal{L}(x)$  and a relaxed bound  $\mathcal{L}^{>k}(x)$ .

$$\text{LLR}^{>k}(x) \equiv \mathcal{L}(x) - \mathcal{L}^{>k}(x), \quad (3)$$

where the exact form of the bounds is,

$$\mathcal{L} = \log p_\theta(x) - D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x)), \quad (4)$$

$$\mathcal{L}^{>k} = \log p_\theta(x) - D_{\text{KL}}(p_\theta(z_{<k}|z_{>k})q_\phi(z_{>k}|x) \| p_\theta(z|x)).$$

In the likelihood ratio  $\log p_\theta(x)$  cancels out and only the KL-divergences from the approximate to the true posterior remain.

$$\begin{aligned} \text{LLR}^{>k}(x) &= -D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x)) \\ &\quad + D_{\text{KL}}(p_\theta(z_{<k}|z_{>k})q_\phi(z_{>k}|x) \| p_\theta(z|x)). \end{aligned} \quad (5)$$



# UNCERTAINTY AND THE MEDICAL INTERVIEW

hierarchical vaes know what they don't know

Likelihood ratios

2024-03-05

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW  
Likelihood ratios  
We suggest to define a likelihood ratio score [9] using the ELBO  $\mathcal{L}(x)$  and a relaxed bound  $\mathcal{L}^{>k}(x)$ .  

$$\text{LLR}^{>k}(x) \equiv \mathcal{L}(x) - \mathcal{L}^{>k}(x), \quad (3)$$
  
where the exact form of the bounds is,  

$$\begin{aligned} \mathcal{L} &= \log p_\theta(x) - D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x)), \\ \mathcal{L}^{>k} &= \log p_\theta(x) - D_{\text{KL}}(p_\theta(z_{<k}|z_{>k})q_\phi(z_{>k}|x) \| p_\theta(z|x)). \end{aligned} \quad (4)$$
  
In the likelihood ratio  $\log p_\theta(x)$  cancels out and only the KL-divergences from the approximate to the true posterior remain.  

$$\begin{aligned} \text{LLR}^{>k}(x) &= -D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x)) \\ &\quad + D_{\text{KL}}(p_\theta(z_{<k}|z_{>k})q_\phi(z_{>k}|x) \| p_\theta(z|x)). \end{aligned} \quad (5)$$

The importance weighted autoencoder (IWAE) bound is tight with the true likelihood in the limit of infinite samples,  $S \rightarrow \infty$  [8],

$$\mathcal{L}_S = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[ \log \frac{1}{N} \sum_{s=1}^S \frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})} \right] \leq \log p_\theta(\mathbf{x}), \quad (6)$$

Consequently, by importance sampling the ELBO,  $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})) \rightarrow 0$  and our likelihood ratio reduces to the KL-divergence of  $\mathcal{L}^{>k}$ .

$$LLR_S^{>k}(\mathbf{x}) \rightarrow D_{KL}(p(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})q(\mathbf{z}_{>k}|\mathbf{x}) \parallel p(\mathbf{z}|\mathbf{x})). \quad (7)$$

$LLR_S^{>k}(\mathbf{x})$  performs KL-divergence-based OOD detection using top-most latent variables.

## UNCERTAINTY AND THE MEDICAL INTERVIEW

- └ hierarchical vae know what they don't know
- └ Importance sampling the ELBO

2024-03-05

HIERARCHICAL VAEs KNOW WHAT THEY DON'T KNOW  
**Importance sampling the ELBO**

The importance weighted autoencoder (IWAE) bound is tight with the true likelihood in the limit of infinite samples,  $S \rightarrow \infty$  [8].

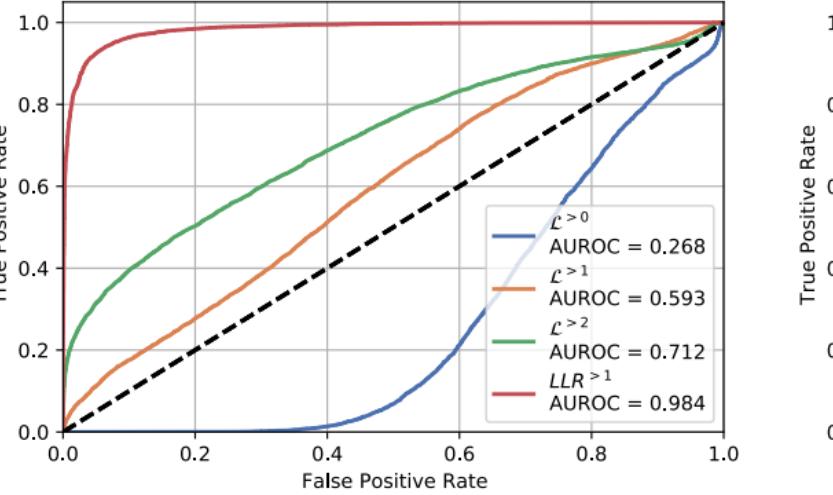
$$\mathcal{L}_S = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[ \log \frac{1}{N} \sum_{s=1}^S \frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})} \right] \leq \log p_\theta(\mathbf{x}), \quad (6)$$

Consequently, by importance sampling the ELBO,  $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})) \rightarrow 0$  and our likelihood ratio reduces to the KL-divergence of  $\mathcal{L}^{>k}$ .

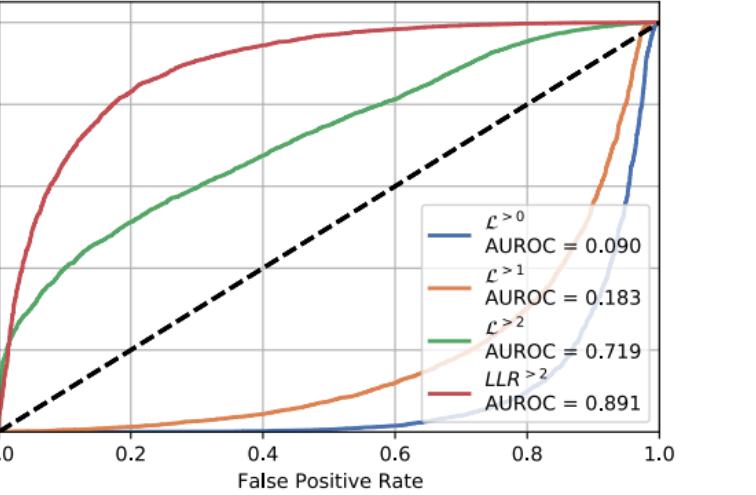
$$LLR_S^{>k}(\mathbf{x}) \rightarrow D_{KL}(p(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})q(\mathbf{z}_{>k}|\mathbf{x}) \parallel p(\mathbf{z}|\mathbf{x})). \quad (7)$$

$LLR_S^{>k}(\mathbf{x})$  performs KL-divergence-based OOD detection using top-most latent variables.

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW  
ROC curves with  $\mathcal{L}^{>k}$  and  $LLR^{>k}$



(a) FashionMNIST HVAE evaluated on MNIST



(b) CIFAR10 BIVA evaluated on SVHN

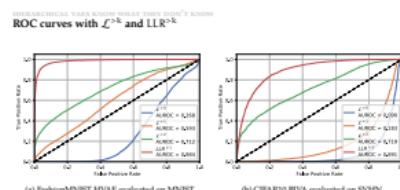


2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

hierarchical vae's know what they don't know

ROC curves with  $\mathcal{L}^{>k}$  and  $LLR^{>k}$



## Results on CIFAR10/SVHN

Method	AUROC↑	AUPRC↑	FPR80↓
<b>CIFAR10 (in) / SVHN (out)</b>			
<b>Use prior knowledge of OOD</b>			
Backgr. contrast. LR (PixelCNN) [47]	0.930	0.881	0.066
Backgr. contrast. LR (VAE) [59]	0.265	-	-
Outlier exposure [23]	0.984	-	-
Input complexity (S, Glow) [51]	0.950	-	-
Input complexity (S, PixelCNN++) [51]	0.929	-	-
Input complexity (S, HVAE) (Ours) [51]	0.833	0.855	0.344
<b>Use in-distribution data labels <math>y</math></b>			
Mahalanobis distance [36]	0.991	-	-
<b>No OOD-specific assumptions</b>			
- <i>Ensembles</i>			
WAIC, 5 models, Glow [12]	1.000	-	-
WAIC, 5 models, PixelCNN [47]	0.628	0.616	0.657
- <i>Not ensembles</i>			
Likelihood regret [59]	0.875	-	-
LLR $>^2$ + HVAE (ours)	0.811	0.837	0.394
LLR $>^2$ + BIVA (ours)	<b>0.891</b>	<b>0.875</b>	<b>0.172</b>



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

↳ hierarchical vaes know what they don't know

↳ Results on CIFAR10/SVHN

Method	AUROC <sub>SVHN</sub>	AUPRC <sub>SVHN</sub>	FPR80 <sub>SVHN</sub>
<b>CIFAR10 (in) / SVHN (out)</b>			
Use prior knowledge of OOD			
Backgr. contrast. LR (PixelCNN) [47]	0.930	0.881	0.066
Backgr. contrast. LR (VAE) [59]	0.265	-	-
Outlier exposure [23]	0.984	-	-
Input complexity (S, Glow) [51]	0.950	-	-
Input complexity (S, PixelCNN++) [51]	0.929	-	-
Input complexity (S, HVAE) (Ours) [51]	0.833	0.855	0.344
Use in-distribution data labels $y$			
Mahalanobis distance [36]	0.991	-	-
<b>No OOD-specific assumptions</b>			
- <i>Ensembles</i>			
WAIC, 5 models, Glow [12]	1.000	-	-
WAIC, 5 models, PixelCNN [47]	0.628	0.616	0.657
- <i>Not ensembles</i>			
Likelihood regret [59]	0.875	-	-
LLR $>^2$ + HVAE (ours)	0.811	0.837	0.394
LLR $>^2$ + BIVA (ours)	<b>0.891</b>	<b>0.875</b>	<b>0.172</b>

- Key observations:
  - The likelihood of a generative model is not a good score for OOD detection [41].
  - Strong correlations between some latent variables for different datasets.
- Provide explanation for why the likelihood fails for OODD for HVAEs.
- Proposed a likelihood-ratio score,  $LLR^{>k}$ , that uses the conditional prior for the bottom-most latent variables in the hierarchy and showed its effectiveness.

2024-03-05

UNCERTAINTY AND THE MEDICAL INTERVIEW

hierarchical vae's know what they don't know

Conclusions

• Key observations:

- The likelihood of a generative model is not a good score for OOD detection [41].
- Strong correlations between some latent variables for different datasets.

- Provide explanation for why the likelihood fails for OODD for HVAEs.
- Proposed a likelihood-ratio score,  $LLR^{>k}$ , that uses the conditional prior for the bottom-most latent variables in the hierarchy and showed its effectiveness.

1. Clock says **29 minutes left** give or take

# OVERVIEW Presentation

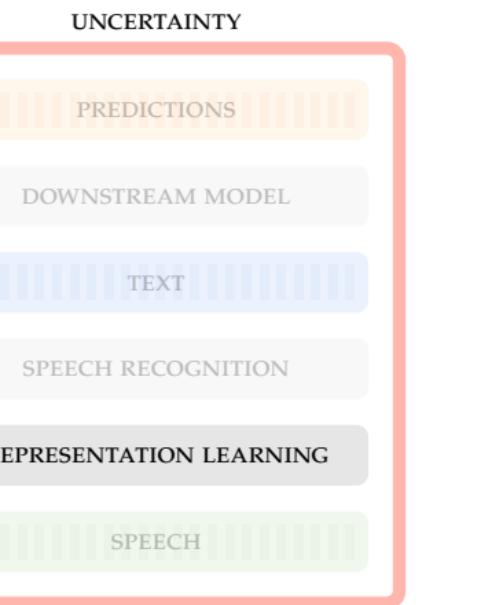
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH  
REPRESENTATION LEARNING

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-  
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

overview

Presentation

# OVERVIEW Presentation

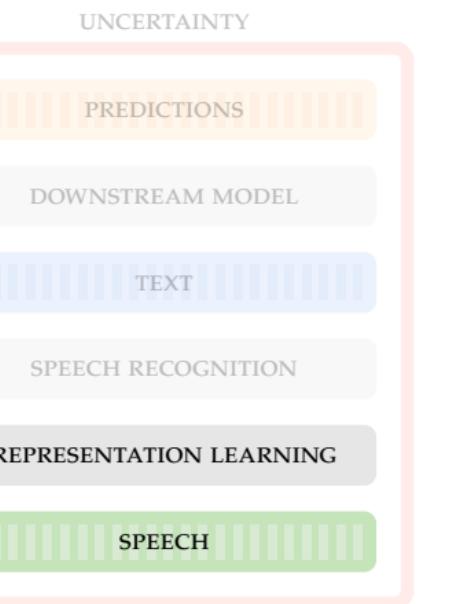
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH  
REPRESENTATION LEARNING

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-  
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION



UNCERTAINTY AND THE MEDICAL INTERVIEW

2024-03-05

overview

Presentation

UNCERTAINTY

PREDICTIONS

DOWNTSTREAM MODEL

TEXT

SPEECH RECOGNITION

REPRESENTATION LEARNING

SPEECH

INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

DISCUSSION AND CONCLUSION

PRESENTATION

UNCERTAINTY

PREDICTIONS

DOWNTSTREAM MODEL

TEXT

SPEECH RECOGNITION

REPRESENTATION LEARNING

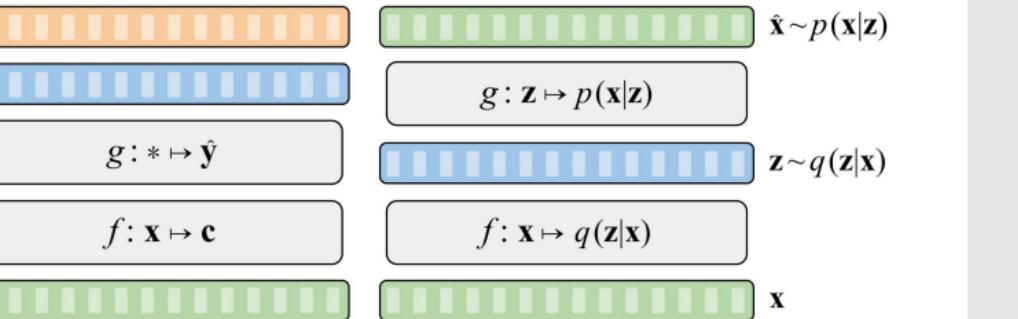
SPEECH

## A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

### Overview: Representation Learning for Speech



- Reviews two learning paradigms:
  - Self-supervised learning (SSL)
  - Probabilistic latent variable models (LVMs)
- Recent developments have been driven by **self-supervised learning**.
- A model-by-model overview for selected self-supervised models.

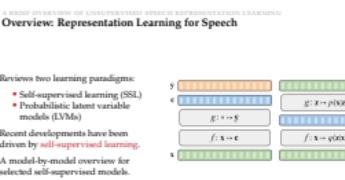


2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

### a brief overview of unsupervised speech representation learning

#### Overview: Representation Learning for Speech



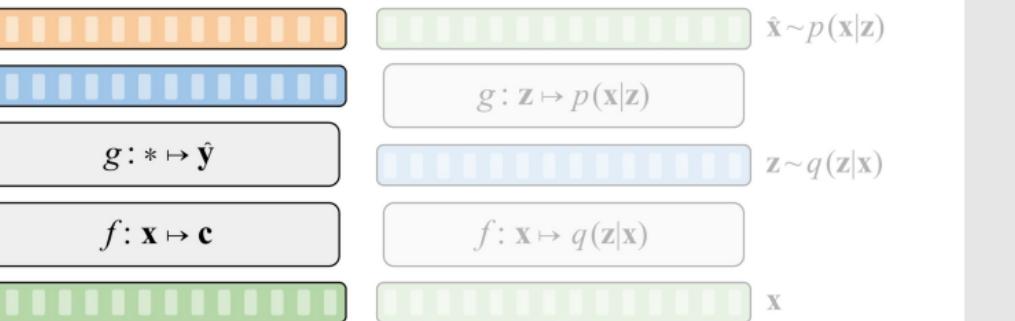
### 1. Focus on the self-supervised learning models.

## A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

### Overview: Representation Learning for Speech



- Reviews two learning paradigms:
  - Self-supervised learning (SSL)
  - Probabilistic latent variable models (LVMs)
- Recent developments have been driven by **self-supervised learning**.
- A model-by-model overview for selected self-supervised models.



## UNCERTAINTY AND THE MEDICAL INTERVIEW

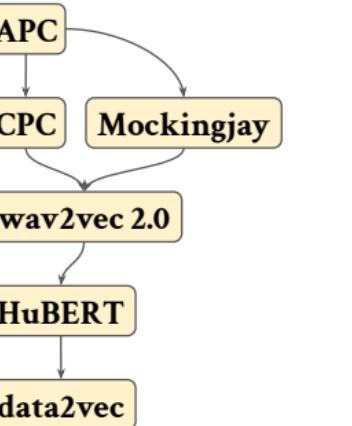
### a brief overview of unsupervised speech representation learning

#### Overview: Representation Learning for Speech

2024-03-05

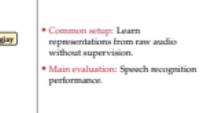
- Reviews two learning paradigms:
  - Self-supervised learning (SSL)
  - Probabilistic latent variable models (LVMs)
- Recent developments have been driven by **self-supervised learning**.
- A model-by-model overview for selected self-supervised models.





- **Common setup:** Learn representations from raw audio without supervision.
- **Main evaluation:** Speech recognition performance.

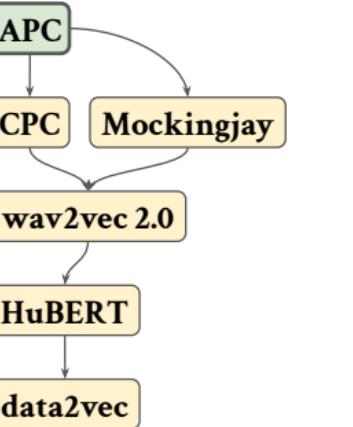
2024-03-05



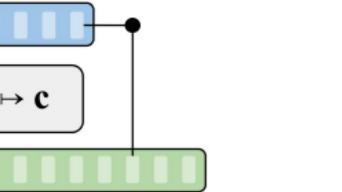
- **Common setup:** Learn representations from raw audio without supervision.
- **Main evaluation:** Speech recognition performance.

1. Clock says 29 minutes left give or take

## A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING Autoregressive Predictive Coding (APC)



- **Task:** Predict future inputs.
- **Input/target:** Log-mel spectrogram.
- **Architecture:** RNN/Transformer decoder.
- **Slow features:** Predict k steps ahead.



2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

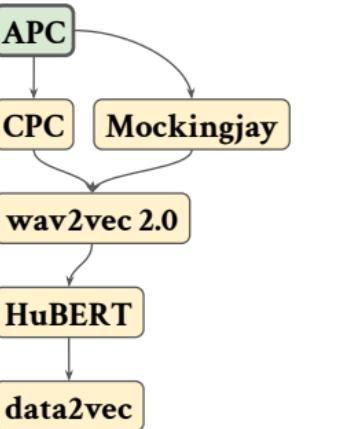
### a brief overview of unsupervised speech representation learning

#### Autoregressive Predictive Coding (APC)

The slide title is "UNCERTAINTY AND THE MEDICAL INTERVIEW". Below it is the subtitle "a brief overview of unsupervised speech representation learning". The main heading is "Autoregressive Predictive Coding (APC)".  
On the right side, there is a legend:

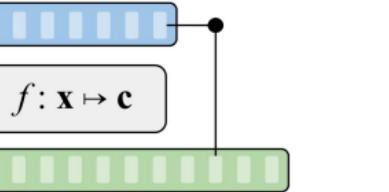
- Task: Predict future inputs.
- Input/target: Log-mel spectrogram.
- Architecture: RNN/Transformer decoder.
- Slow features: Predict k steps ahead.

Below the legend is a small diagram showing a sequence of frames  $x_1, x_2, \dots, x_n$  and a prediction frame  $c_{n+1}$  with the label  $f: x \rightarrow c$ .



- Challenges:

- Encodes only past inputs ✗
- Uses the input as target ✗



2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

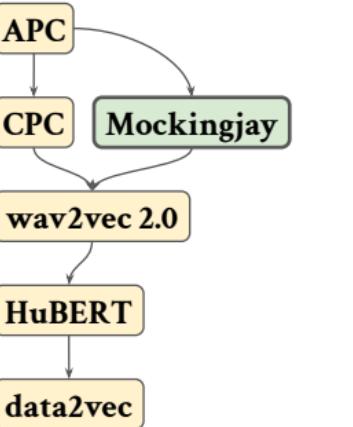
- └ a brief overview of unsupervised speech representation learning
  - └ Autoregressive Predictive Coding (APC)

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING  
**Autoregressive Predictive Coding (APC)**

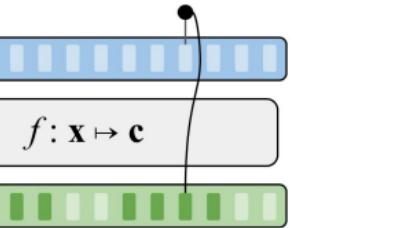
Challenges

- Encodes only past inputs ✗
- Uses the input as target ✗

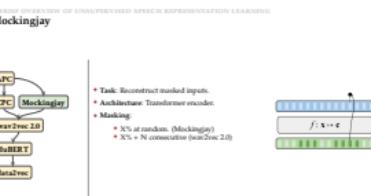
The diagram shows the APC architecture. It consists of three main components: APC, CPC, and Mockingjay. APC is at the top, followed by CPC and Mockingjay. Below them is wav2vec 2.0, then HuBERT, and finally data2vec at the bottom. A legend on the right defines the symbols: a box for 'Challenge', a red 'X' for 'Encodes only past inputs ✗', and another red 'X' for 'Uses the input as target ✗'. To the right of the legend is a small diagram of a neural network layer with the label  $f: \mathbf{x} \mapsto \mathbf{c}$ .

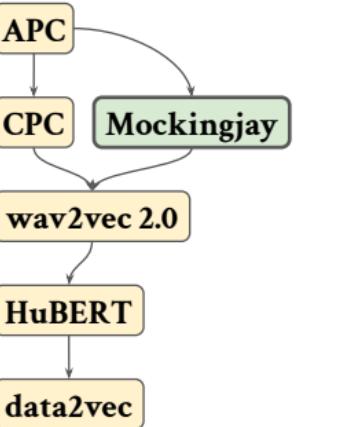


- **Task:** Reconstruct masked inputs.
- **Architecture:** Transformer encoder.
- **Masking:**
  - X% at random. (Mockingjay)
  - X% + N consecutive (wav2vec 2.0)

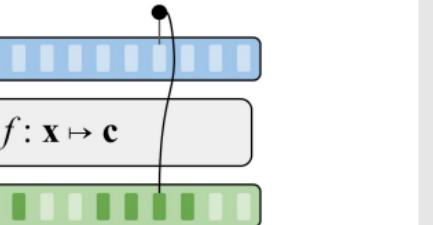


2024-03-05





- Challenges:
  - Encodes the entire input ✓
  - Uses the input as target ✗



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

└ a brief overview of unsupervised speech representation learning

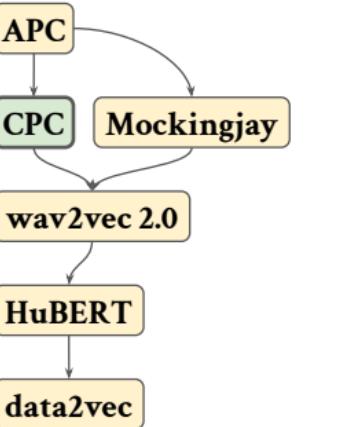
└ Mockingjay

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING  
**Mockingjay**

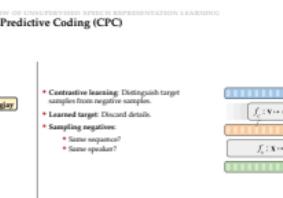
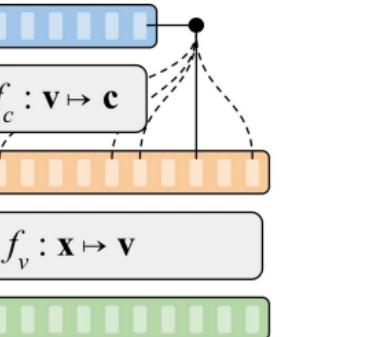
Challenges

- Encodes the entire input ✓
- Uses the input as target ✗

The diagram shows the Mockingjay architecture. It consists of an APC layer, a CPC layer, a Mockingjay layer, a wav2vec 2.0 layer, a HuBERT layer, and a data2vec layer. The Mockingjay layer is highlighted in red. A legend indicates that the Mockingjay layer is used for both encoding the entire input (✓) and using the input as target (✗).

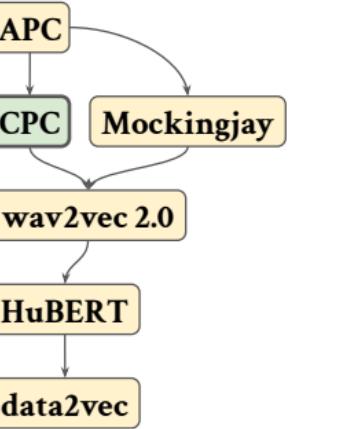


- **Contrastive learning:** Distinguish target samples from negative samples.
- **Learned target:** Discard details.
- **Sampling negatives:**
  - Same sequence?
  - Same speaker?

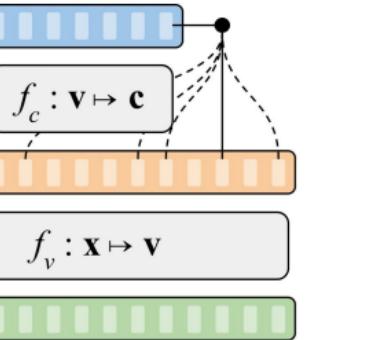


# A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

## Contrastive Predictive Coding (CPC)



- Challenges:
  - Only encodes past inputs ✗
  - Uses a learned target ✓



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

## a brief overview of unsupervised speech representation learning

### Contrastive Predictive Coding (CPC)

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING  
Contrastive Predictive Coding (CPC)

• Challenges:

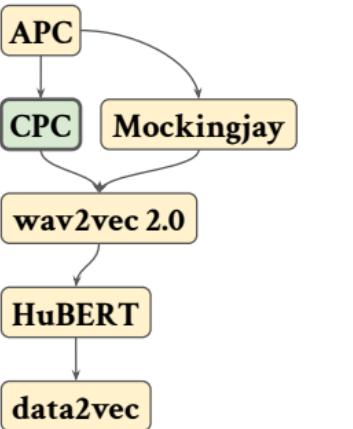
- Only encodes past inputs ✗
- Uses a learned target ✓

The diagram shows the CPC architecture with three parallel layers:  $f_c : \mathbf{v} \mapsto \mathbf{c}$  (blue),  $f_v : \mathbf{x} \mapsto \mathbf{v}$  (orange), and  $f_o : \mathbf{x} \mapsto \mathbf{o}$  (green). A legend indicates:

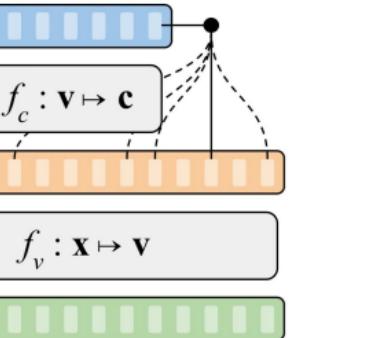
- Challenge: Only encodes past inputs ✗
- Challenge: Uses a learned target ✓

# A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

## Contrastive Predictive Coding (CPC)



- Challenges:
- Only encodes past inputs ✗
- Uses a learned target ✓
- Sampling negatives ✗

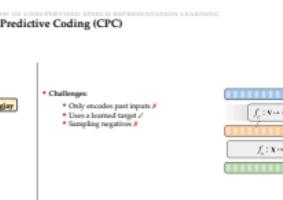


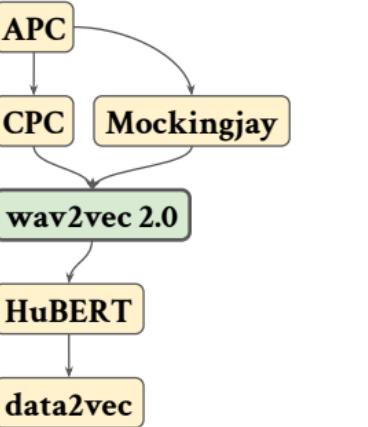
# UNCERTAINTY AND THE MEDICAL INTERVIEW

## a brief overview of unsupervised speech representation learning

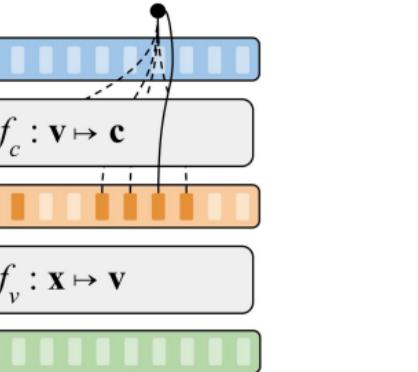
### Contrastive Predictive Coding (CPC)

2024-03-05

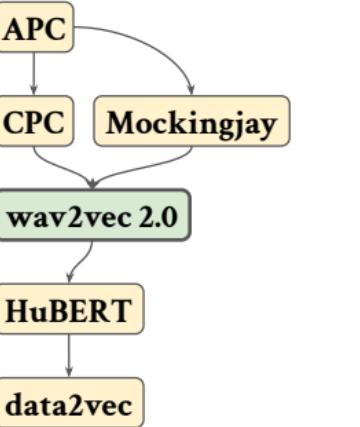




- Masking + contrastive learning.
- **Quantisation:** Better negative samples.
- **Results:**
  - 960 hours: **2.0%** WER.
  - 10 minutes: **4.8%** WER.

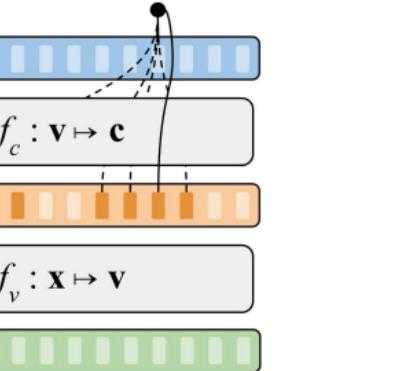


1. Training objective requires identifying the correct quantized latent audio representation in a set of distractors for each masked time step.
2. Quantisation improves negative sampling (requires approximation via Gumbel softmax).



- Challenges:

- Encodes the entire input ✓
- Uses a learned target ✓
- Sampling negatives ✗



2024-03-05

wav2vec 2.0



$f_c : \mathbf{v} \mapsto \mathbf{c}$

$f_v : \mathbf{x} \mapsto \mathbf{v}$

$f_d : \mathbf{X} \mapsto \mathbf{Y}$

$f_u : \mathbf{Y} \mapsto \mathbf{C}$

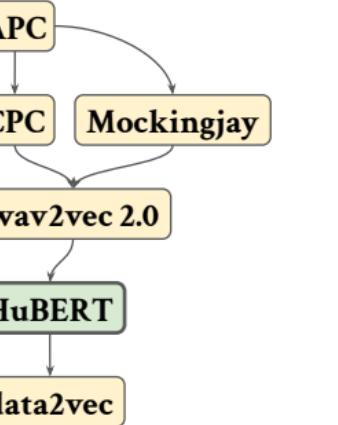
$f_t : \mathbf{C} \mapsto \mathbf{X}$

$f_{\text{HuBERT}} : \mathbf{X} \mapsto \mathbf{Y}$

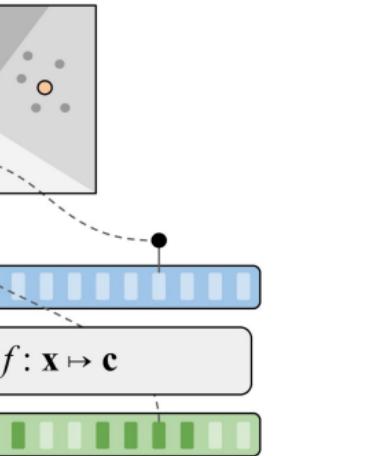
$f_{\text{data2vec}} : \mathbf{Y} \mapsto \mathbf{C}$

$f_{\text{APC}} : \mathbf{C} \mapsto \mathbf{X}$

## A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING Hidden-unit BERT (HuBERT)



- Target: K-means teacher (MFCC frames).
- Training: Cross-entropy loss.
- 1st iteration: K-means on inputs.



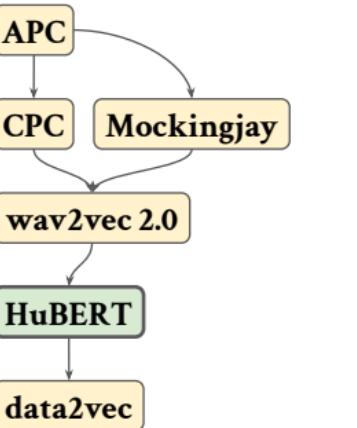
2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

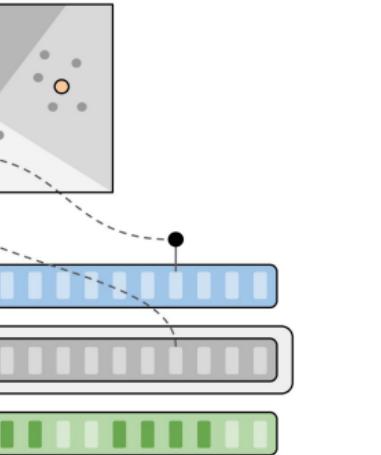
- └ a brief overview of unsupervised speech representation learning
  - └ Hidden-unit BERT (HuBERT)



## A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING Hidden-unit BERT (HuBERT)



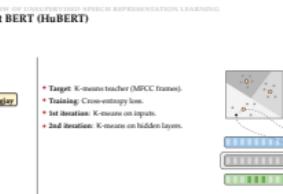
- Target: K-means teacher (MFCC frames).
- Training: Cross-entropy loss.
- 1st iteration: K-means on inputs.
- 2nd iteration: K-means on hidden layers.

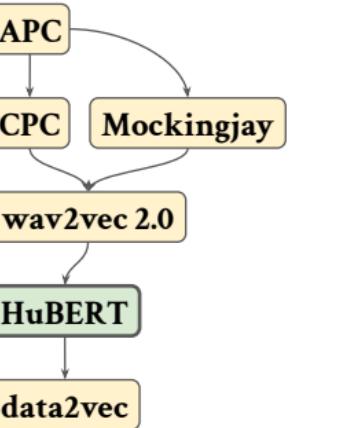


2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

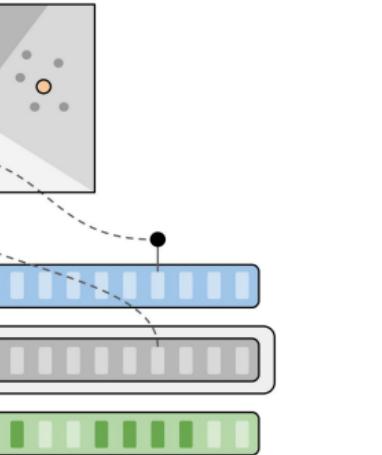
- └ a brief overview of unsupervised speech representation learning
  - └ Hidden-unit BERT (HuBERT)





• Challenges:

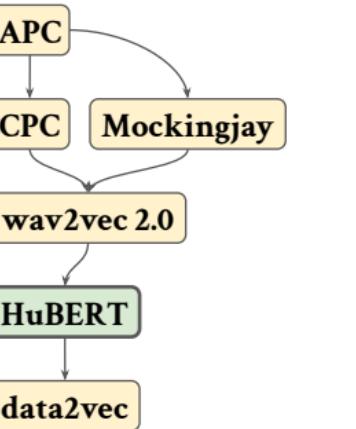
- Encodes the entire input ✓
- Uses a learned target ✓
- No need for negative samples ✓



2024-03-05

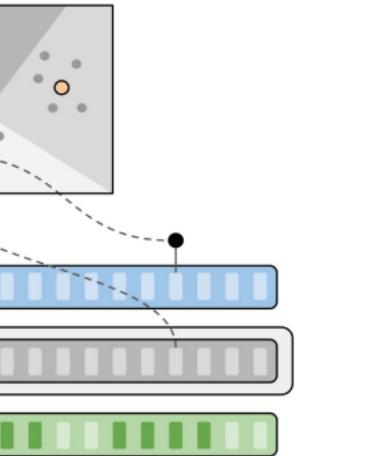


1. HuBERT approach predicts hidden cluster assignments of masked frames
2. Targets are still quantised although we no longer solve a contrastive sampling problem. Might reduce quality.



• Challenges:

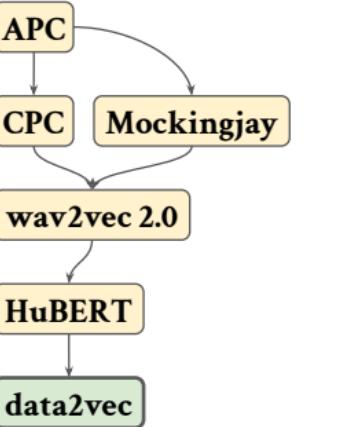
- Encodes the entire input ✓
- Uses a learned target ✓
- No need for negative samples ✓
- Targets updated infrequently ✗
- Quantized targets ✗



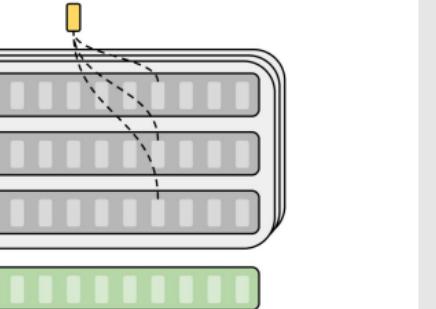
# UNCERTAINTY AND THE MEDICAL INTERVIEW

└ a brief overview of unsupervised speech representation learning  
└ Hidden-unit BERT (HuBERT)

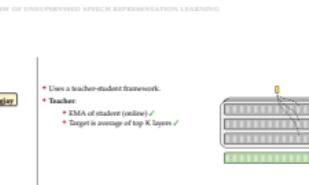


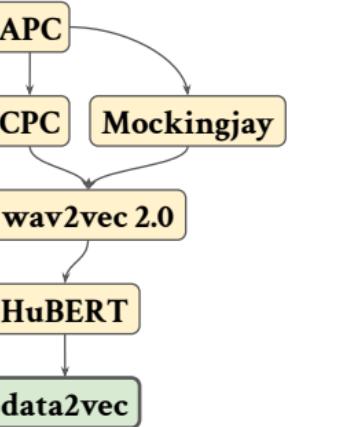


- Uses a teacher-student framework.
- Teacher:
  - EMA of student (online) ✓
  - Target is average of top K layers ✓

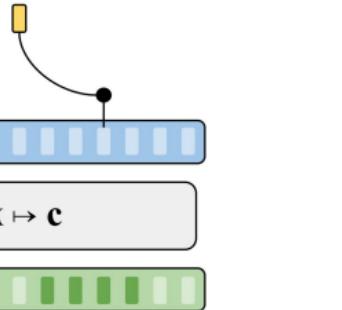


2024-03-05





- Uses a teacher-student framework.
- **Teacher:**
  - EMA of student (online) ✓
  - Target is average of top K layers ✓
- **Student training:** Smooth  $\ell_1$  loss.



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

## a brief overview of unsupervised speech representation learning

### data2vec

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING  
**data2vec**

• Uses a teacher-student framework.  
• Teacher:

- EMA of student (online) ✓
- Target is average of top K layers ✓

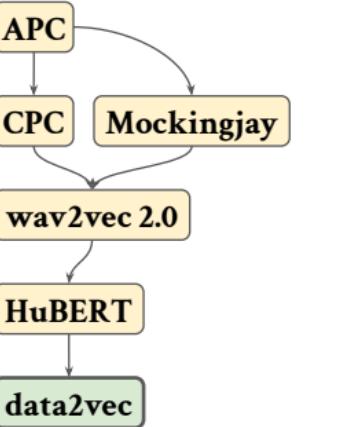
• Student training: Smooth  $\ell_1$  loss.

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING  
**data2vec**

• Uses a teacher-student framework.  
• Teacher:

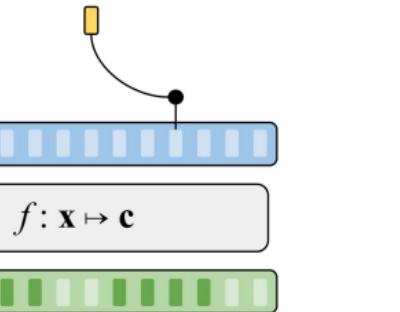
- EMA of student (online) ✓
- Target is average of top K layers ✓

• Student training: Smooth  $\ell_1$  loss.

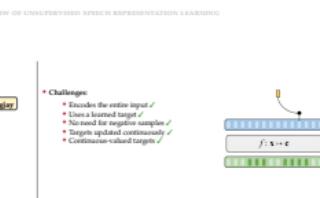


• Challenges:

- Encodes the entire input ✓
- Uses a learned target ✓
- No need for negative samples ✓
- Targets updated continuously ✓
- Continuous-valued targets ✓



2024-03-05



- **Main conclusions:**
  - The most popular self-supervised speech models can be compactly described by a few core design choices.
  - Many of these design choices are mirrored in earlier work on speech embedding models.
- **Open questions and limitations:**
  - Which design choices benefit which downstream tasks?
  - It is difficult to compare methods as model size and evaluation procedures differ widely between papers.

## UNCERTAINTY AND THE MEDICAL INTERVIEW

- └ a brief overview of unsupervised speech representation learning
- └ Conclusions

2024-03-05

- **Main conclusions:**
  - The most popular self-supervised speech models can be compactly described by a few core design choices.
  - Many of these design choices are mirrored in earlier work on speech embedding models.
- **Open questions and limitations:**
  - Which design choices benefit which downstream tasks?
  - It is difficult to compare methods as model size and evaluation procedures differ widely between papers.

# OVERVIEW Presentation

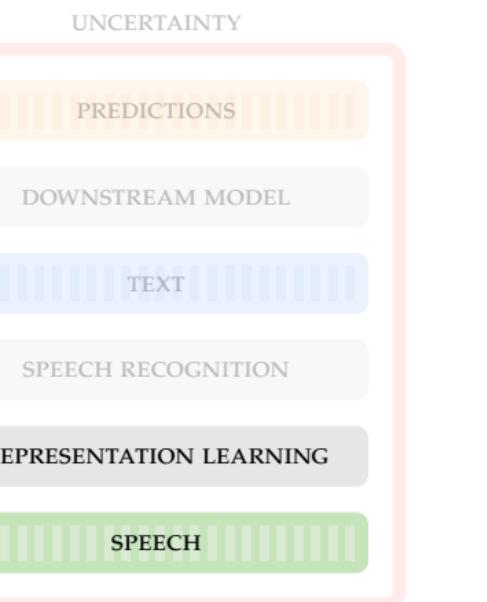
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH  
REPRESENTATION LEARNING

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-  
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

overview

Presentation

# OVERVIEW Presentation

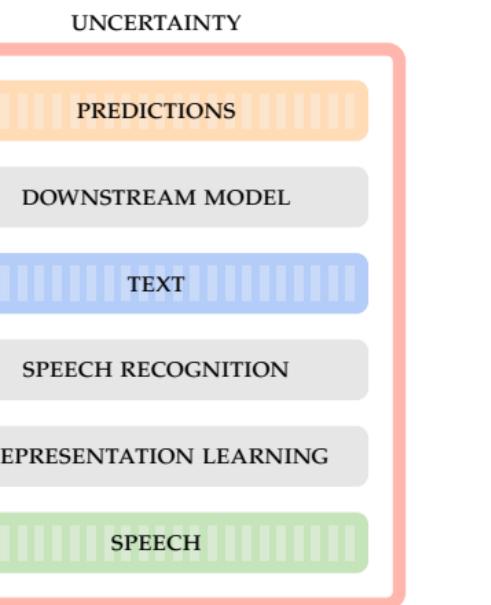
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH  
REPRESENTATION LEARNING

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-  
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION



UNCERTAINTY AND THE MEDICAL INTERVIEW

2024-03-05

overview

Presentation



- Stroke is the second leading cause of death (11.6%) and third leading cause of death and disability combined (5.7%) worldwide [20, 30, 35].
- Effective treatment is very **time-sensitive** [4, 56].
- The gateway to **ambulance transport and hospital admittance** is through **prehospital telehealth services**.
- **Mobile stroke units** have made it possible to deliver advanced treatment faster [22, 42].
- The effectiveness of mobile stroke units hinges on **call-taker recognition of stroke** [22, 42].
- Approximately half of all patients with stroke do not receive the correct triage for their condition from call-takers [7, 45, 58].

## UNCERTAINTY AND THE MEDICAL INTERVIEW

- 2024-03-05
- └ a retrospective study on machine learning-assisted stroke recognition for medical helpline calls
    - └ Stroke

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION  
FOR MEDICAL HELPLINE CALLS  
**Stroke**

- Stroke is the second leading cause of death (11.6%) and third leading cause of death and disability combined (5.7%) worldwide [20, 30, 35].
- Effective treatment is very **time-sensitive** [4, 56].
- The gateway to **ambulance transport and hospital admittance** is through **prehospital telehealth services**.
- **Mobile stroke units** have made it possible to deliver advanced treatment faster [22, 42].
- The effectiveness of mobile stroke units hinges on **call-taker recognition of stroke** [22, 42].
- Approximately half of all patients with stroke do not receive the correct triage for their condition from call-takers [7, 45, 58].

## The study

- Collaboration between **Corti** and the **Copenhagen Emergency Medical Services (CEMS)** ("Region Hovedstadens Akutberedskab").
- CEMS provides prehospital emergency services in the Capital Region of Denmark (1.9M people).
- CEMS operates the 1-1-2 emergency line (similar to 9-1-1) and the 1813 medical helpline (non-life-threatening conditions when the general practitioner is unavailable).
- We wanted to investigate if a machine learning model could assist call-takers of 1813 in recognizing stroke.



2024-03-05

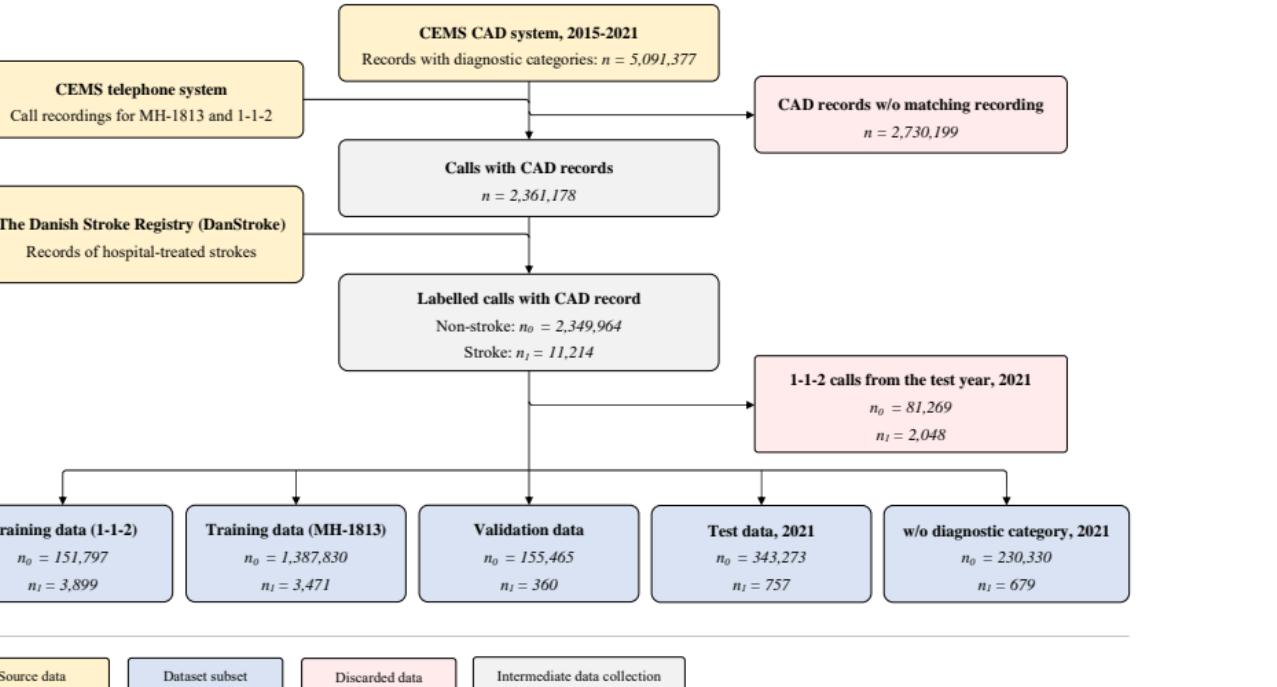
## UNCERTAINTY AND THE MEDICAL INTERVIEW

- └ a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

- └ The study

# A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

## Population selection and datasets



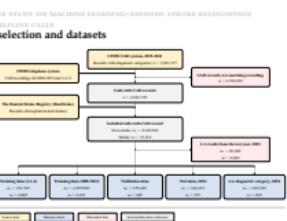
2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

- a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

## Population selection and datasets

1. Test data is MH-1813 2021.
2. All 1-1-2 data is used for training except 2021.
3. Validation data is sampled with stratified sampling from MH-1813 from 2015-2020.



## Population characteristics of test set

Subset	All calls	Stroke calls	Non-stroke
<b>Num. calls</b>	344,030	757	343,273
<b>Female</b>	190,974 (55.51%)	349 (46.10%)	190,625 (55.53%)
<b>Male</b>	153,050 (44.49%)	408 (53.90%)	152,642 (44.47%)
<b>65+ years</b>	65,652 (19.08%)	555 (73.32%)	65,097 (18.96%)
<b>Age (mean ± std.)</b>	44.31 ± 20.10	71.51 ± 13.41	44.25 ± 20.08

## UNCERTAINTY AND THE MEDICAL INTERVIEW

a retrospective study on machine learning-assisted stroke recognition  
for medical helpline calls

Population characteristics of test set

2024-03-05

Subset	All calls	Stroke calls	Non-stroke
Num. calls	344,030	757	343,273
Female	190,974 (55.51%)	349 (46.10%)	190,625 (55.53%)
Male	153,050 (44.49%)	408 (53.90%)	152,642 (44.47%)
65+ years	65,652 (19.08%)	555 (73.32%)	65,097 (18.96%)
Age (mean ± std.)	44.31 ± 20.10	71.51 ± 13.41	44.25 ± 20.08

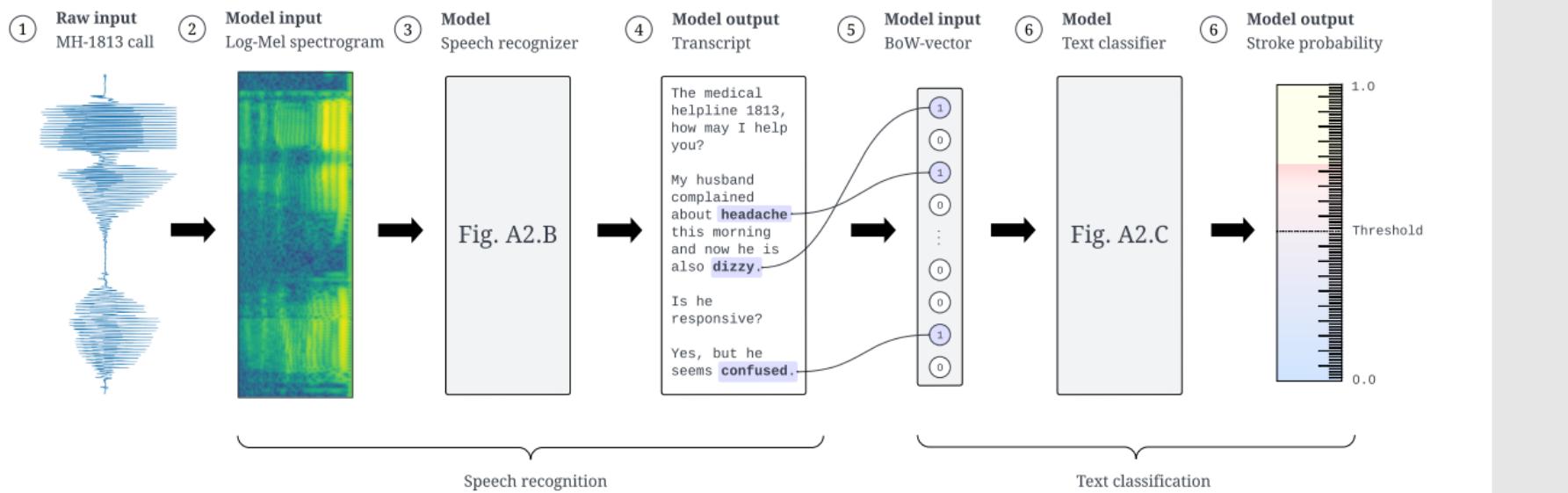
1. Prevalence of stroke is less than a quarter percent, one in every 400 calls.
2. The mean age of stroke calls is 71.5 years, older than general callers.
3. Males are a bit more likely to call with a stroke compared to females.
4. Other datasets (training, validation) have similar characteristics.

# A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

## Model design



### A. Schematic Overview of Stroke Classification Pipeline

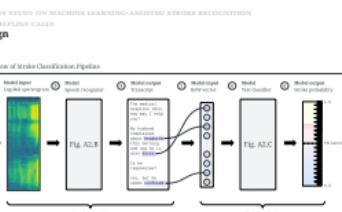


# UNCERTAINTY AND THE MEDICAL INTERVIEW

a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

Model design

1. Same structure as the decision-support system sketched in the overview slides.

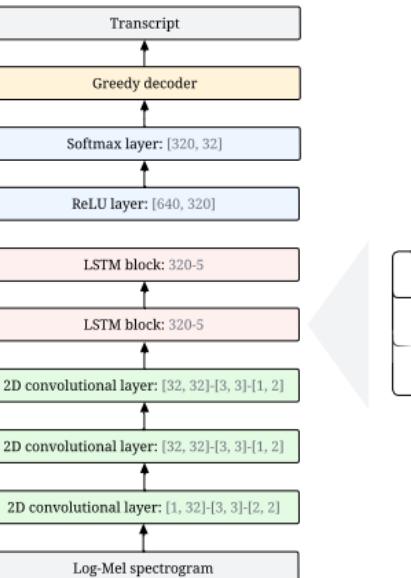


# A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

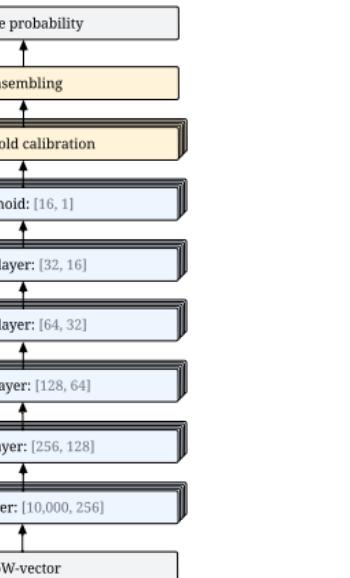
## Model design



B. Speech Recognition Model



C. Text Classification Model

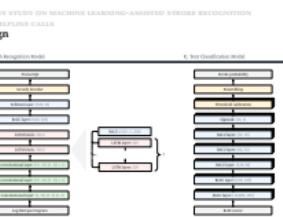


2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

- └ a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

- └ Model design

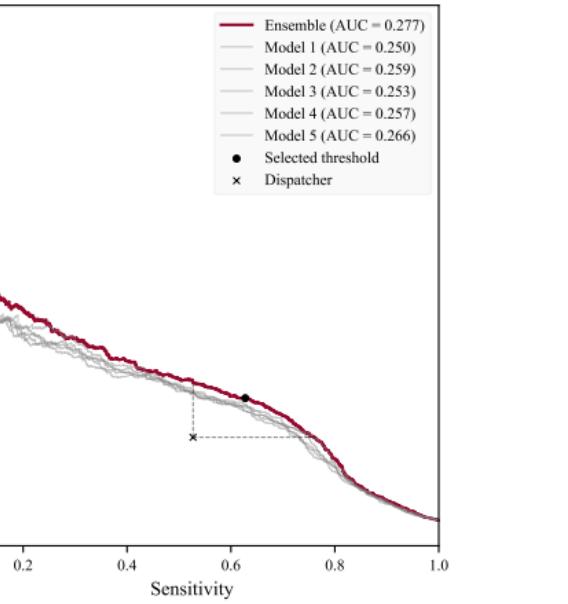
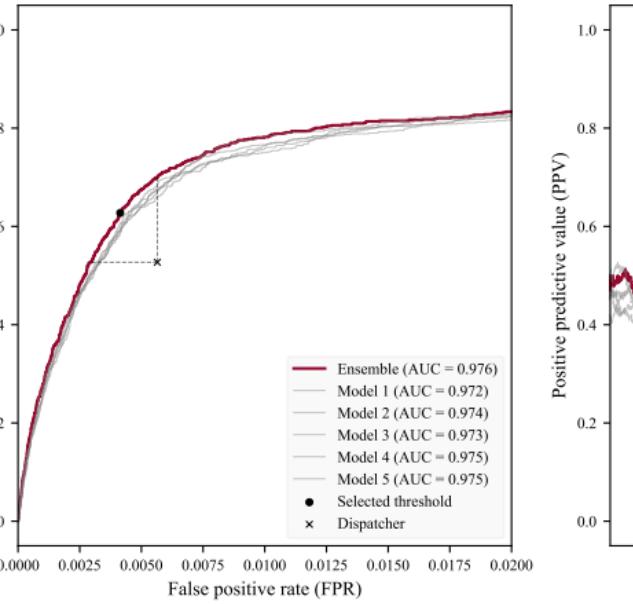


1. ASR model somewhat old-school, but performant for this task with noisy calls.
2. Classification model is ensemble of five models.

## A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

### Main results

ROC curve and PPV-sensitivity curve (precision-recall curve). Models 1-5 are the individual models that make up the ensemble model.



corti

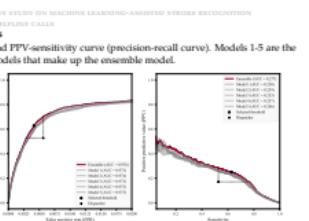
DTU

2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

### Main results



## Main results

Confusion matrices of predictions for call takers and the model on the test set. Numbers for the model are given as the rounded mean over eleven runs.

		Ground truth labels	
		Positives	Negatives
Call taker predictions	Positives	True positives 399	False positives 1,938
	Negatives	False negatives 358	True negatives 341,335

		Ground truth labels	
		Positives	Negatives
Model predictions	Positives	True positives 477	False positives 1,440
	Negatives	False negatives 280	True negatives 341,833

## UNCERTAINTY AND THE MEDICAL INTERVIEW

a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

>Main results

2024-03-05

Confusion matrices of predictions for call takers and the model on the test set. Numbers for the model are given as the rounded mean over eleven runs.			
Call takers		Model	
		Positives	Negatives
True positives	399	477	477
False positives	1,938	1,440	1,440
True negatives	341,335	341,833	341,833
False negatives	358	280	280

1. In absolute numbers, the model correctly identifies 78 cases of stroke missed by call-takers.
2. The model also reduces the number of false positives by about 500 from 1938 to 1440.

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION  
FOR MEDICAL HELPLINE CALLS

## Main results

MH-1813 test set performance in demographic subgroups (age/sex) [mean (95% CI)].

Subset	Predictor	F1-score [%] ↑	Sensitivity [%] ↑	PPV [%] ↑	FOR [%] ↓ (1 - specificity)	FPR [%] ↓ (1 - NPV)
Overall	<b>Call-takers</b>	25.8 (23.7-27.9)	52.7 (49.2-56.4)	17.1 (15.5-18.6)	0.105 (0.094-0.116)	0.565 (0.539-0.590)
	<b>Model</b>	35.7 (35.0-36.4)	63.0 (62.0-64.1)	24.9 (24.3-25.5)	0.082 (0.079-0.085)	0.419 (0.413-0.426)
18-64 years	<b>Call-takers</b>	15.9 (13.1-18.5)	50.5 (43.6-57.2)	9.40 (7.61-11.2)	0.036 (0.028-0.043)	0.353 (0.331-0.375)
	<b>Model</b>	22.9 (21.8-24.0)	54.1 (52.1-56.3)	14.5 (13.8-15.3)	0.033 (0.031-0.035)	0.231 (0.226-0.236)
65+ years	<b>Call-takers</b>	32.9 (30.1-35.7)	53.5 (49.4-57.6)	23.7 (21.4-26.0)	0.401 (0.352-0.449)	1.467 (1.373-1.560)
	<b>Model</b>	42.8 (41.9-43.7)	66.3 (65.1-67.5)	31.6 (30.8-32.4)	0.290 (0.278-0.303)	1.224 (1.198-1.249)
Male	<b>Call-takers</b>	30.2 (27.2-33.3)	53.9 (49.1-58.9)	21.0 (18.5-23.5)	0.124 (0.105-0.141)	0.542 (0.506-0.580)
	<b>Model</b>	39.0 (38.0-40.1)	63.7 (62.3-65.2)	28.1 (27.3-29.0)	0.097 (0.093-0.102)	0.435 (0.425-0.445)
Female	<b>Call-takers</b>	21.9 (19.1-24.6)	51.3 (46.0-56.6)	13.9 (12.0-15.8)	0.090 (0.076-0.103)	0.582 (0.547-0.616)
	<b>Model</b>	32.4 (31.4-33.4)	62.3 (60.7-63.8)	21.9 (21.1-22.7)	0.069 (0.066-0.073)	0.407 (0.399-0.416)



## UNCERTAINTY AND THE MEDICAL INTERVIEW

a retrospective study on machine learning-assisted stroke recognition  
for medical helpline calls

>Main results

2024-03-05

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS					
Main results					
MH-1813 test set performance in demographic subgroups (age/sex) [mean (95% CI)].					
Subset	Predictor	F1-score [%] ↑	Sensitivity [%] ↑	PPV [%] ↑	FOR [%] ↓ (1 - specificity)
Overall	Call-takers	25.8 (23.7-27.9)	52.7 (49.2-56.4)	17.1 (15.5-18.6)	0.105 (0.094-0.116)
Overall	Model	35.7 (35.0-36.4)	63.0 (62.0-64.1)	24.9 (24.3-25.5)	0.082 (0.079-0.085)
18-64 years	Call-takers	15.9 (13.1-18.5)	50.5 (43.6-57.2)	9.40 (7.61-11.2)	0.036 (0.028-0.043)
18-64 years	Model	22.9 (21.8-24.0)	54.1 (52.1-56.3)	14.5 (13.8-15.3)	0.033 (0.031-0.035)
65+ years	Call-takers	32.9 (30.1-35.7)	53.5 (49.4-57.6)	23.7 (21.4-26.0)	0.401 (0.352-0.449)
65+ years	Model	42.8 (41.9-43.7)	66.3 (65.1-67.5)	31.6 (30.8-32.4)	0.290 (0.278-0.303)
Male	Call-takers	30.2 (27.2-33.3)	53.9 (49.1-58.9)	21.0 (18.5-23.5)	0.124 (0.105-0.141)
Male	Model	39.0 (38.0-40.1)	63.7 (62.3-65.2)	28.1 (27.3-29.0)	0.097 (0.093-0.102)
Female	Call-takers	21.9 (19.1-24.6)	51.3 (46.0-56.6)	13.9 (12.0-15.8)	0.090 (0.076-0.103)
Female	Model	32.4 (31.4-33.4)	62.3 (60.7-63.8)	21.9 (21.1-22.7)	0.069 (0.066-0.073)

## Occlusion analysis — Which features are evidence?



Features with positive ranking score ( $r^{(w)} > 0$ ) computed on stroke positive predictions ( $D = 1,897$ )					
Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$	Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$
1.	Ambulance	1,680	16.	Difficulties speaking	44
2.	Blood clot	895	17.	Hemorrhagic stroke	133
3.	Left	1,108	18.	Hand	297
4.	Right	1,050	19.	The ambulance	521
5.	Double vision	84	20.	Slurred speech	58
6.	The words	344	21.	Blood clots	224
7.	Suddenly	783	22.	Fast	663
8.	Arm	709	23.	Express	44
9.	Side	1,139	24.	Blood thinner	259
10.	Stroke	117	25.	Incoherent	15
11.	Double	113	26.	Lopsided	211
12.	Control	134	27.	Reduced	528
13.	Call	39	28.	Hangs	628
14.	Numb	94	29.	Transient	48
15.	Minutes	763	30.	Not making sense	14

[Recognition, Symptom, Urgency/Time]

## UNCERTAINTY AND THE MEDICAL INTERVIEW

— a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

— Occlusion analysis — Which features are evidence?

2024-03-05

Occlusion analysis — Which features are evidence? (D = 1,897)					
Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$	Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$
1.	Ambulance	1,680	16.	Difficulties speaking	44
2.	Blood clot	895	17.	Hemorrhagic stroke	133
3.	Left	1,108	18.	Hand	297
4.	Right	1,050	19.	The ambulance	521
5.	Double vision	84	20.	Slurred speech	58
6.	The words	344	21.	Blood clots	224
7.	Suddenly	783	22.	Fast	663
8.	Arm	709	23.	Express	44
9.	Side	1,139	24.	Blood thinner	259
10.	Stroke	117	25.	Incoherent	15
11.	Double	113	26.	Lopsided	211
12.	Control	134	27.	Reduced	528
13.	Call	39	28.	Hangs	628
14.	Numb	94	29.	Transient	48
15.	Minutes	763	30.	Not making sense	14

[Recognition, Symptom, Urgency/Time]

- Wanted to examine the features that were important for predictions.
- Performed an occlusion analysis where we remove one word from the input at a time.
- Sort the words by their impact on the model's output logit.

—

## Occlusion analysis — Which features are evidence?



Features with positive ranking score ( $r^{(w)} > 0$ ) computed on stroke positive predictions ( $D = 1,897$ )					
Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$	Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$
1.	Ambulance	1,680	16.	Difficulties speaking	44
2.	Blood clot	895	17.	Hemorrhagic stroke	133
3.	Left	1,108	18.	Hand	297
4.	Right	1,050	19.	The ambulance	521
5.	Double vision	84	20.	Slurred speech	58
6.	The words	344	21.	Blood clots	224
7.	Suddenly	783	22.	Fast	663
8.	Arm	709	23.	Express	44
9.	Side	1,139	24.	Blood thinner	259
10.	Stroke	117	25.	Incoherent	15
11.	Double	113	26.	Lopsided	211
12.	Control	134	27.	Reduced	528
13.	Call	39	28.	Hangs	628
14.	Numb	94	29.	Transient	48
15.	Minutes	763	30.	Not making sense	14

[Recognition, Symptom, Urgency/Time]

## UNCERTAINTY AND THE MEDICAL INTERVIEW

— a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

— Occlusion analysis — Which features are evidence?

2024-03-05

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS					
Occlusion analysis — Which features are evidence?					
Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$	Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$
1.	Ambulance	1,680	16.	Difficulties speaking	44
2.	Blood clot	895	17.	Hemorrhagic stroke	133
3.	Left	1,108	18.	Hand	297
4.	Right	1,050	19.	The ambulance	521
5.	Double vision	84	20.	Slurred speech	58
6.	The words	344	21.	Blood clots	224
7.	Suddenly	783	22.	Fast	663
8.	Arm	709	23.	Express	44
9.	Side	1,139	24.	Blood thinner	259
10.	Stroke	117	25.	Incoherent	15
11.	Double	113	26.	Lopsided	211
12.	Control	134	27.	Reduced	528
13.	Call	39	28.	Hangs	628
14.	Numb	94	29.	Transient	48
15.	Minutes	763	30.	Not making sense	14

[Recognition, Symptom, Urgency/Time]

- Wanted to examine the features that were important for predictions.
- Performed an occlusion analysis where we remove one word from the input at a time.
- Sort the words by their impact on the model's output logit.

## Occlusion analysis — Which features are evidence?



Features with positive ranking score ( $r^{(w)} > 0$ ) computed on stroke positive predictions ( $D = 1,897$ )					
Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$	Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$
1.	Ambulance	1,680	16.	Difficulties speaking	44
2.	Blood clot	895	17.	Hemorrhagic stroke	133
3.	Left	1,108	18.	Hand	297
4.	Right	1,050	19.	The ambulance	521
5.	Double vision	84	20.	Slurred speech	58
6.	The words	344	21.	Blood clots	224
7.	Suddenly	783	22.	Fast	663
8.	Arm	709	23.	Express	44
9.	Side	1,139	24.	Blood thinner	259
10.	Stroke	117	25.	Incoherent	15
11.	Double	113	26.	Lopsided	211
12.	Control	134	27.	Reduced	528
13.	Call	39	28.	Hangs	628
14.	Numb	94	29.	Transient	48
15.	Minutes	763	30.	Not making sense	14

## UNCERTAINTY AND THE MEDICAL INTERVIEW

— a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

— Occlusion analysis — Which features are evidence?

2024-03-05

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS			
Occlusion analysis — Which features are evidence?			
<small>Features with positive ranking score (<math>r^{(w)} &gt; 0</math>) computed on stroke positive predictions (<math>D = 1,897</math>)</small>			
Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$	Rank
1.	Ambulance	1,680	36.
2.	Blood clot	895	17.
3.	Left	1,108	133
4.	Right	1,050	297
5.	Double vision	84	521
6.	The words	344	224
7.	Suddenly	783	663
8.	Arm	709	44
9.	Side	1,139	259
10.	Stroke	117	15
11.	Double	113	211
12.	Control	134	528
13.	Call	39	628
14.	Numb	94	48
15.	Minutes	763	14
<small>[Recognition, Symptom, Urgency/Time]</small>			

1. Wanted to examine the features that were important for predictions.
2. Performed an occlusion analysis where we remove one word from the input at a time.
3. Sort the words by their impact on the model's output logit.

## Occlusion analysis — Which features are evidence?



Features with positive ranking score ( $r^{(w)} > 0$ ) computed on stroke positive predictions ( $D = 1,897$ )					
Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$	Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$
1.	Ambulance	1,680	16.	Difficulties speaking	44
2.	Blood clot	895	17.	Hemorrhagic stroke	133
3.	Left	1,108	18.	Hand	297
4.	Right	1,050	19.	The ambulance	521
5.	Double vision	84	20.	Slurred speech	58
6.	The words	344	21.	Blood clots	224
7.	Suddenly	783	22.	Fast	663
8.	Arm	709	23.	Express	44
9.	Side	1,139	24.	Blood thinner	259
10.	Stroke	117	25.	Incoherent	15
11.	Double	113	26.	Lopsided	211
12.	Control	134	27.	Reduced	528
13.	Call	39	28.	Hangs	628
14.	Numb	94	29.	Transient	48
15.	Minutes	763	30.	Not making sense	14

[Recognition, Symptom, Urgency/Time]

## UNCERTAINTY AND THE MEDICAL INTERVIEW

— a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

— Occlusion analysis — Which features are evidence?

2024-03-05

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS			
Occlusion analysis — Which features are evidence?			
<small>Features with positive ranking score (<math>r^{(w)} &gt; 0</math>) computed on stroke positive predictions (<math>D = 1,897</math>)</small>			
Rank	Word, $w$ (translated)	Occurrences, $D^{(w)}$	Rank
1.	Ambulance	1,680	16.
2.	Blood clot	895	17.
3.	Left	1,108	18.
4.	Right	1,050	19.
5.	Double vision	84	20.
6.	The words	344	21.
7.	Suddenly	783	22.
8.	Arm	709	23.
9.	Side	1,139	24.
10.	Stroke	117	25.
11.	Double	113	26.
12.	Control	134	27.
13.	Call	39	28.
14.	Numb	94	29.
15.	Minutes	763	30.

[Recognition, Symptom, Urgency/Time]

1. Wanted to examine the features that were important for predictions.
2. Performed an occlusion analysis where we remove one word from the input at a time.
3. Sort the words by their impact on the model's output logit.

## Occlusion analysis – Which features are counter-evidence?



Features with negative ranking score ( $r^{(w)} < 0$ ) computed on stroke-negative predictions (D = 342,133)					
Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>	Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>
1.	Tetanus	4,378	16.	The pharmacy	10,085
2.	Pregnant	8,749	17.	The stomach	42,105
3.	Cut	7,592	18.	Psychiatric	3,688
4.	Bandage	4,561	19.	Pneumonia	7,597
5.	Amager (a location)	23,776	20.	Stomach pain	10,551
6.	O'clock	94,436	21.	Stool	19,155
7.	The emergency room	42,809	22.	The ribs	3,928
8.	The police	2,903	23.	Bleed	10,501
9.	Swollen	60,559	24.	Bleeding	24,313
10.	Over the counter (OTC)	4,641	25.	Ribs	2,941
11.	The neck	30,151	26.	Broken	19,415
12.	Fever	112,586	27.	Inflammation	10,050
13.	Prescription	5,450	28.	Common cold	8,127
14.	Centimeter	12,026	29.	Morning or morrow	78,558
15.	The knee	8,875	30.	Swelling	17,762

[Recognition, Symptom, Urgency/Time]

## UNCERTAINTY AND THE MEDICAL INTERVIEW

↳ a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

↳ Occlusion analysis – Which features are counter-evidence?

2024-03-05

Occlusion analysis – Which features are counter-evidence?					
Features with negative ranking score ( $r^{(w)} < 0$ ) computed on stroke-negative predictions (D = 342,133)					
Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>	Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>
1.	Tetanus	8,079	16.	The pharmacy	41,000
2.	Pregnant	8,749	17.	The stomach	42,105
3.	Cut	7,592	18.	Psychiatric	3,688
4.	Bandage	4,561	19.	Pneumonia	7,597
5.	Amager (a location)	23,776	20.	Stomach pain	10,551
6.	O'clock	94,436	21.	Bleed	19,155
7.	The emergency room	42,809	22.	The ribs	3,928
8.	The police	2,903	23.	Bleeding	24,313
9.	Swollen	60,559	24.	Ribs	2,941
10.	Over the counter (OTC)	4,641	25.	Broken	19,415
11.	The neck	30,151	26.	Inflammation	10,050
12.	Fever	112,586	27.	Common cold	8,127
13.	Prescription	5,450	28.	Morning or morrow	78,558
14.	Centimeter	12,026	29.	Swelling	17,762
15.	The knee	8,875	30.	Boiling	27,742

[Recognition, Symptom, Urgency/Time]

## Occlusion analysis – Which features are counter-evidence?



Features with negative ranking score ( $r^{(w)} < 0$ ) computed on stroke-negative predictions (D = 342,133)					
Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>	Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>
1.	Tetanus	4,378	16.	The pharmacy	10,085
2.	Pregnant	8,749	17.	The stomach	42,105
3.	Cut	7,592	18.	Psychiatric	3,688
4.	Bandage	4,561	19.	Pneumonia	7,597
5.	Amager (a location)	23,776	20.	Stomach pain	10,551
6.	O'clock	94,436	21.	Stool	19,155
7.	The emergency room	42,809	22.	The ribs	3,928
8.	The police	2,903	23.	Bleed	10,501
9.	Swollen	60,559	24.	Bleeding	24,313
10.	Over the counter (OTC)	4,641	25.	Ribs	2,941
11.	The neck	30,151	26.	Broken	19,415
12.	Fever	112,586	27.	Inflammation	10,050
13.	Prescription	5,450	28.	Common cold	8,127
14.	Centimeter	12,026	29.	Morning or morrow	78,558
15.	The knee	8,875	30.	Swelling	17,762

[Recognition, Symptom, Urgency/Time]

## UNCERTAINTY AND THE MEDICAL INTERVIEW

↳ a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

↳ Occlusion analysis – Which features are counter-evidence?

2024-03-05

Occlusion analysis – Which features are counter-evidence?					
Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>	Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>
1.	Arteries	8,079	16.	The pharmacy	41,068
2.	Pregnant	8,749	17.	The stomach	42,105
3.	Bandage	4,561	18.	Therapeutics	7,597
4.	Amager (a location)	23,776	19.	Over-the-counter	39,155
5.	O'clock	94,436	20.	Head	3,048
6.	The emergency room	42,809	21.	The ribs	3,928
7.	The police	2,903	22.	Bladder	24,313
8.	Swollen	60,559	23.	Backache	7,597
9.	Over the counter (OTC)	4,641	24.	Bleeding	10,050
10.	The neck	30,151	25.	Broken	19,415
11.	Fever	112,586	26.	Common cold	8,127
12.	Prescription	5,450	27.	Common condition	39,155
13.	Centimeter	12,026	28.	Common cold	8,127
14.	The knee	8,875	29.	Morning or morrow	78,558
15.	Handwriting	8,079	30.	Swelling	17,762

[Recognition, Symptom, Urgency/Time]

## Occlusion analysis – Which features are counter-evidence?



Features with negative ranking score ( $r^{(w)} < 0$ ) computed on stroke-negative predictions (D = 342,133)					
Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>	Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>
1.	Tetanus	4,378	16.	The pharmacy	10,085
2.	Pregnant	8,749	17.	The stomach	42,105
3.	Cut	7,592	18.	Psychiatric	3,688
4.	Bandage	4,561	19.	Pneumonia	7,597
5.	Amager (a location)	23,776	20.	Stomach pain	10,551
6.	O'clock	94,436	21.	Stool	19,155
7.	The emergency room	42,809	22.	The ribs	3,928
8.	The police	2,903	23.	Bleed	10,501
9.	Swollen	60,559	24.	Bleeding	24,313
10.	Over the counter (OTC)	4,641	25.	Ribs	2,941
11.	The neck	30,151	26.	Broken	19,415
12.	Fever	112,586	27.	Inflammation	10,050
13.	Prescription	5,450	28.	Common cold	8,127
14.	Centimeter	12,026	29.	Morning or tomorrow	78,558
15.	The knee	8,875	30.	Swelling	17,762

## UNCERTAINTY AND THE MEDICAL INTERVIEW

a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

↳ Occlusion analysis – Which features are counter-evidence?

2024-03-05

Features with negative ranking score ( $r^{(w)} < 0$ ) computed on stroke-negative predictions (D = 342,133)					
Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>	Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>
1.	Tetanus	8,078	16.	The pharmacy	41,000
2.	Pregnant	8,749	17.	The stomach	41,000
3.	Cut	7,592	18.	Psychiatric	3,688
4.	Bandage	4,561	19.	Pneumonia	7,597
5.	Amager (a location)	23,776	20.	Stomach pain	10,551
6.	O'clock	94,436	21.	Bleed	19,155
7.	The emergency room	42,809	22.	The ribs	3,928
8.	The police	2,903	23.	Bleeding	24,313
9.	Swollen	60,559	24.	Ribs	2,941
10.	Over the counter (OTC)	4,641	25.	Broken	19,415
11.	The neck	30,151	26.	Inflammation	10,050
12.	Fever	112,586	27.	Common cold	8,127
13.	Prescription	5,450	28.	Morning or tomorrow	78,558
14.	Centimeter	12,026	29.	Swelling	17,762
15.	The knee	8,875	30.	Blushing or sweating	76,908

[Recognition, Symptom, Urgency/Time]

## Occlusion analysis – Which features are counter-evidence?



Features with negative ranking score ( $r^{(w)} < 0$ ) computed on stroke-negative predictions (D = 342,133)					
Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>	Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>
1.	Tetanus	4,378	16.	The pharmacy	10,085
2.	Pregnant	8,749	17.	The stomach	42,105
3.	Cut	7,592	18.	Psychiatric	3,688
4.	Bandage	4,561	19.	Pneumonia	7,597
5.	Amager (a location)	23,776	20.	Stomach pain	10,551
6.	O'clock	94,436	21.	Stool	19,155
7.	The emergency room	42,809	22.	The ribs	3,928
8.	The police	2,903	23.	Bleed	10,501
9.	Swollen	60,559	24.	Bleeding	24,313
10.	Over the counter (OTC)	4,641	25.	Ribs	2,941
11.	The neck	30,151	26.	Broken	19,415
12.	Fever	112,586	27.	Inflammation	10,050
13.	Prescription	5,450	28.	Common cold	8,127
14.	Centimeter	12,026	29.	Morning or morrow	78,558
15.	The knee	8,875	30.	Swelling	17,762

[Recognition, Symptom, Urgency/Time]

## UNCERTAINTY AND THE MEDICAL INTERVIEW

↳ a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

↳ Occlusion analysis – Which features are counter-evidence?

2024-03-05

Occlusion analysis – Which features are counter-evidence?					
Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>	Rank	Word, w (translated)	Occurrences, D <sup>(w)</sup>
1.	Stroke	8,079	16.	The pharmacy	41,068
2.	Pregnant	8,749	17.	The stomach	42,105
3.	Amager	23,776	18.	Psychiatric	3,688
4.	Bandage	4,561	19.	Pneumonia	7,597
5.	O'clock	94,436	20.	Stomach pain	10,551
6.	The emergency room	42,809	21.	Stool	19,155
7.	The ribs	3,928	22.	The ribs	3,928
8.	Bleed	10,501	23.	Stool	19,155
9.	Bleeding	24,313	24.	Swelling	24,313
10.	Over the counter (OTC)	4,641	25.	Ribs	2,941
11.	The neck	30,151	26.	Broken	19,415
12.	Fever	112,586	27.	Inflammation	10,050
13.	Prescription	5,450	28.	Common cold	8,127
14.	Centimeter	12,026	29.	Morning or morrow	78,558
15.	The knee	8,875	30.	Swelling	17,762

[Recognition, Symptom, Urgency/Time]

## Future work

- Machine learning
  - Learning to predict directly from audio data (SSL).
  - Learning to defer to predict methods [57].
- Clinical applications
  - Mental health: Screening for suicide risk in emergency and medical helpline calls.
  - Maternity ward: Screening for serious pregnancy complications.



2024-03-05

**UNCERTAINTY AND THE MEDICAL INTERVIEW**  
└ a retrospective study on machine learning-assisted stroke recognition  
  for medical helpline calls  
    └ Future work

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION  
FOR MEDICAL HELPLINE CALLS  
**Future work**

- Machine learning
  - Learning to predict directly from audio data (SSL).
  - Learning to defer to predict methods [57].
- Clinical applications
  - Mental health: Screening for suicide risk in emergency and medical helpline calls.
  - Maternity ward: Screening for serious pregnancy complications.

# OVERVIEW Presentation

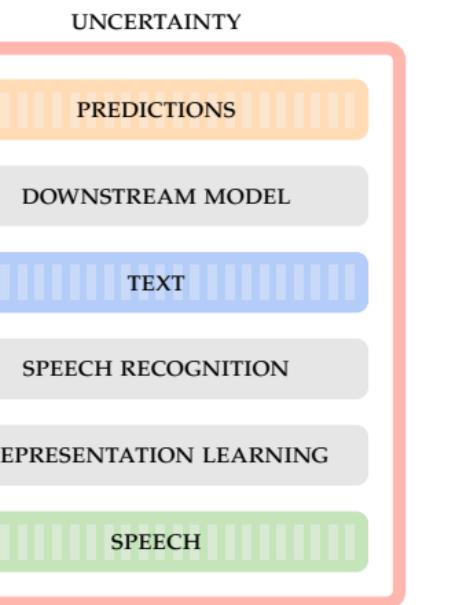
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH  
REPRESENTATION LEARNING

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-  
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION



UNCERTAINTY AND THE MEDICAL INTERVIEW

2024-03-05

overview

Presentation

DTU

UNCERTAINTY

PREDICTIONS

DOWNSTREAM MODEL

TEXT

SPEECH RECOGNITION

REPRESENTATION LEARNING

SPEECH

INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH

REPRESENTATION LEARNING

A RETROSPECTIVE STUDY ON MACHINE LEARNING-  
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

DISCUSSION AND CONCLUSION

# OVERVIEW Presentation

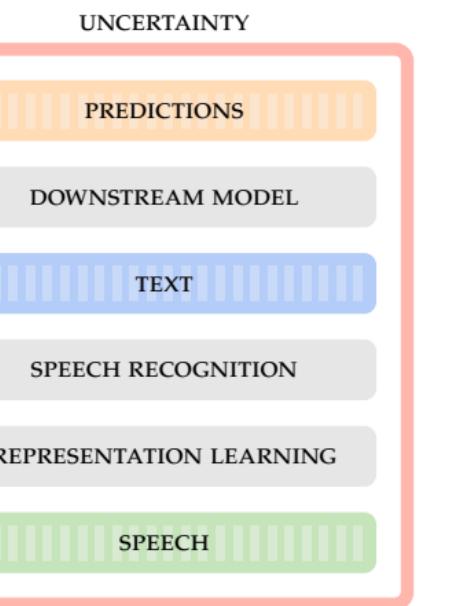
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

CHAPTER 4 HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH  
REPRESENTATION LEARNING

CHAPTER 9 A RETROSPECTIVE STUDY ON MACHINE LEARNING-  
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

CHAPTER 10 DISCUSSION AND CONCLUSION



UNCERTAINTY AND THE MEDICAL INTERVIEW

2024-03-05

overview

Presentation

UNCERTAINTY

PREDICTIONS

DOWNSTREAM MODEL

TEXT

SPEECH RECOGNITION

REPRESENTATION LEARNING

SPEECH

DISCUSSION AND CONCLUSION

INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH  
REPRESENTATION LEARNING

A RETROSPECTIVE STUDY ON MACHINE LEARNING-  
ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

DISCUSSION AND CONCLUSION



## The broad picture: The thesis topic since 2020



### 2020 Project start

- Out-of-distribution detection with generative models: Mysterious new topic.
- Speech representation/recognition: Inflection point between supervised methods and new self-supervised approaches.

### └ discussion

#### └ The broad picture: The thesis topic since 2020

2024-03-05

## The broad picture: The thesis topic since 2020



### 2020 Project start

- Out-of-distribution detection with generative models: Mysterious new topic.
- Speech representation/recognition: Inflection point between supervised methods and new self-supervised approaches.

### 2024 Project end

- Out-of-distribution detection is a mature field with a wide range of methods.
- Self-supervised learning is the dominant paradigm in speech recognition - challenged by weak labelling.

### └ discussion

#### └ The broad picture: The thesis topic since 2020

2024-03-05

## DISCUSSION

### The broad picture: The thesis topic since 2020



#### 2020 Project start

- Out-of-distribution detection with generative models: Mysterious new topic.
- Speech representation/recognition: Inflection point between supervised methods and new self-supervised approaches.

#### 2024 Project end

- Out-of-distribution detection is a mature field with a wide range of methods.
- Self-supervised learning is the dominant paradigm in speech recognition - challenged by weak labelling.
- Clinical research is increasingly becoming interested in the use of machine learning.

## UNCERTAINTY AND THE MEDICAL INTERVIEW

### ↳ discussion

#### ↳ The broad picture: The thesis topic since 2020

2024-03-05

## What lies ahead



### • Selective out-of-distribution detection

Two pairs of distributions may have identical divergence, but in different dimensions. How do we control features in black-box models?

### • Self-supervised learning in the wild

Does the recent progress on academic datasets translate to the real-world setting?

Speech recognition has been the cornerstone benchmarking task. How do we target spoken language understanding directly?

### • Large language models in medical dialogue

LLMs will likely play a central role in the future of medical documentation and communication. How do we get a grip of their uncertainty?

## └ discussion

### └ What lies ahead

2024-03-05

1. We saw how certain features can dominate an uncertainty score. How do we properly control which features we want to use to detect out-of-distribution data?
2. Self-supervised learning has made great strides in academic benchmarks. How does this translate to real-world settings?
3. LLMs will be important for medical documentation and communication going forward. What does uncertainty look like in this context?
4. Will explainability be enough to satisfy regulatory requirements?

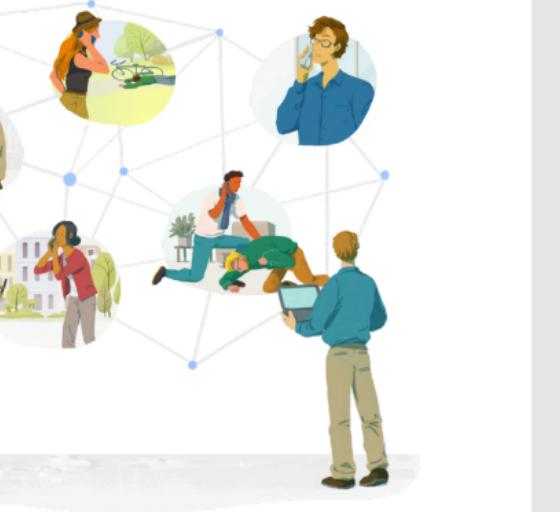
- Selective out-of-distribution detection  
Two pairs of distributions may have identical divergence, but in different dimensions. How do we control features in black-box models?
- Self-supervised learning in the wild  
Does the recent progress on academic datasets translate to the real-world setting?  
Speech recognition has been the cornerstone benchmarking task. How do we target spoken language understanding directly?
- Large language models in medical dialogue  
LLMs will likely play a central role in the future of medical documentation and communication. How do we get a grip of their uncertainty?

## DISCUSSION

# The role of uncertainty in an operational decision support system

- Are true uncertainty estimates really feasible?  
Pragmatism versus idealism.
- Role of explainability compared to uncertainty estimates.
- European Parliamentary Research Services [18]:

*"Future AI solutions for healthcare should be implemented by integrating uncertainty estimation, a relatively new field of research that aims to provide clinicians with clinically useful indications on the degree of confidence in AI predictions"*



2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

### ↳ discussion

#### ↳ The role of uncertainty in an operational decision support system

1. LVMs were difficult to scale to the problems we care about.
2. LVMs did not convincingly outperform more pragmatic approaches (e.g. deferring).
3. Bayesian methods, deferring to predict, discriminative uncertainty.
4. What will happen if uncertainty estimates become a regulatory requirement?

DISCUSSION  
The role of uncertainty in an operational decision support system

- Are true uncertainty estimates really feasible?  
Pragmatism versus idealism.
- Role of explainability compared to uncertainty estimates.
- European Parliamentary Research Services [18]:

*"Future AI solutions for healthcare should be implemented by integrating uncertainty estimation, a relatively new field of research that aims to provide clinicians with clinically useful indications on the degree of confidence in AI predictions"*



Thank you for your attention.



## BIBLIOGRAPHY

### Bibliography I

- [1] Aksan, E., Hilliges, O., "STCN: Stochastic Temporal Convolutional Networks". In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA, 2019 (cited on pages 113, 114).
- [2] Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., Auli, M., *Data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language*. Facebook AI Research blog, 2022 (cited on page 114).
- [3] Baevski, A., Zhou, H., Mohamed, A., Auli, M., "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*. Virtual, 2020. arXiv: 2006.11477 (cited on page 114).
- [4] Berge, E., Whiteley, W., Audebert, H., De Marchis, G. M., Fonseca, A. C., Padiglioni, C., Pérez de la Ossa, N., Strbian, D., Tsivgoulis, G., Turc, G., "European Stroke Organisation (ESO) Guidelines on Intravenous Thrombolysis for Acute Ischaemic Stroke". In: *European Stroke Journal* 6.1 (2021) (cited on page 65).
- [5] Bishop, C. M. "Novelty Detection and Neural-Network Validation". In: *IEE Proceedings - Vision, Image and Signal Processing* 141.4 (1994). ISSN: 1350245x, 13597108 (cited on pages 31, 32).
- [6] Blomberg, S. N., Christensen, H. C., Lippert, F., Ersbøll, A. K., Torp-Petersen, C., Sayre, M. R., Kudenchuk, P. J., Folke, F., "Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest during Calls to Emergency Medical Services: A Randomized Clinical Trial". In: *JAMA Network Open* 4.1 (2021) (cited on page 115).
- [7] Bohm, K., Kurland, L., "The Accuracy of Medical Dispatch - A Systematic Review". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 26 (2018) (cited on page 65).



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

## bibliography

## Bibliography

- BIBLIOGRAPHY  
Bibliography I
- [1] Aksan, E., Hilliges, O., "STCN: Stochastic Temporal Convolutional Networks". In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA, 2019 (cited on pages 113, 114).
  - [2] Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., Auli, M., *Data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language*. Facebook AI Research blog, 2022 (cited on page 114).
  - [3] Baevski, A., Zhou, H., Mohamed, A., Auli, M., "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*. Virtual, 2020. arXiv: 2006.11477 (cited on page 114).
  - [4] Berge, E., Whiteley, W., Audebert, H., De Marchis, G. M., Fonseca, A. C., Padiglioni, C., Pérez de la Ossa, N., Strbian, D., Tsivgoulis, G., Turc, G., "European Stroke Organisation (ESO) Guidelines on Intravenous Thrombolysis for Acute Ischaemic Stroke". In: *European Stroke Journal* 6.1 (2021) (cited on page 65).
  - [5] Bishop, C. M. "Novelty Detection and Neural-Network Validation". In: *IEE Proceedings - Vision, Image and Signal Processing* 141.4 (1994). ISSN: 1350245x, 13597108 (cited on pages 31, 32).
  - [6] Blomberg, S. N., Christensen, H. C., Lippert, F., Ersbøll, A. K., Torp-Petersen, C., Sayre, M. R., Kudenchuk, P. J., Folke, F., "Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest during Calls to Emergency Medical Services: A Randomized Clinical Trial". In: *JAMA Network Open* 4.1 (2021) (cited on page 115).
  - [7] Bohm, K., Kurland, L., "The Accuracy of Medical Dispatch - A Systematic Review". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 26 (2018) (cited on page 65).

# Bibliography II

- [8] Burda, Y., Grosse, R., Salakhutdinov, R. R., "Importance Weighted Autoencoders". In: *Proceedings of the 4th International Conference on Learning Representations (ICLR)*. San Juan, Puerto Rico, 2016 (cited on page 37).
- [9] Buse, A. "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note". In: *The American Statistician* 36 (3a 1982) (cited on pages 36, 104).
- [10] Carver, N., Gupta, V., Hipskind, J. E., "Medical Errors". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024. pmid: 28613514 (cited on pages 7, 8).
- [11] Child, R. "Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images". In: *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. 2021 (cited on page 135).
- [12] Choi, H., Jang, E., Alemi, A. A., WAIC, but Why? Generative Ensembles for Robust Anomaly Detection. 2019. arXiv: 1810.01392 (cited on page 39).
- [13] Chung, Y.-A., Hsu, W.-N., Tang, H., Glass, J., *An Unsupervised Autoregressive Model for Speech Representation Learning*. 2019. arXiv: 1904.03240 (cited on page 114).
- [14] Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., Bengio, Y., "A Recurrent Latent Variable Model for Sequential Data". In: *Proceedings of the 29th Conference on Neural Information Processing Systems (NeurIPS)*. Montréal, Québec, Canada, 2015 (cited on pages 113, 114).
- [15] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. arXiv: 1810.04805. (Visited on 11 February 2019) (cited on page 117).



## bibliography

### Bibliography

2024-03-05

- BIBLIOGRAPHY**
- Bibliography II**
- [8] Burda, Y., Grosse, R., Salakhutdinov, R. R., "Importance Weighted Autoencoders". In: *Proceedings of the 4th International Conference on Learning Representations (ICLR)*. San Juan, Puerto Rico, 2016 (cited on page 27).
  - [9] Buse, A. "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note". In: *The American Statistician* 36 (3a 1982) (cited on pages 36, 104).
  - [10] Carver, N., Gupta, V., Hipskind, J. E., "Medical Errors". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024.
  - [11] Child, R. "Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images". In: *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. 2021 (cited on page 135).
  - [12] Choi, H., Jang, E., Alemi, A. A., WAIC, but Why? Generative Ensembles for Robust Anomaly Detection. 2019. arXiv: 1810.01392 (cited on page 39).
  - [13] Chung, Y.-A., Hsu, W.-N., Tang, H., Glass, J., *An Unsupervised Autoregressive Model for Speech Representation Learning*. 2019. arXiv: 1904.03240 (cited on page 114).
  - [14] Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., Bengio, Y., "A Recurrent Latent Variable Model for Sequential Data". In: *Proceedings of the 29th Conference on Neural Information Processing Systems (NeurIPS)*. Montréal, Québec, Canada, 2015 (cited on pages 113, 114).
  - [15] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. arXiv: 1810.04805. (Visited on 11 February 2019) (cited on page 117).

# Bibliography III

- [16] Dieng, A. B., Kim, Y., Rush, A. M., Blei, D. M., "Avoiding Latent Variable Collapse with Generative Skip Models". In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*. Volume 89. Naha, Okinawa, Japan: PMLR, 2019 (cited on page 130).
- [17] Ebbers, J., Heymann, J., Drude, L., Glarner, T., Haeb-Umbach, R., Raj, B., "Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery". In: *Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2017 (cited on page 113).
- [18] European Parliament, Directorate-General for Parliamentary Research Services, Lekadir, K., Quaglio, G., Tselioudis Garmendia, A., Gallin, C., *Artificial Intelligence in Healthcare – Applications, Risks, and Ethical and Societal Impacts*. European Parliament, 2022 (cited on page 89).
- [19] Fraccaro, M., Sønderby, S. K., Paquet, U., Winther, O., "Sequential Neural Models with Stochastic Layers". In: *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*. Barcelona, Spain, 2016 (cited on pages 113, 114).
- [20] GBD 2019 Stroke Collaborators, "Global, Regional, and National Burden of Stroke and Its Risk Factors, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019". In: *The Lancet Neurology* 20.10 (2021). ISSN: 1474-4422 (cited on page 65).
- [21] Glarner, T., Hanebrink, P., Ebbers, J., Haeb-Umbach, R., "Full Bayesian Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery". In: *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech)*. Hyderabad, India: ISCA, 2018 (cited on page 113).



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

- [bibliography](#)
- [Bibliography](#)

- BIBLIOGRAPHY**  
**Bibliography III**
- [14] Dong, A. B., Kim, Y., Rush, A. M., Blei, D. M., "Avoiding Latent Variable Collapse with Generative Skip Models". In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*. Volume 89. Naha, Okinawa, Japan: PMLR, 2019 (cited on page 130).
- [15] Ebbers, J., Heymann, J., Drude, L., Glarner, T., Haeb-Umbach, R., Raj, B., "Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery". In: *Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2017 (cited on page 113).
- [16] European Parliament, Directorate-General for Parliamentary Research Services, Lekadir, K., Quaglio, G., Tselioudis Garmendia, A., Gallin, C., *Artificial Intelligence in Healthcare – Applications, Risks, and Ethical and Societal Impacts*. European Parliament, 2022 (cited on page 89).
- [17] Fraccaro, M., Sønderby, S. K., Paquet, U., Winther, O., "Sequential Neural Models with Stochastic Layers". In: *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*. Barcelona, Spain, 2016 (cited on pages 113, 114).
- [18] GBD 2019 Stroke Collaborators, "Global, Regional, and National Burden of Stroke and Its Risk Factors, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019". In: *The Lancet Neurology* 20.10 (2021). ISSN: 1474-4422 (cited on page 65).
- [19] Glarner, T., Hanebrink, P., Ebbers, J., Haeb-Umbach, R., "Full Bayesian Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery". In: *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech)*. Hyderabad, India: ISCA, 2018 (cited on page 113).

## BIBLIOGRAPHY

### Bibliography IV

- [22] Hariharan, P., Tariq, M. B., Grotta, J. C., Czap, A. L., "Mobile Stroke Units: Current Evidence and Impact". In: *Current Neurology and Neuroscience Reports* 22.1 (2022) (cited on page 65).
- [23] Hendrycks, D., Mazeika, M., Dietterich, T. G., "Deep Anomaly Detection with Outlier Exposure". In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA, 2019 (cited on page 39).
- [24] Horwitz, L. I., Jenq, G. Y., Brewster, U. C., Chen, C., Kanade, S., Van Ness, P. H., Araujo, K. L. B., Ziaeian, B., Moriarty, J. P., Fogerty, R., Krumholz, H. M., "Comprehensive Quality of Discharge Summaries at an Academic Medical Center". In: *Journal of hospital medicine : an official publication of the Society of Hospital Medicine* 8.8 (2013). ISSN: 1553-5592. pmid: 23526813 (cited on pages 9, 10).
- [25] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units". In: (2021) (cited on page 114).
- [26] Hsu, W.-N., Zhang, Y., Glass, J., *Learning Latent Representations for Speech Generation and Transformation*. 2017. arXiv: 1704.04222 (cited on pages 113, 114).
- [27] Hsu, W.-N., Zhang, Y., Glass, J., "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data". In: *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA, 2017 (cited on pages 113, 114).
- [28] Ioffe, S., Szegedy, C., "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the International Conference on Machine Learning (ICML)*. Lille, France, 2015. arXiv: 1502.03167 (cited on page 130).



## UNCERTAINTY AND THE MEDICAL INTERVIEW

### bibliography

### Bibliography

2024-03-05

- BIBLIOGRAPHY  
Bibliography IV
- [22] Hariharan, P., Tariq, M. B., Grotta, J. C., Czap, A. L., "Mobile Stroke Units: Current Evidence and Impact". In: *Current Neurology and Neuroscience Reports* 22.1 (2022) (cited on page 65).
  - [23] Hendrycks, D., Mazeika, M., Dietterich, T. G., "Deep Anomaly Detection with Outlier Exposure". In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA, 2019 (cited on page 39).
  - [24] Horwitz, L. I., Jenq, G. Y., Brewster, U. C., Chen, C., Kanade, S., Van Ness, P. H., Araujo, K. L. B., Ziaeian, B., Moriarty, J. P., Fogerty, R., Krumholz, H. M., "Comprehensive Quality of Discharge Summaries at an Academic Medical Center". In: *Journal of hospital medicine : an official publication of the Society of Hospital Medicine* 8.8 (2013). ISSN: 1553-5592. pmid: 23526813 (cited on pages 9, 10).
  - [25] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units". In: (2021) (cited on page 114).
  - [26] Hsu, W.-N., Zhang, Y., Glass, J., *Learning Latent Representations for Speech Generation and Transformation*. 2017. arXiv: 1704.04222 (cited on pages 113, 114).
  - [27] Hsu, W.-N., Zhang, Y., Glass, J., "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data". In: *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA, 2017 (cited on pages 113, 114).
  - [28] Ioffe, S., Szegedy, C., "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the International Conference on Machine Learning (ICML)*. Lille, France, 2015. arXiv: 1502.03167 (cited on page 130).

## BIBLIOGRAPHY Bibliography V



- [29] Joukes, E., Abu-Hanna, A., Cornet, R., De Keizer, N., "Time Spent on Dedicated Patient Care and Documentation Tasks Before and After the Introduction of a Structured and Standardized Electronic Health Record". In: *Applied Clinical Informatics* 09.01 (2018). ISSN: 1869-0327 (cited on pages 9, 10).
- [30] Katan, M., Luft, A., "Global Burden of Stroke". In: *Seminars in Neurology*. Volume 38. 02. Thieme Medical Publishers, 2018 (cited on page 65).
- [31] Khurana, S., Joty, S. R., Ali, A., Glass, J., "A Factorial Deep Markov Model for Unsupervised Disentangled Representation Learning from Speech". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, 2019. ISBN: 978-1-4799-8131-1 (cited on pages 113, 114).
- [32] Khurana, S., Laurent, A., Hsu, W.-N., Chorowski, J., Lancucki, A., Marxer, R., Glass, J., *A Convolutional Deep Markov Model for Unsupervised Speech Representation Learning*. 2020. arXiv: 2006.02547 (cited on pages 113, 114).
- [33] Kingma, D. P., Welling, M., "Auto-Encoding Variational Bayes". In: *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. Banff, AB, Canada, 2014. arXiv: 1312.6114 (cited on pages 33, 124).
- [34] Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M., "Improved Variational Inference with Inverse Autoregressive Flow". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS'16)*. Barcelona, Spain, 2016. ISBN: 978-1-5108-3881-9 (cited on page 130).

## UNCERTAINTY AND THE MEDICAL INTERVIEW

2024-03-05

- └ bibliography
- └ Bibliography

- └ BIBLIOGRAPHY
- └ Bibliography V
- [29] Joukes, E., Abu-Hanna, A., Cornet, R., De Keizer, N., "Time Spent on Dedicated Patient Care and Documentation Tasks Before and After the Introduction of a Structured and Standardized Electronic Health Record". In: *Applied Clinical Informatics* 09.01 (2018). ISSN: 1869-0327 (cited on pages 9, 10).
- [30] Katan, M., Luft, A., "Global Burden of Stroke". In: *Seminars in Neurology*. Volume 38. 02. Thieme Medical Publishers, 2018 (cited on page 65).
- [31] Khurana, S., Joty, S. R., Ali, A., Glass, J., "A Factorial Deep Markov Model for Unsupervised Disentangled Representation Learning from Speech". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, 2019. ISBN: 978-1-4799-8131-1 (cited on pages 113, 114).
- [32] Khurana, S., Laurent, A., Hsu, W.-N., Chorowski, J., Lancucki, A., Marxer, R., Glass, J., *A Convolutional Deep Markov Model for Unsupervised Speech Representation Learning*. 2020. arXiv: 2006.02547 (cited on pages 113, 114).
- [33] Kingma, D. P., Welling, M., "Auto-Encoding Variational Bayes". In: *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. Banff, AB, Canada, 2014. arXiv: 1312.6114 (cited on pages 33, 124).
- [34] Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M., "Improved Variational Inference with Inverse Autoregressive Flow". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS'16)*. Barcelona, Spain, 2016. ISBN: 978-1-5108-3881-9 (cited on page 130).

## BIBLIOGRAPHY

### Bibliography VI

- [35] Kyu, H. H., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., "Global, Regional, and National Disability-Adjusted Life-Years (DALYs) for 359 Diseases and Injuries and Healthy Life Expectancy (HALE) for 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017". In: *The Lancet* 392.10159 (2018) (cited on page 65).
- [36] Lee, K., Lee, K., Lee, H., Shin, J., "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks". In: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*. Montréal, Québec, Canada, 2018 (cited on page 39).
- [37] Ling, S., Liu, Y., "DeCoAR 2.0: Deep Contextualized Acoustic Representations with Vector Quantization". 2020. arXiv: 2012.06659 (cited on page 114).
- [38] Liu, A. H., Chung, Y.-A., Glass, J., "Non-Autoregressive Predictive Coding for Learning Speech Representations from Local Dependencies". 2020. arXiv: 2011.00406 (cited on page 114).
- [39] Liu, A. T., Yang, S.-w., Chi, P.-H., Hsu, P.-c., Lee, H.-y., "Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020 (cited on page 114).
- [40] Maaløe, L., Fraccaro, M., Liévin, V., Winther, O., "BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling". In: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada, 2019 (cited on pages 33, 103, 130).



2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

## bibliography

## Bibliography

- BIBLIOGRAPHY  
Bibliography VI
- [35] Kyu, H. H., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., "Global, Regional, and National Disability-Adjusted Life-Years (DALYs) for 359 Diseases and Injuries and Healthy Life Expectancy (HALE) for 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017". In: *The Lancet* 392.10159 (2018) (cited on page 65).
  - [36] Lee, K., Lee, K., Lee, H., Shin, J., "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks". In: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*. Montréal, Québec, Canada, 2018 (cited on page 39).
  - [37] Ling, S., Liu, Y., "DeCoAR 2.0: Deep Contextualized Acoustic Representations with Vector Quantization". 2020. arXiv: 2012.06659 (cited on page 114).
  - [38] Liu, A. H., Chung, Y.-A., Glass, J., "Non-Autoregressive Predictive Coding for Learning Speech Representations from Local Dependencies". 2020. arXiv: 2011.00406 (cited on page 114).
  - [39] Liu, A. T., Yang, S.-w., Chi, P.-H., Hsu, P.-c., Lee, H.-y., "Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020 (cited on page 114).
  - [40] Maaløe, L., Fraccaro, M., Liévin, V., Winther, O., "BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling". In: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada, 2019 (cited on pages 33, 103, 130).

## BIBLIOGRAPHY

### Bibliography VII

- [41] Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., Lakshminarayanan, B., "Do Deep Generative Models Know What They Don't Know?" In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA, 2019. arXiv: 1810.09136 (cited on page 40).
- [42] Navi, B. B., Audebert, H. J., Alexandrov, A. W., Cadilhac, D. A., Grotta, J. C., PRESTO (Prehospital Stroke Treatment Organization) Writing Group, "Mobile Stroke Units: Evidence, Gaps, and next Steps". In: *Stroke* 53.6 (2022) (cited on page 65).
- [43] Oord, A., Li, Y., Vinyals, O., *Representation Learning with Contrastive Predictive Coding*. 2018. arXiv: 1807.03748 (cited on page 114).
- [44] Oord, A., Vinyals, O., Kavukcuoglu, K., "Neural Discrete Representation Learning". In: *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA, 2018 (cited on pages 113, 114).
- [45] Oostema, J. A., Carle, T., Talia, N., Reeves, M., "Dispatcher Stroke Recognition Using a Stroke Screening Tool: A Systematic Review". In: *Cerebrovascular Diseases* 42.5-6 (2016) (cited on page 65).
- [46] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., "Improving Language Understanding by Generative Pre-Training". In: (2018) (cited on page 117).
- [47] Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., Lakshminarayanan, B., "Likelihood Ratios for Out-of-Distribution Detection". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada, 2019 (cited on page 39).



2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

- ### bibliography

  - ### Bibliography

- BIBLIOGRAPHY
- Bibliography VII
- [41] Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., Lakshminarayanan, B., "Do Deep Generative Models Know What They Don't Know?" In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA, 2019. arXiv: 1810.09136 (cited on page 40).
- [42] Navi, B. B., Audebert, H. J., Alexandrov, A. W., Cadilhac, D. A., Grotta, J. C., PRESTO (Prehospital Stroke Treatment Organization) Writing Group, "Mobile Stroke Units: Evidence, Gaps, and next Steps". In: *Stroke* 53.6 (2022) (cited on page 65).
- [43] Oord, A., Li, Y., Vinyals, O., *Representation Learning with Contrastive Predictive Coding*. 2018. arXiv: 1807.03748 (cited on page 114).
- [44] Oord, A., Vinyals, O., Kavukcuoglu, K., "Neural Discrete Representation Learning". In: *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA, 2018 (cited on pages 113, 114).
- [45] Oostema, J. A., Carle, T., Talia, N., Reeves, M., "Dispatcher Stroke Recognition Using a Stroke Screening Tool: A Systematic Review". In: *Cerebrovascular Diseases* 42.5-6 (2016) (cited on page 65).
- [46] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., "Improving Language Understanding by Generative Pre-Training". In: (2018) (cited on page 117).
- [47] Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., Lakshminarayanan, B., "Likelihood Ratios for Out-of-Distribution Detection". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada, 2019 (cited on page 39).

## BIBLIOGRAPHY

### Bibliography VIII

- [48] Rezende, D. J., Mohamed, S., Wierstra, D., "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *Proceedings of the 31st International Conference on Machine Learning (ICML)*. Volume 32. Beijing, China: PMLR, 2014 (cited on page 33).
- [49] Salimans, T., Kingma, D. P., "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks". In: *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*. Barcelona, Spain, 2016 (cited on page 130).
- [50] Schneider, S., Baevski, A., Collobert, R., Auli, M., "Wav2vec: Unsupervised Pre-training for Speech Recognition". In: *Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech)*. Graz, Austria: ISCA, 2019. arXiv: 1904.05862 (cited on page 114).
- [51] Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Núñez, J. F., Luque, J., "Input Complexity and Out-of-Distribution Detection with Likelihood-Based Generative Models". In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. Addis Ababa, Ethiopia, 2020 (cited on page 39).
- [52] Sinsky, C., Colligan, L., Li, L., Prgomet, M., Reynolds, S., Goeders, L., Westbrook, J., Tutty, M., Blike, G., "Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties". In: *Annals of Internal Medicine* 165.11 (2016). ISSN: 1539-3704. pmid: 27595430 (cited on pages 9, 10).
- [53] Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., Winther, O., "Ladder Variational Autoencoders". In: *Proceedings of the 29th Conference on Neural Information Processing Systems (NeurIPS)*. Barcelona, Spain, 2016 (cited on page 130).



2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

- └ bibliography
  - └ Bibliography

### BIBLIOGRAPHY

#### Bibliography VIII

- [48] Rezende, D. J., Mohamed, S., Wierstra, D., "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *Proceedings of the 31st International Conference on Machine Learning (ICML)*. Volume 32. Beijing, China: PMLR, 2014 (cited on page 33).
- [49] Salimans, T., Kingma, D. P., "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks". In: *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*. Barcelona, Spain, 2016 (cited on page 130).
- [50] Schneider, S., Baevski, A., Collobert, R., Auli, M., "Wav2vec: Unsupervised Pre-training for Speech Recognition". In: *Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech)*. Graz, Austria: ISCA, 2019. arXiv: 1904.05862 (cited on page 114).
- [51] Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Núñez, J. F., Luque, J., "Input Complexity and Out-of-Distribution Detection with Likelihood-Based Generative Models". In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. Addis Ababa, Ethiopia, 2020 (cited on page 39).
- [52] Sinsky, C., Colligan, L., Li, L., Prgomet, M., Reynolds, S., Goeders, L., Westbrook, J., Tutty, M., Blike, G., "Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties". In: *Annals of Internal Medicine* 165.11 (2016). ISSN: 1539-3704. pmid: 27595430 (cited on pages 9, 10).
- [53] Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., Winther, O., "Ladder Variational Autoencoders". In: *Proceedings of the 29th Conference on Neural Information Processing Systems (NeurIPS)*. Barcelona, Spain, 2016 (cited on page 130).

# Bibliography IX



- [54] Starmer, A. J., Spector, N. D., Srivastava, R., West, D. C., Rosenbluth, G., Allen, A. D., Noble, E. L., Tse, L. L., Dalal, A. K., Keohane, C. A., "Changes in Medical Errors after Implementation of a Handoff Program". In: *New England Journal of Medicine* 371.19 (2014) (cited on pages 7, 8).
- [55] Tipping, M. D., Forth, V. E., O'Leary, K. J., Malkenson, D. M., Magill, D. B., Englert, K., Williams, M. V., "Where Did the Day Go?—A Time-Motion Study of Hospitalists". In: *Journal of Hospital Medicine* 5.6 (2010). issn: 1553-5606. pmid: 20803669 (cited on pages 9, 10).
- [56] Turc, G., Bhogal, P., Fischer, U., Khatri, P., Lobotesis, K., Mazighi, M., Schellinger, P. D., Toni, D., De Vries, J., White, P., "European Stroke Organisation (ESO)-European Society for Minimally Invasive Neurological Therapy (ESMINT) Guidelines on Mechanical Thrombectomy in Acute Ischemic Stroke". In: *Journal of Neurointerventional Surgery* 11.8 (2019) (cited on page 65).
- [57] Verma, R., Nalisnick, E., "Calibrated Learning to Defer with One-vs-All Classifiers". In: *International Conference on Machine Learning*. PMLR, 2022 (cited on page 82).
- [58] Viereck, S., Møller, T. P., Iversen, H. K., Christensen, H., Lippert, F., "Medical Dispatchers Recognise Substantial Amount of Acute Stroke during Emergency Calls". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 24 (2016) (cited on page 65).
- [59] Xiao, Z., Yan, Q., Amit, Y., "Likelihood Regret: An Out-of-Distribution Detection Score for Variational Auto-Encoder". In: *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*. Virtual, 2020 (cited on page 39).

## bibliography

### Bibliography

2024-03-05

- [54] Starmer, A. J., Spector, N. D., Srivastava, R., West, D. C., Rosenbluth, G., Allen, A. D., Noble, E. L., Tse, L. L., Dalal, A. K., Keohane, C. A., "Changes in Medical Errors after Implementation of a Handoff Program". In: *New England Journal of Medicine* 371.19 (2014) (cited on pages 7, 8).
- [55] Tipping, M. D., Forth, V. E., O'Leary, K. J., Malkenson, D. M., Magill, D. B., Englert, K., Williams, M. V., "Where Did the Day Go?—A Time-Motion Study of Hospitalists". In: *Journal of Hospital Medicine* 5.6 (2010). issn: 1553-5606. pmid: 20803669 (cited on pages 9, 10).
- [56] Turc, G., Bhogal, P., Fischer, U., Khatri, P., Lobotesis, K., Mazighi, M., Schellinger, P. D., Toni, D., De Vries, J., White, P., "European Stroke Organisation (ESO)-European Society for Minimally Invasive Neurological Therapy (ESMINT) Guidelines on Mechanical Thrombectomy in Acute Ischemic Stroke". In: *Journal of Neurointerventional Surgery* 11.8 (2019) (cited on page 65).
- [57] Verma, R., Nalisnick, E., "Calibrated Learning to Defer with One-vs-All Classifiers". In: *International Conference on Machine Learning*. PMLR, 2022 (cited on page 82).
- [58] Viereck, S., Møller, T. P., Iversen, H. K., Christensen, H., Lippert, F., "Medical Dispatchers Recognise Substantial Amount of Acute Stroke during Emergency Calls". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 24 (2016) (cited on page 65).
- [59] Xiao, Z., Yan, Q., Amit, Y., "Likelihood Regret: An Out-of-Distribution Detection Score for Variational Auto-Encoder". In: *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*. Virtual, 2020 (cited on page 39).

## Extra slides

• Introduction  
• Hierarchical VAEs Know What They Don't Know  
• A Brief Overview of Unsupervised Speech Representation Learning  
• VAE background and the Ladder VAE



- Introduction
- Hierarchical VAEs Know What They Don't Know
- A Brief Overview of Unsupervised Speech Representation Learning
- VAE background and the Ladder VAE

# Reliability of machine learning systems



- **Data:** Quality, quantity, diversity, bias, privacy, ethics.
- **Task:** Context, domain, language, culture, purpose.

└ introduction

└ Reliability of machine learning systems

1. So what are the challenges holding us back in implementing such systems?
  2. – Strong requirements of data.
    - Tasks that span different contexts, domains, languages, and cultures.
    - Regulatory requirements for transparency and accountability.
    - Trust and understanding of predictions.

- Data: Quality, quantity, diversity, bias, privacy, ethics.
- Task: Context, domain, language, culture, purpose.

## Reliability of machine learning systems



- **Data:** Quality, quantity, diversity, bias, privacy, ethics.
- **Task:** Context, domain, language, culture, purpose.
- **Interpretability** of how a model works (transparency, accountability, regulation).
- **Explainability** of model predictions (trust, understanding, feedback).
- **Fairness** in treatment of different groups of people.
- **Robustness** to noise, outliers, distribution shift, and adversarial attacks.

└ introduction

└ Reliability of machine learning systems

2024-03-05

- **Data:** Quality, quantity, diversity, bias, privacy, ethics.
- **Task:** Context, domain, language, culture, purpose.
- **Interpretability** of how a model works (transparency, accountability, regulation).
- **Explainability** of model predictions (trust, understanding, feedback).
- **Fairness** in treatment of different groups of people.
- **Robustness** to noise, outliers, distribution shift, and adversarial attacks.

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW  
An alternative likelihood bound,  $\mathcal{L}^{>k}$



An alternative version of the ELBO that only partially uses the approximate posterior can be written as [40]

$$\mathcal{L}^{>k}(x; \theta, \phi) = \mathbb{E}_{p_{\theta}(z_{\leq k}|z_{>k})q_{\phi}(z_{>k}|x)} \left[ \log \frac{p_{\theta}(x|z)p_{\theta}(z_{>k})}{q_{\phi}(z_{>k}|x)} \right] \quad (8)$$

Here, we have replaced the approximate posterior  $q_{\phi}(z|x)$  with a different proposal distribution that combines part of the approximate posterior with the conditional prior, namely

$$p_{\theta}(z_{\leq k}|z_{>k})q_{\phi}(z_{>k}|x)$$

This bound uses the conditional prior for the lowest latent variables in the hierarchy.

UNCERTAINTY AND THE MEDICAL INTERVIEW

- 2024-03-05  
└ hierarchical vaes know what they don't know  
└ An alternative likelihood bound,  $\mathcal{L}^{>k}$

An alternative version of the ELBO that only partially uses the approximate posterior can be written as [40]

$$\mathcal{L}^{>k}(x; \theta, \phi) = \mathbb{E}_{p_{\theta}(z_{\leq k}|z_{>k})q_{\phi}(z_{>k}|x)} \left[ \log \frac{p_{\theta}(x|z)p_{\theta}(z_{>k})}{q_{\phi}(z_{>k}|x)} \right] \quad (8)$$

Here, we have replaced the approximate posterior  $q_{\phi}(z|x)$  with a different proposal distribution that combines part of the approximate posterior with the conditional prior, namely

$$p_{\theta}(z_{\leq k}|z_{>k})q_{\phi}(z_{>k}|x)$$

This bound uses the conditional prior for the lowest latent variables in the hierarchy.

1. So can we come up with a new bound that does not use the lowest latent variables in the same way?
2. So we could use this for OOD detection (as done in BIVA).

## Likelihood ratios

We can use our new bound to compute the score used in a standard likelihood ratio test [9].

$$\text{LLR}^{>k}(x) \equiv \mathcal{L}(x) - \mathcal{L}^{>k}(x). \quad (9)$$

We can inspect what this likelihood-ratio measures by considering the exact form of our bounds.

$$\mathcal{L} = \log p_\theta(x) - D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x)), \quad (10)$$

$$\mathcal{L}^{>k} = \log p_\theta(x) - D_{\text{KL}}(p_\theta(z_{\leq k}|z_{>k})q_\phi(z_{>k}|x) \| p_\theta(z|x)).$$

In the likelihood ratio the reconstruction terms cancel out and only the KL-divergences from the approximate to the true posterior remain.

$$\begin{aligned} \text{LLR}^{>k}(x) &= -D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x)) \\ &\quad + D_{\text{KL}}(p_\theta(z_{\leq k}|z_{>k})q_\phi(z_{>k}|x) \| p_\theta(z|x)). \end{aligned} \quad (11)$$

# UNCERTAINTY AND THE MEDICAL INTERVIEW

hierarchical vaes know what they don't know

Likelihood ratios

2024-03-05

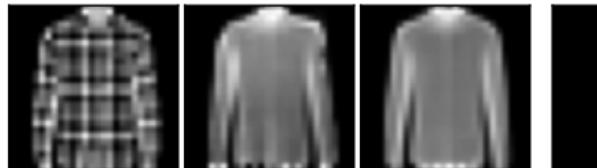
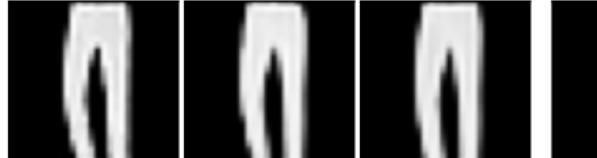
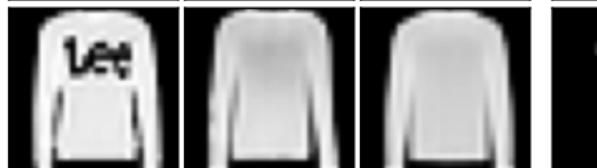
HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW  
Likelihood ratios  
We can use our new bound to compute the score used in a standard likelihood ratio test [9].  
 $\text{LLR}^{>k}(x) \equiv \mathcal{L}(x) - \mathcal{L}^{>k}(x).$  (9)  
 We can inspect what this likelihood-ratio measures by considering the exact form of our bounds.  
 $\mathcal{L} = \log p_\theta(x) - D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x)),$   
 $\mathcal{L}^{>k} = \log p_\theta(x) - D_{\text{KL}}(p_\theta(z_{\leq k}|z_{>k})q_\phi(z_{>k}|x) \| p_\theta(z|x)).$   
 In the likelihood ratio the reconstruction terms cancel out and only the KL-divergences from the approximate to the true posterior remain.  
 $\text{LLR}^{>k}(x) = -D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x))$   
 $+ D_{\text{KL}}(p_\theta(z_{\leq k}|z_{>k})q_\phi(z_{>k}|x) \| p_\theta(z|x)).$  (11)

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

## Reconstructions of ID and OOD data

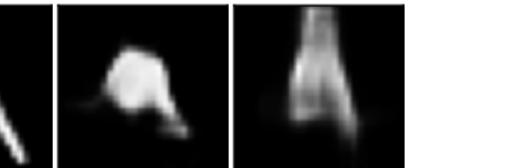
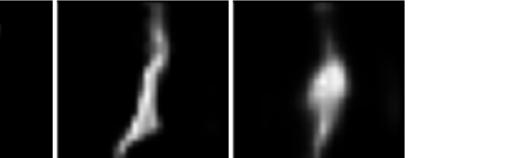
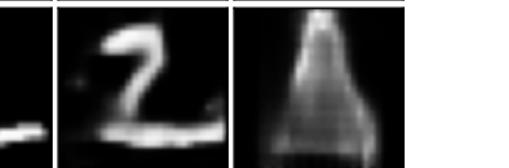
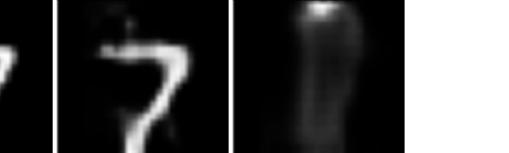
In-distribution

Example Reconstruction Latent recon.



Out-of-distribution

Example Reconstruction Latent recon.

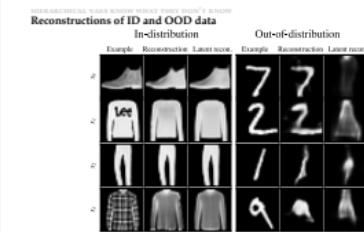


2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

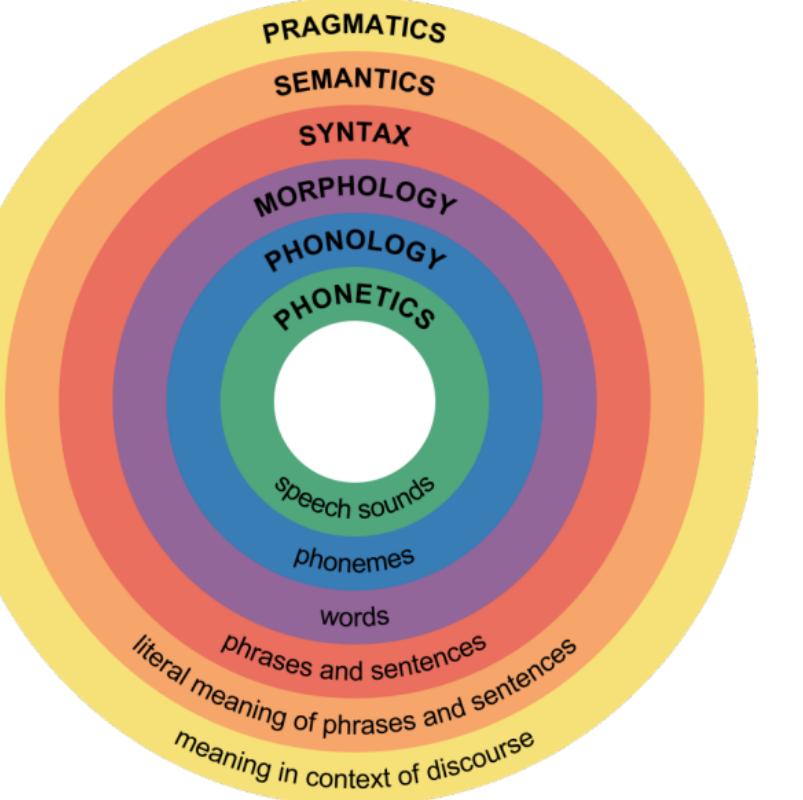
hierarchical vae's know what they don't know

Reconstructions of ID and OOD data



HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

## Hierarchy of speech features



## UNCERTAINTY AND THE MEDICAL INTERVIEW

hierarchical vaes know what they don't know

Hierarchy of speech features

2024-03-05

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW  
Hierarchy of speech features



# Results on diverse datasets

OOD dataset	Metric	AUROC↑	AUPRC↑	FPR80↓
<b>Trained on CIFAR10</b>				
SVHN	LLR <sup>&gt;2</sup>	0.811	0.837	0.394
CIFAR10	LLR <sup>&gt;1</sup>	0.469	0.479	0.835
<b>Trained on SVHN</b>				
CIFAR10	LLR <sup>&gt;1</sup>	0.939	0.950	0.052
SVHN	LLR <sup>&gt;1</sup>	0.489	0.484	0.799



OOD dataset	Metric	AUROC↑	AUPRC↑	FPR80↓
<b>Trained on FashionMNIST</b>				
MNIST	LLR <sup>&gt;1</sup>	0.986	0.987	0.011
notMNIST	LLR <sup>&gt;1</sup>	0.998	0.998	0.000
KMNIST	LLR <sup>&gt;1</sup>	0.974	0.977	0.017
Omniglot28x28	LLR <sup>&gt;2</sup>	1.000	1.000	0.000
Omniglot28x28Inverted	LLR <sup>&gt;1</sup>	0.954	0.954	0.050
SmallNORB28x28	LLR <sup>&gt;2</sup>	0.999	0.999	0.002
SmallNORB28x28Inverted	LLR <sup>&gt;2</sup>	0.941	0.946	0.069
FashionMNIST	LLR <sup>&gt;1</sup>	0.488	0.496	0.811
<b>Trained on MNIST</b>				
FashionMNIST	LLR <sup>&gt;1</sup>	0.999	0.999	0.000
notMNIST	LLR <sup>&gt;1</sup>	1.000	0.999	0.000
KMNIST	LLR <sup>&gt;1</sup>	0.999	0.999	0.000
Omniglot28x28	LLR <sup>&gt;1</sup>	1.000	1.000	0.000
Omniglot28x28Inverted	LLR <sup>&gt;1</sup>	0.944	0.953	0.057
SmallNORB28x28	LLR <sup>&gt;1</sup>	1.000	1.000	0.000
SmallNORB28x28Inverted	LLR <sup>&gt;1</sup>	0.985	0.987	0.000
MNIST	LLR <sup>&gt;2</sup>	0.515	0.507	0.792

# UNCERTAINTY AND THE MEDICAL INTERVIEW

↳ hierarchical vae's know what they don't know

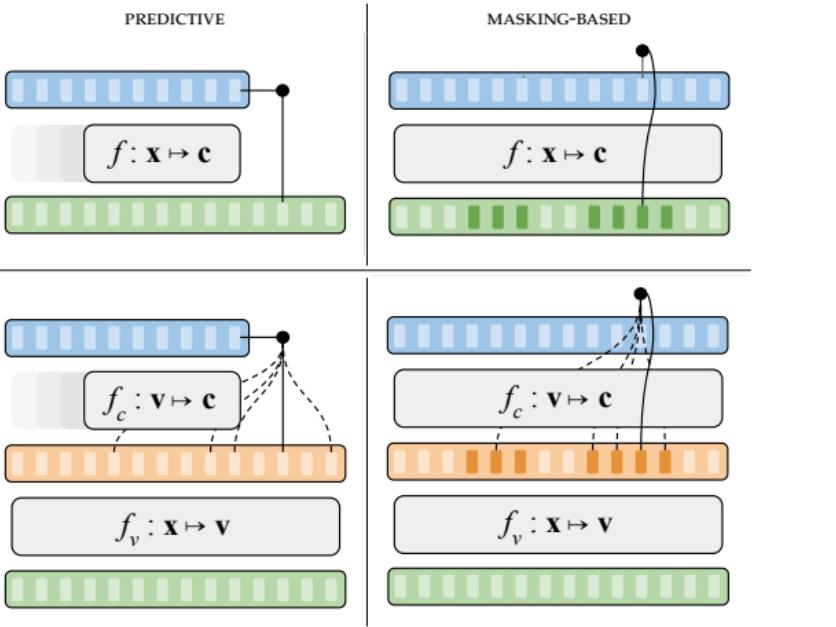
↳ Results on diverse datasets

2024-03-05

OOD dataset	Metric	AUROC↑	AUPRC↑	FPR80↓
<b>Trained on FashionMNIST</b>				
MNIST	LLR <sup>&gt;1</sup>	0.998	0.997	0.012
notMNIST	LLR <sup>&gt;1</sup>	0.998	0.998	0.000
KMNIST	LLR <sup>&gt;1</sup>	1.000	1.000	0.007
Omniglot28x28	LLR <sup>&gt;1</sup>	1.000	1.000	0.000
Omniglot28x28Inverted	LLR <sup>&gt;1</sup>	0.994	0.994	0.000
SmallNORB28x28	LLR <sup>&gt;1</sup>	1.000	1.000	0.000
SmallNORB28x28Inverted	LLR <sup>&gt;1</sup>	0.994	0.994	0.000
SVHN	LLR <sup>&gt;1</sup>	0.994	0.994	0.000
<b>Trained on SVHN</b>				
CIFAR10	LLR <sup>&gt;1</sup>	0.990	0.992	0.012
SVHN	LLR <sup>&gt;1</sup>	0.489	0.494	0.799
<b>Trained on MNIST</b>				
FashionMNIST	LLR <sup>&gt;1</sup>	0.999	0.999	0.000
notMNIST	LLR <sup>&gt;1</sup>	1.000	0.999	0.000
KMNIST	LLR <sup>&gt;1</sup>	0.999	0.999	0.000
Omniglot28x28	LLR <sup>&gt;1</sup>	1.000	1.000	0.000
Omniglot28x28Inverted	LLR <sup>&gt;1</sup>	0.994	0.993	0.007
SmallNORB28x28	LLR <sup>&gt;1</sup>	1.000	1.000	0.000
SmallNORB28x28Inverted	LLR <sup>&gt;1</sup>	0.994	0.994	0.000
MNIST	LLR <sup>&gt;2</sup>	0.918	0.927	0.792

## Types of self-supervised speech representation learning methods

Schematic of self-supervised methods. Each subfigure illustrates the loss computation for a single time-step. The temporal subscript has been left out for simplicity.

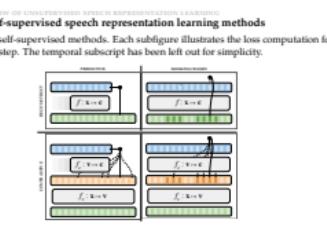


# UNCERTAINTY AND THE MEDICAL INTERVIEW

└ a brief overview of unsupervised speech representation learning

└ Types of self-supervised speech representation learning methods

2024-03-05



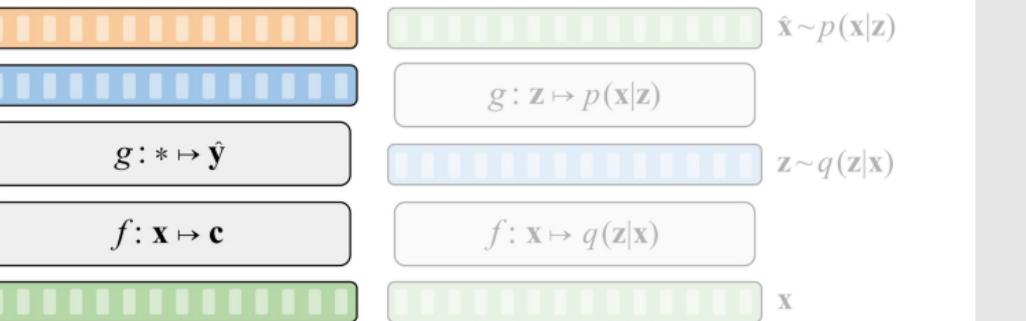
A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING  
Types of self-supervised speech representation learning methods  
Schematic of self-supervised methods. Each subfigure illustrates the loss computation for a single time-step. The temporal subscript has been left out for simplicity.

## A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

### Overview: Representation Learning for Speech



- We focus on two primary categories:
  - Self-supervised learning (SSL)
  - Probabilistic latent variable models (LVMs)
- Recent developments have been driven by self-supervised learning.
- A model-by-model overview: Focus on speech recognition.

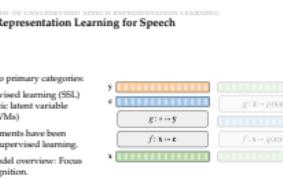


## UNCERTAINTY AND THE MEDICAL INTERVIEW

### a brief overview of unsupervised speech representation learning

### Overview: Representation Learning for Speech

2024-03-05



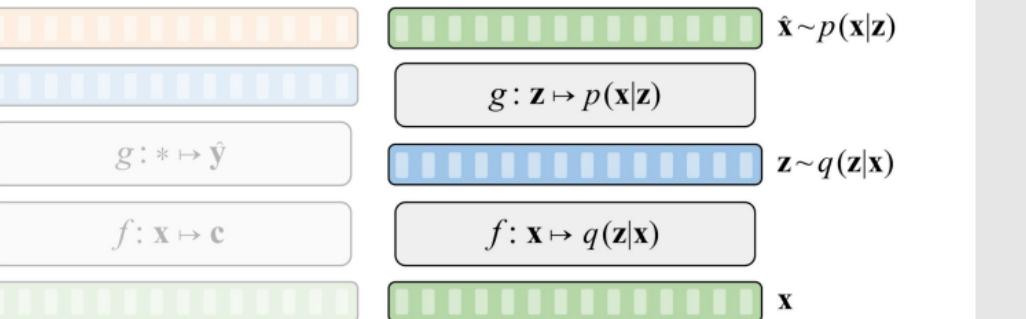
- We focus on two primary categories:
  - Self-supervised learning (SSL)
  - Probabilistic latent variable models (LVMs)
- Recent developments have been driven by self-supervised learning.
- A model-by-model overview: Focus on speech recognition.

## A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

### Overview: Representation Learning for Speech



- We focus on two primary categories:
  - Self-supervised learning (SSL)
  - Probabilistic latent variable models (LVMs)
- Recent developments have been driven by self-supervised learning.
- A model-by-model overview: Focus on speech recognition.



## UNCERTAINTY AND THE MEDICAL INTERVIEW

### a brief overview of unsupervised speech representation learning

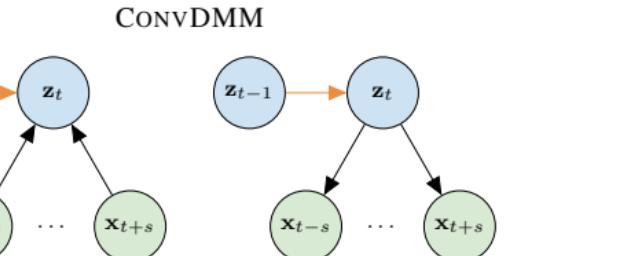
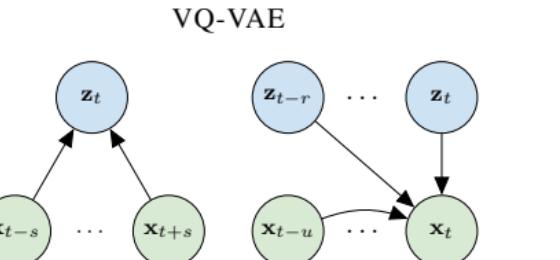
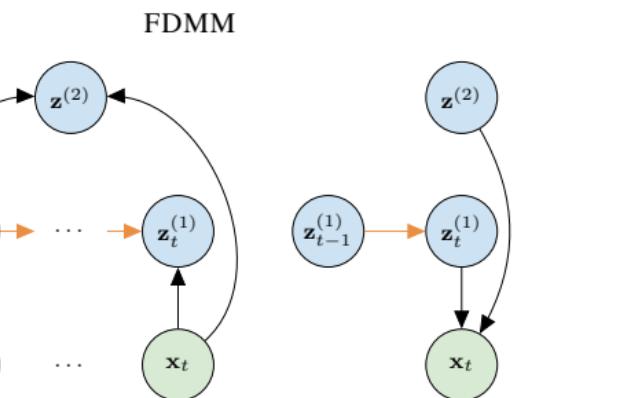
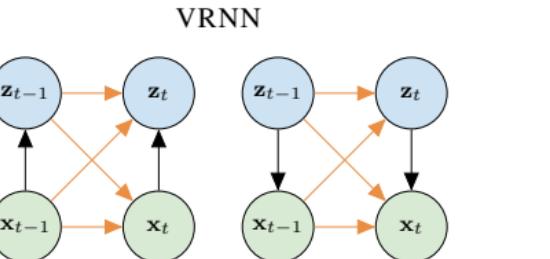
### Overview: Representation Learning for Speech

2024-03-05

- We focus on two primary categories:
  - Self-supervised learning (SSL)
  - Probabilistic latent variable models (LVMs)
- Recent developments have been driven by self-supervised learning.
- A model-by-model overview: Focus on speech recognition.



A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING  
**Graphical models for LVMs**



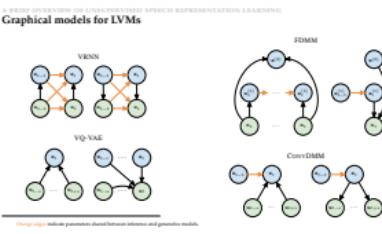
Orange edges indicate parameters shared between inference and generative models.

UNCERTAINTY AND THE MEDICAL INTERVIEW

└ a brief overview of unsupervised speech representation learning

└ Graphical models for LVMs

2024-03-05



# A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

## Overview of LVM probabilistic components



TYPE	FORM
OBSERVATION MODEL	
<b>ARX</b>	Autoregressive on $x_t$ $p(x_t x_{1:t-1})$
<b>LOC</b>	Local latent variable $p(x_t z_{1:t})$
<b>GLB</b>	Global latent variable $p(x_t z)$
PRIOR	
<b>ARX</b>	Autoregressive on $x_t$ $p(z_t x_{1:t-1})$
<b>ARZ</b>	Autoregressive on $z_t$ $p(z_t z_{1:t-1})$
<b>IND</b>	Locally independent $z_t$ $p(z_t)$
<b>GLB</b>	Global latent variable $p(z)$
INFERENCE MODEL	
<b>ARZ</b>	Autoregressive on $z_t$ $q(z_t z_{1:t-1})$
<b>FLT</b>	Filtering $q(z_t x_{1:t})$
<b>LSM</b>	Local smoothing $q(z_t x_{t-r:t+r})$
<b>GSM</b>	Global smoothing $q(z_t x_{1:T})$
<b>GLB</b>	Global latent variable $q(z x_{1:T})$

# UNCERTAINTY AND THE MEDICAL INTERVIEW

- └ a brief overview of unsupervised speech representation learning
- └ Overview of LVM probabilistic components

Type	Form
Observation model	
<b>ARX</b>	Autoregressive on $x_t$ $p(x_t x_{1:t-1})$
<b>LOC</b>	Local latent variable $p(x_t z_{1:t})$
<b>GLB</b>	Global latent variable $p(x_t z)$
Prior	
<b>ARX</b>	Autoregressive on $x_t$ $p(z_t x_{1:t-1})$
<b>ARZ</b>	Autoregressive on $z_t$ $p(z_t z_{1:t-1})$
<b>IND</b>	Locally independent $z_t$ $p(z_t)$
<b>GLB</b>	Global latent variable $p(z)$
Inference model	
<b>ARZ</b>	Autoregressive on $z_t$ $q(z_t z_{1:t-1})$
<b>FLT</b>	Filtering $q(z_t x_{1:t})$
<b>LSM</b>	Local smoothing $q(z_t x_{t-r:t+r})$
<b>GSM</b>	Global smoothing $q(z_t x_{1:T})$
<b>GLB</b>	Global latent variable $q(z x_{1:T})$

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING  
**Classification of selected LVMs for speech**



MODEL	OBSERVATION			PRIOR				INFERENCE					
	ARX	LOC	GLB	ARX	ARZ	IND	GLB	ARZ	FLT	LSM	GSM	GLB	HIE
VRNN [14]	✓	✓	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗
SRNN [19]	✓	✓	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗
HMM-VAE [17]	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✓
ConvVAE [26]	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗
FHVAE [27]	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓	✓	✓
VQ-VAE [44]	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
BHMM-VAE [21]	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗
STCN [1]	✗	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
FDMM [31]	✗	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✓	✓
ConvDMM [32]	✗	✓	✗	✗	✓	✗	✗	✓	✗	✓	✗	✗	✗

# UNCERTAINTY AND THE MEDICAL INTERVIEW

- └ a brief overview of unsupervised speech representation learning
- └ Classification of selected LVMs for speech

2024-03-05

MODEL	OBSERVATION				PRIOR				INFERENCE				
	ARX	LOC	GLB	ARX IND GLB	ARZ	IND	GLB	ARZ FLT LSM GSM GLB	ARZ	FLT	LSM	GSM	GLB
VRNN [14]	✓	✓	✗	✓	✓	✗	✗	✓	✓	✓	✗	✗	✗
SRNN [19]	✓	✓	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗
HMM-VAE [17]	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✓
ConvVAE [26]	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗
FHVAE [27]	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓	✓	✓
VQ-VAE [44]	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
BHMM-VAE [21]	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗
STCN [1]	✗	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
FDMM [31]	✗	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✓	✓
ConvDMM [32]	✗	✓	✗	✗	✓	✗	✗	✓	✗	✓	✗	✗	✗

A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING  
**Comparison of LVMs and SSL methods**



MODEL	MODEL AND TASK DESIGN					RESOLUTION			USAGE		
	MSK	PRD	CON	REC	QTZ	GEN	LOC	GLB	VAR	FRZ	FTN
<b>SELF-SUPERVISED MODELS</b>											
CPC [43]	✗	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗
APC [13]	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗
wav2vec [50]	✗	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗
Mockingjay [39]	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✓
wav2vec 2.0 [3]	✓	✗	✓	✗	✓	✗	✓	✗	✗	✗	✓
NPC [38]	✓	✗	✗	✓	✓	✗	✓	✗	✗	✓	✗
DeCoAR 2.0 [37]	✓	✗	✗	✓	✓	✗	✓	✗	✗	✓	✗
HuBERT [25]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✓
data2vec [2]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓
<b>LATENT VARIABLE MODELS</b>											
VRNN [14]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
SRNN [19]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
ConvVAE [26]	✗	✗	✗	✓	✗	✓	✗	✓	✗	✓	✗
FHVAE [27]	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✗
VQ-VAE [44]	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗
STCN [1]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
FDMM [31]	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✗
ConvDMM [32]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗

# UNCERTAINTY AND THE MEDICAL INTERVIEW

- a brief overview of unsupervised speech representation learning
- Comparison of LVMs and SSL methods

2024-03-05

Method	Model and Task Design			Resolution	Usage
	MSK	PRD	CON		
CPC [43]	✓	✓	✓	✓	✓
APC [13]	✓	✓	✗	✓	✓
wav2vec [50]	✓	✓	✓	✓	✓
Mockingjay [39]	✓	✗	✗	✓	✓
Mockingjay 2.0 [39]	✓	✗	✗	✓	✓
data2vec [2]	✓	✗	✗	✓	✓
VRNN [14]	✗	✗	✗	✓	✓
SRNN [19]	✗	✗	✗	✓	✓
ConvVAE [26]	✗	✗	✗	✓	✓
FHVAE [27]	✗	✗	✗	✓	✓
VQ-VAE [44]	✗	✗	✗	✓	✓
STCN [1]	✗	✗	✗	✓	✓
FDMM [31]	✗	✗	✗	✓	✓
ConvDMM [32]	✗	✗	✗	✓	✓

## Simulated prospective study

I. When is the model prediction presented to the call-taker?

1. Notify the call-taker after the call ends.
2. Notify the call-taker during the call.

II. How does prediction influence the diagnostic code the call-taker assigns to the call?

- A. Call-takers mirror model positives.
- B. Call-takers mirror model negatives.
- C. Call-takers mirror model predictions (corresponds to main results of the model itself).

To simulate the online scenario (2.), we stream the transcript to the model and make predictions every 50 words. A stroke positive is triggered only when three consecutive positive predictions are made. This is similar to the strategy implemented for a previous RCT on cardiac arrest [6].

## UNCERTAINTY AND THE MEDICAL INTERVIEW

a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

Simulated prospective study

2024-03-05

## Simulated prospective study



Predictor	Call-taker	Model		Call-taker supported by the model (simulated)				
When	During call	After call	During call	During call	After call	During call	After call	During call
Method	-	-	-	neg → pos	neg → pos	pos → neg	pos → neg	
F1-score [%] ↑	25.8 (23.7-27.9)	35.7 (35.0-36.4)	33.1 (32.4-33.7)	28.9 (28.3-29.5)	27.6 (27.0-28.1)	33.3 (32.5-34.1)	32.7 (31.8-33.5)	
Sensitivity [%] ↑	52.7 (49.2-56.4)	63.0 (62.0-64.1)	58.7 (57.7-59.8)	72.4 (71.5-73.3)	72.3 (71.4-73.3)	43.4 (42.3-44.5)	39.1 (38.1-40.1)	
PPV [%] ↑	17.1 (15.5-18.6)	24.9 (24.3-25.5)	23.0 (22.5-23.6)	18.0 (17.6-18.4)	17.0 (16.7-17.4)	27.0 (26.3-27.8)	28.1 (27.3-28.9)	
FOR [%] ↓ (1 - NPV)	0.105 (0.094-0.116)	0.082 (0.079-0.085)	0.091 (0.088-0.094)	0.061 (0.059-0.064)	0.061 (0.059-0.064)	0.125 (0.121-0.129)	0.134 (0.131-0.138)	
FPR [%] ↓ (1 - specificity)	0.565 (0.539-0.590)	0.419 (0.413-0.426)	0.432 (0.426-0.439)	0.726 (0.717-0.735)	0.776 (0.767-0.786)	0.258 (0.253-0.263)	0.221 (0.216-0.226)	

## UNCERTAINTY AND THE MEDICAL INTERVIEW

a retrospective study on machine learning-assisted stroke recognition for medical helpline calls

Simulated prospective study

2024-03-05

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS								
Simulated prospective study								
Predictor	Call-taker	Model	Call-taker supported by the model (simulated)					
When	During call	After call	During call	During call	After call	During call	After call	During call
Method	-	-	-	neg → pos	neg → pos	pos → neg	pos → neg	
F1-score [%] ↑	25.8 (23.7-27.9)	35.7 (35.0-36.4)	33.1 (32.4-33.7)	28.9 (28.3-29.5)	27.6 (27.0-28.1)	33.3 (32.5-34.1)	32.7 (31.8-33.5)	
Sensitivity [%] ↑	52.7 (49.2-56.4)	63.0 (62.0-64.1)	58.7 (57.7-59.8)	72.4 (71.5-73.3)	72.3 (71.4-73.3)	43.4 (42.3-44.5)	39.1 (38.1-40.1)	
PPV [%] ↑	17.1 (15.5-18.6)	24.9 (24.3-25.5)	23.0 (22.5-23.6)	18.0 (17.6-18.4)	17.0 (16.7-17.4)	27.0 (26.3-27.8)	28.1 (27.3-28.9)	
FOR [%] ↓ (1 - NPV)	0.105 (0.094-0.116)	0.082 (0.079-0.085)	0.091 (0.088-0.094)	0.061 (0.059-0.064)	0.061 (0.059-0.064)	0.125 (0.121-0.129)	0.134 (0.131-0.138)	
FPR [%] ↓ (1 - specificity)	0.565 (0.539-0.590)	0.419 (0.413-0.426)	0.432 (0.426-0.439)	0.726 (0.717-0.735)	0.776 (0.767-0.786)	0.258 (0.253-0.263)	0.221 (0.216-0.226)	

## Fine-tuning a large language model

- Large language models are effective in a wide range of NLP tasks [15, 46].
- Might BERT be useful for recognizing stroke?

Subset	Predictor	F1-score [%] ↑	Sensitivity [%] ↑	PPV [%] ↑	FOR [%] ↓ (1 - NPV)	FPR [%] ↓ (1 - specificity)
Overall	<b>Call-takers</b>	25.8 (23.7-27.9)	52.7 (49.2-56.4)	17.1 (15.5-18.6)	0.105 (0.094-0.116)	0.565 (0.539-0.590)
	<b>MLP</b>	35.7 (35.0-36.4)	63.0 (62.0-64.1)	24.9 (24.3-25.5)	0.082 (0.079-0.085)	0.419 (0.413-0.426)
	<b>BERT (fine-tuned)</b>	33.8 (31.5-36.2)	57.5 (53.9-60.9)	23.9 (21.9-25.9)	0.094 (0.084-0.104)	0.403 (0.381-0.424)

## UNCERTAINTY AND THE MEDICAL INTERVIEW

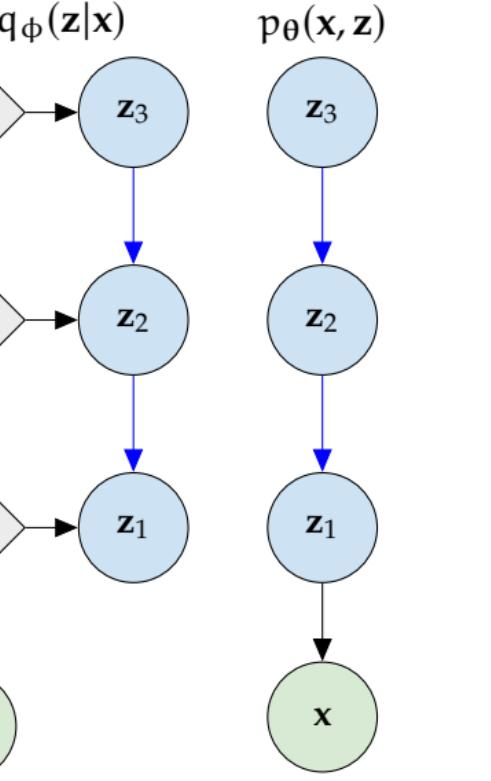
- 2024-03-05
- └ a retrospective study on machine learning-assisted stroke recognition for medical helpline calls
    - └ Fine-tuning a large language model

- Large language models are effective in a wide range of NLP tasks [15, 46].
- Might BERT be useful for recognizing stroke?

Subset	Predictor	F1-score [%] ↑	Sensitivity [%] ↑	PPV [%] ↑	FOR [%] ↓ (1 - NPV)	FPR [%] ↓ (1 - specificity)
Overall	Call-takers	25.8 (23.7-27.9)	52.7 (49.2-56.4)	17.1 (15.5-18.6)	0.105 (0.094-0.116)	0.565 (0.539-0.590)
	MLP	35.7 (35.0-36.4)	63.0 (62.0-64.1)	24.9 (24.3-25.5)	0.082 (0.079-0.085)	0.419 (0.413-0.426)
	BERT (fine-tuned)	33.8 (31.5-36.2)	57.5 (53.9-60.9)	23.9 (21.9-25.9)	0.094 (0.084-0.104)	0.403 (0.381-0.424)

## The Ladder Variational Autoencoder (LVAE)

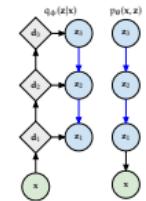
This is a Ladder VAE with three latent variables.



## UNCERTAINTY AND THE MEDICAL INTERVIEW

vae background and the ladder variational autoencoder (lvae)

The Ladder Variational Autoencoder (LVAE)



Understanding the Ladder VAE is not so much about understanding the specific architecture as it is about:

- a) understanding the reason for choosing it,
- b) which other options were available,
- c) and why they don't work as well.

VAE BACKGROUND AND THE LADDER VARIATIONAL AUTOENCODER (LVAE)  
**A generative model from latent variables**



Suppose data  $\mathbf{x} \sim p(\mathbf{x})$  is generated via some underlying *latent* variable  $\mathbf{z}$ . Then

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \quad (12)$$

We would like to

- a) *infer* the latent variables  $\mathbf{z}$  given  $\mathbf{x}$
- b) *generate* the observed variable  $\mathbf{x}$  given  $\mathbf{z}$



2024-03-05

**UNCERTAINTY AND THE MEDICAL INTERVIEW**  
└ vae background and the ladder variational autoencoder (lvae)  
└ A generative model from latent variables

VAE BACKGROUND AND THE LADDER VARIATIONAL AUTOENCODER (LVAE)  
**A generative model from latent variables**

Suppose data  $\mathbf{x} \sim p(\mathbf{x})$  is generated via some underlying latent variable  $\mathbf{z}$ . Then  
$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \quad (12)$$

We would like to

- a) infer the latent variables  $\mathbf{z}$  given  $\mathbf{x}$
- b) generate the observed variable  $\mathbf{x}$  given  $\mathbf{z}$



**Exact inference**

We can choose some simple model for  $p(x, z)$  (denoted  $p_\theta$  and parameterized by  $\theta$ ).

Bayes theorem then gives us the true model posterior,

$$p_\theta(z|x) = \frac{p_\theta(x,z)}{p_\theta(x)} = \frac{p_\theta(x|z)p_\theta(z)}{\int p_\theta(x|z)p_\theta(z)dz}. \quad (13)$$

But this only works if we can integrate over the latent variable  $z$  and the model  $p_\theta(x|z)$ .



2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

- └ vae background and the ladder variational autoencoder (lvae)

- └ Exact inference

We can choose some simple model for  $p(x, z)$  (denoted  $p_\theta$  and parameterized by  $\theta$ ).

Bayes theorem then gives us the true model posterior,

$$p_\theta(z|x) = \frac{p_\theta(x,z)}{p_\theta(x)} = \frac{p_\theta(x|z)p_\theta(z)}{\int p_\theta(x|z)p_\theta(z)dz} \quad (13)$$

But this only works if we can integrate over the latent variable  $z$  and the model  $p_\theta(x|z)$ .



## Exact inference becomes intractable



Suppose we want to model complex data where  $|x| = D \gg 1$ , such as images, audio or graphs.

We might choose to parameterize  $p(x, z)$  with a neural network model with parameters  $\theta$  and try to *learn*  $p_\theta(x) \approx p(x)$ .

$$p_\theta(x, z) = p_\theta(x|z)p(z) \quad (14)$$

where we could choose  $p(z) = \mathcal{N}(0, 1)$ .

However, this comes at the cost of making integrals over the latent variables intractable and hence, **we can no longer directly compute  $p_\theta(x)$** .

## UNCERTAINTY AND THE MEDICAL INTERVIEW

vae background and the ladder variational autoencoder (lvae)

Exact inference becomes intractable

2024-03-05

Suppose we want to model complex data where  $|x| = D \gg 1$ , such as images, audio or graphs.  
We might choose to parameterize  $p(x, z)$  with a neural network model with parameters  $\theta$  and try to *learn*  $p_\theta(x) \approx p(x)$

$p_\theta(x, z) = p_\theta(x|z)p(z)$   
where we could choose  $p(z) = \mathcal{N}(0, 1)$ .

However, this comes at the cost of making integrals over the latent variables intractable and hence, **we can no longer directly compute  $p_\theta(x)$** .

**Variational inference and the ELBO**Introduce some *variational* distribution  $q_\phi(z|x)$  (parameterized by  $\phi$ )

$$\begin{aligned}
 \log p(x) &= \log \int p_\theta(x, z) dz \\
 &= \log \int q_\phi(z|x) \frac{p_\theta(x, z)}{q_\phi(z|x)} dz \\
 &\geq \int q_\phi(z|x) \log \frac{p_\theta(x, z)}{q_\phi(z|x)} dz \\
 &= \mathbb{E}_{q_\phi(z|x)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right] \\
 &= \mathbb{E}_{q_\phi(z|x)} \left[ \log p_\theta(x|z) + \log p_\theta(z) - \log q_\phi(z|x) \right] \\
 &\equiv \underbrace{\mathcal{L}(x; \theta, \phi)}_{\text{evidence lower bound}}
 \end{aligned} \tag{15}$$

**UNCERTAINTY AND THE MEDICAL INTERVIEW**

└ vae background and the ladder variational autoencoder (lvae)

└ Variational inference and the ELBO

2024-03-05

VAE BACKGROUND AND THE LADDER VARIATIONAL AUTOENCODER (LVAE)  
**Variational inference and the ELBO**  
 Introduce some *variational* distribution  $q_\phi(z|x)$  (parameterized by  $\phi$ )

$$\begin{aligned}
 \log p(x) &= \log \int p_\theta(x, z) dz \\
 &= \log \int q_\phi(z|x) \frac{p_\theta(x, z)}{q_\phi(z|x)} dz \\
 &> \int q_\phi(z|x) \log \frac{p_\theta(x, z)}{q_\phi(z|x)} dz \\
 &= \mathbb{E}_{q_\phi(z|x)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right] \\
 &= \mathbb{E}_{q_\phi(z|x)} \left[ \log p_\theta(x|z) + \log p_\theta(z) - \log q_\phi(z|x) \right] \\
 &\equiv \underbrace{\mathcal{L}(x; \theta, \phi)}_{\text{evidence lower bound}}
 \end{aligned} \tag{15}$$

## The variational autoencoder (VAE)

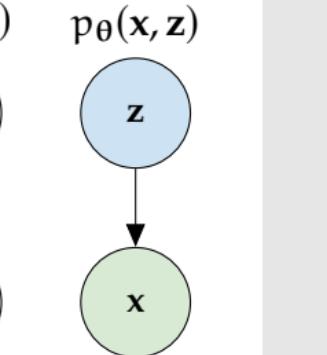


Suppose we parameterize  $q_\phi(z|x)$  by a model with parameters  $\phi$ . Then we can optimize both  $\{\theta, \phi\}$  jointly by maximizing the ELBO.

$$\{\theta^*, \phi^*\} = \arg \max_{\{\theta, \phi\}} \mathcal{L}(x; \theta, \phi) \quad (16)$$

The result is the VAE [33] consisting of

- a) an *inference model* (or *encoder*)  $q_\phi(z|x)$  which approximates the intractable true posterior  $p(z|x)$ .
- b) a *generative model* (or *decoder*)  $p_\theta(x|z)$  which can generate new samples from the prior  $p(z)$ , or "reconstruct" with proposals from  $q_\phi(z|x)$ .



## UNCERTAINTY AND THE MEDICAL INTERVIEW

vae background and the ladder variational autoencoder (lvae)

The variational autoencoder (VAE)

2024-03-05

Suppose we parameterize  $q_\phi(z|x)$  by a model with parameters  $\phi$ . Then we can optimize both  $\{\theta, \phi\}$  jointly by maximizing the ELBO.

$$\{\theta^*, \phi^*\} = \arg \max_{\{\theta, \phi\}} \mathcal{L}(x; \theta, \phi) \quad (16)$$

The result is the VAE [33] consisting of

- a) an *inference model* (or *encoder*)  $q_\phi(z|x)$  which approximates the intractable true posterior  $p(z|x)$ .
- b) a *generative model* (or *decoder*)  $p_\theta(x|z)$  which can generate new samples from the prior  $p(z)$ , or "reconstruct" with proposals from  $q_\phi(z|x)$ .



## Limitations of VAEs

The VAE uses a mean field approximation for most common variational posteriors.

$$q(\mathbf{z}) = \prod_i q(z_i) \quad (17)$$

- Assumes independence between latent variables.
- Model cannot learn covariance between latents.
- This limits expressivity as we cannot expect to always match the true posterior well.

*E.g. a vehicle's color is often dependent on its type (fire truck, police car, bus, taxi etc.) but these are coded independently in  $\mathbf{z}$ .*

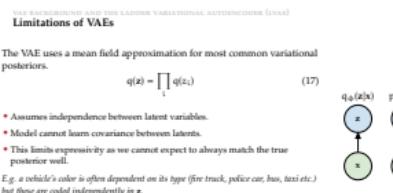
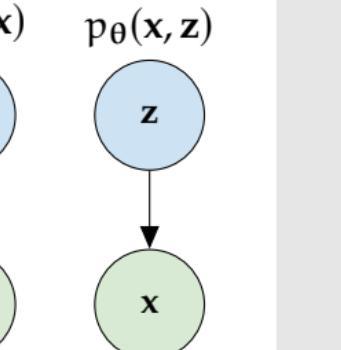


2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

vae background and the ladder variational autoencoder (lvae)

Limitations of VAEs



## VAE BACKGROUND AND THE LADDER VARIATIONAL AUTOENCODER (LVAE)

### Hierarchical VAE

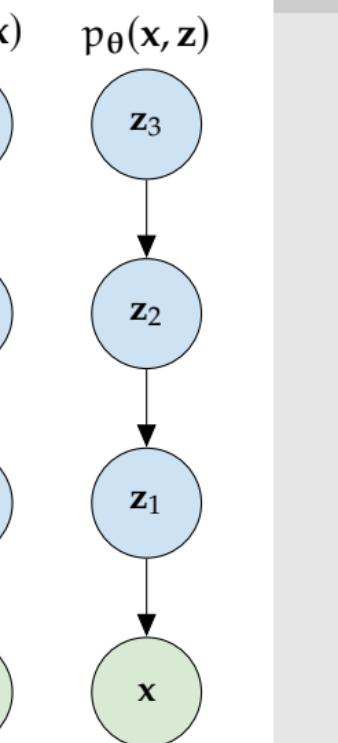
To avoid these limitations, we can introduce a hierarchy of additional latent variables  $\mathbf{z} = \mathbf{z}_1, \dots, \mathbf{z}_L$ . For  $L = 3$ ,

$$p_{\theta}(\mathbf{x}|\mathbf{z}) = p_{\theta}(\mathbf{x}|\mathbf{z}_1)p_{\theta}(\mathbf{z}_1|\mathbf{z}_2)p(\mathbf{z}_3). \quad (18)$$

We can straightforwardly generalize the inference model,

$$q_{\phi}(\mathbf{z}|\mathbf{x}) = q_{\phi}(\mathbf{z}_1|\mathbf{x})q_{\phi}(\mathbf{z}_2|\mathbf{z}_1)q_{\phi}(\mathbf{z}_3|\mathbf{z}_2). \quad (19)$$

This is called *bottom-up* inference.



## UNCERTAINTY AND THE MEDICAL INTERVIEW

vae background and the ladder variational autoencoder (lvae)

Hierarchical VAE

2024-03-05

VAE BACKGROUND AND THE LADDER VARIATIONAL AUTOENCODER (LVAE)  
Hierarchical VAE

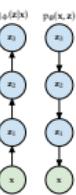
To avoid these limitations, we can introduce a hierarchy of additional latent variables  $\mathbf{z} = \mathbf{z}_1, \dots, \mathbf{z}_L$ . For  $L = 3$ ,

$$p_{\theta}(\mathbf{x}|\mathbf{z}) = p_{\theta}(\mathbf{x}|\mathbf{z}_1)p_{\theta}(\mathbf{z}_1|\mathbf{z}_2)p(\mathbf{z}_3).$$

We can straightforwardly generalize the inference model,

$$q_{\phi}(\mathbf{z}|\mathbf{x}) = q_{\phi}(\mathbf{z}_1|\mathbf{x})q_{\phi}(\mathbf{z}_2|\mathbf{z}_1)q_{\phi}(\mathbf{z}_3|\mathbf{z}_2).$$

This is called *bottom-up* inference.



## Challenges of hierarchical VAEs

[View slide](#)



Consider a simple model  $p_{\text{simple}}(x)$  without any latent variables. We can rewrite the likelihood as,

$$\begin{aligned} \mathbb{E}_{p(x)} [\log p_{\text{simple}}(x)] &= \mathbb{E}_{p(x)} \left[ \log \left( p(x) \frac{p_{\text{simple}}(x)}{p(x)} \right) \right] \\ &= - \underbrace{\mathcal{H}(p)}_{\text{data entropy}} - \underbrace{D_{\text{KL}}(p(x) || p_{\text{simple}}(x))}_{\text{divergence from data distribution}} \end{aligned}$$

where  $\mathcal{H}(p) = \mathbb{E}_{p(x)}[\log p(x)]$ .

# UNCERTAINTY AND THE MEDICAL INTERVIEW

└ vae background and the ladder variational autoencoder (lvae)

└ Challenges of hierarchical VAEs

2024-03-05

Consider a simple model  $p_{\text{simple}}(x)$  without any latent variables. We can rewrite the likelihood as,

$$\begin{aligned} \mathbb{E}_{p(x)} [\log p_{\text{simple}}(x)] &= \mathbb{E}_{p(x)} \left[ \log \left( p(x) \frac{p_{\text{simple}}(x)}{p(x)} \right) \right] \\ &= - \underbrace{\mathcal{H}(p)}_{\text{data entropy}} - \underbrace{D_{\text{KL}}(p(x) || p_{\text{simple}}(x))}_{\text{divergence from data distribution}} \end{aligned}$$

where  $\mathcal{H}(p) = \mathbb{E}_{p(x)}[\log p(x)]$ .

## Challenges of hierarchical VAEs

Let's do the same for a VAE  $p_\theta(x)$  with ELBO given by

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)} \left[ \log \frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)} \right]. \quad (20)$$

The expectation over the data becomes,

$$\begin{aligned} \mathbb{E}_{p(x)} [\log p_\theta(x)] &\geq \mathbb{E}_{p(x)} \left[ \log \left( p(x) \frac{p_\theta(x)}{p(x)} \right) \right] \\ &= -\mathcal{H}(p) - D_{KL}(p(x)||p_\theta(x)) \\ &\quad - \underbrace{\mathbb{E}_{p(x)} [D_{KL}(q_\phi(z|x)||p_\theta(z|x))]}_{\text{divergence from true posterior}} \end{aligned} \quad (21)$$

Compared to the simple model, we incur an **additional cost** for the latent variable given by the divergence from the true model posterior.



2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

vae background and the ladder variational autoencoder (lvae)

Challenges of hierarchical VAEs

VAE BACKGROUND AND THE LADDER VARIATIONAL AUTOENCODER (LVAE)  
Challenges of hierarchical VAEs  
Let's do the same for a VAE  $p_\theta(x)$  with ELBO given by

$$\log p_\theta(x) > \mathbb{E}_{q_\phi(z|x)} \left[ \log \frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)} \right]. \quad (20)$$

The expectation over the data becomes,

$$\begin{aligned} \mathbb{E}_{p(x)} [\log p_\theta(x)] &> \mathbb{E}_{p(x)} \left[ \log \left( p(x) \frac{p_\theta(x)}{p(x)} \right) \right] \\ &= -\mathcal{H}(p) - D_{KL}(p(x)||p_\theta(x)) \\ &\quad - \mathbb{E}_{p(x)} [D_{KL}(q_\phi(z|x)||p_\theta(z|x))] \end{aligned} \quad (21)$$

Compared to the simple model, we incur an **additional cost** for the latent variable given by the divergence from the true model posterior.

## Challenges of hierarchical VAEs

For a hierarchical VAE with two latent variables

$$\begin{aligned}
 \mathbb{E}_{p(x)} [\log p_{\theta}(x)] &\geq \mathbb{E}_{p(x)} \left[ \log \left( p(x) \frac{p_{\theta}(x)}{p(x)} \right) \right] \\
 &= -\mathcal{H}(p) - D_{KL} (p(x) || p_{\theta}(x)) \\
 &\quad - \mathbb{E}_{p(x)} \mathbb{E}_{q_{\phi}(z_1|x)} [D_{KL} (q_{\phi}(z_2|z_1) || p_{\theta}(z_2|z_1))] \\
 &\quad - \mathbb{E}_{p(x)} [D_{KL} (q_{\phi}(z_1|x) || p_{\theta}(z_1|x))] \tag{22}
 \end{aligned}$$

The cost is **incurred for each additional latent** we add to the hierarchy.

Hence, each latent will only be used by the model if we get an **opposite equal or larger improvement** in the ELBO.



└ vae background and the ladder variational autoencoder (lvae)

└ Challenges of hierarchical VAEs

2024-03-05

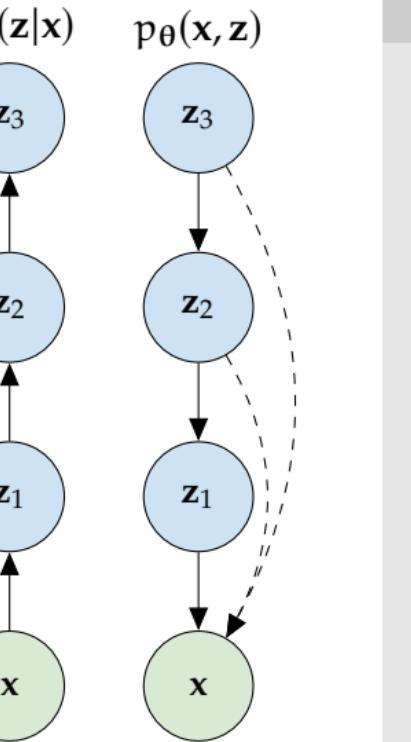
VAE BACKGROUND AND THE LADDER VARIATIONAL AUTOENCODER (LVAE)  
Challenges of hierarchical VAEs

For a hierarchical VAE with two latent variables

$$\begin{aligned}
 \mathbb{E}_{p(x)} [\log p_{\theta}(x)] &\geq \mathbb{E}_{p(x)} \left[ \log \left( p(x) \frac{p_{\theta}(x)}{p(x)} \right) \right] \\
 &= -\mathcal{H}(p) - D_{KL} (p(x) || p_{\theta}(x)) \\
 &\quad - \mathbb{E}_{p(x)} \mathbb{E}_{q_{\phi}(z_1|x)} [D_{KL} (q_{\phi}(z_2|z_1) || p_{\theta}(z_2|z_1))] \\
 &\quad - \mathbb{E}_{p(x)} [D_{KL} (q_{\phi}(z_1|x) || p_{\theta}(z_1|x))] \tag{22}
 \end{aligned}$$

The cost is incurred for each additional latent we add to the hierarchy.  
Hence, each latent will only be used by the model if we get an opposite equal or larger improvement in the ELBO.

## (Failing to) Dodge the challenges



There's a few common ways to try and dodge these issues:

- Skip connections [16, 40]
- Free bits in the KL-terms [34]
- Deterministic warmup [53]
- Batch normalization/weight normalization [28, 49]

None of them work for more than around 5 latent variables.

## UNCERTAINTY AND THE MEDICAL INTERVIEW

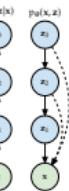
vae background and the ladder variational autoencoder (lvae)

(Failing to) Dodge the challenges

2024-03-05

- There's a few common ways to try and dodge these issues:
- Skip connections [16, 40]
  - Free bits in the KL-terms [34]
  - Deterministic warmup [53]
  - Batch normalization/weight normalization [28, 49]

None of them work for more than around 5 latent variables.

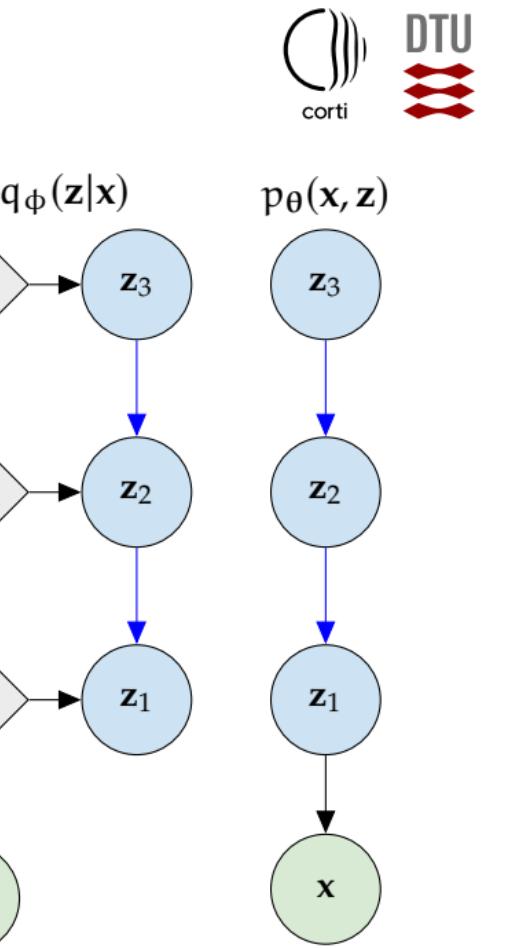


## The Ladder VAE

The LVAE introduces a *top-down* inference path,

$$q_{\phi}(\mathbf{z}|\mathbf{x}) = q_{\phi}(\mathbf{z}_L|\mathbf{x}) \prod_{i=1}^{L-1} q_{\phi}(\mathbf{z}_i|\mathbf{z}_{i+1}). \quad (23)$$

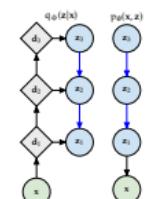
It is aided by a **deterministic bottom-up path** and **parameter sharing** between the inference and generative models.



The LVAE introduces a *top-down* inference path,

$$q_{\phi}(\mathbf{z}|\mathbf{x}) = q_{\phi}(\mathbf{z}_L|\mathbf{x}) \prod_{i=1}^{L-1} q_{\phi}(\mathbf{z}_i|\mathbf{z}_{i+1}). \quad (23)$$

It is aided by a **deterministic bottom-up path** and **parameter sharing** between the inference and generative models.



Model	$\geq \log p(x)$
VAE + NF (L = 1)	-85.10
IWAE (L = 2, K = 1)	-85.33
IWAE (L = 2, K = 50)	-82.90
VAE + VGP (L = 2)	-81.90
<hr/>	
LVAE (L = 5)	-82.12
LVAE + finetuning (L = 5)	-81.84
LVAE + finetuning (L = 5, K = 10)	-81.74

Table 1: Results on dynamically binarized MNIST.

Model	$\geq \log p(x)$
VAE + NF (L = 1)	-85.10
IWAE (L = 2, K = 1)	-85.33
IWAE (L = 2, K = 50)	-82.90
VAE + VGP (L = 2)	-81.90
<hr/>	
LVAE (L = 5)	-82.12
LVAE + finetuning (L = 5)	-81.84
LVAE + finetuning (L = 5, K = 10)	-81.74

Table 1: Results on dynamically binarized MNIST.

# Latent space activation

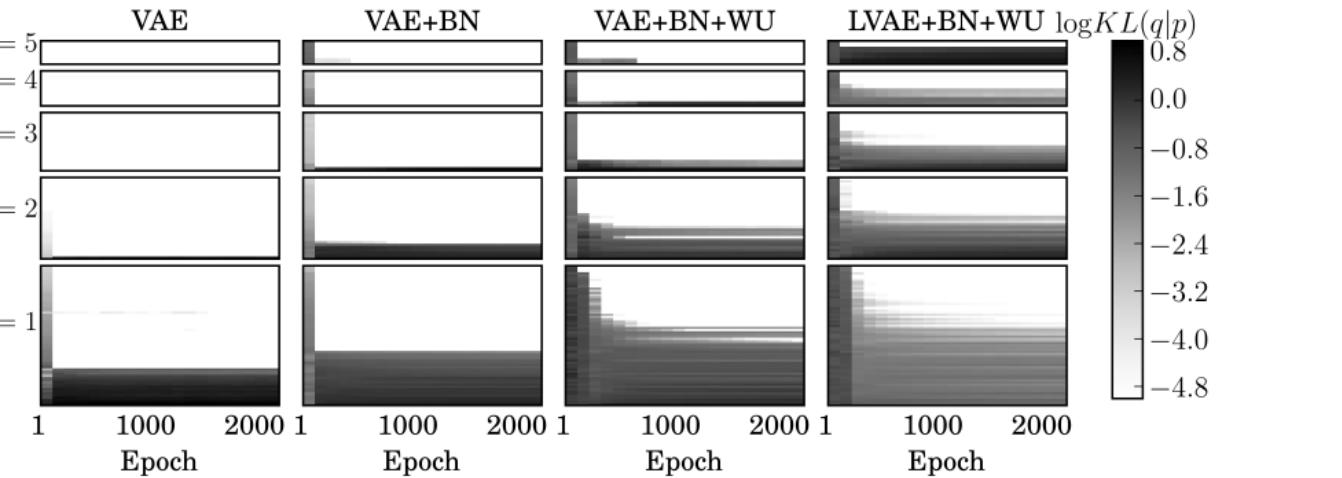


Figure 2: Element-wise KL-divergences in each latent variable  $q(z_i|z_{i-1})$  with  $z_0 \equiv x$ .

# UNCERTAINTY AND THE MEDICAL INTERVIEW

vae background and the ladder variational autoencoder (lvae)

Latent space activation

2024-03-05

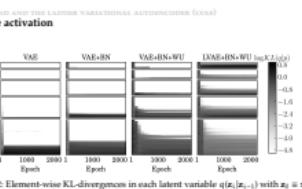


Figure 2: Element-wise KL-divergences in each latent variable  $q(z_i|z_{i-1})$  with  $z_0 \equiv x$ .

## Latent space representation

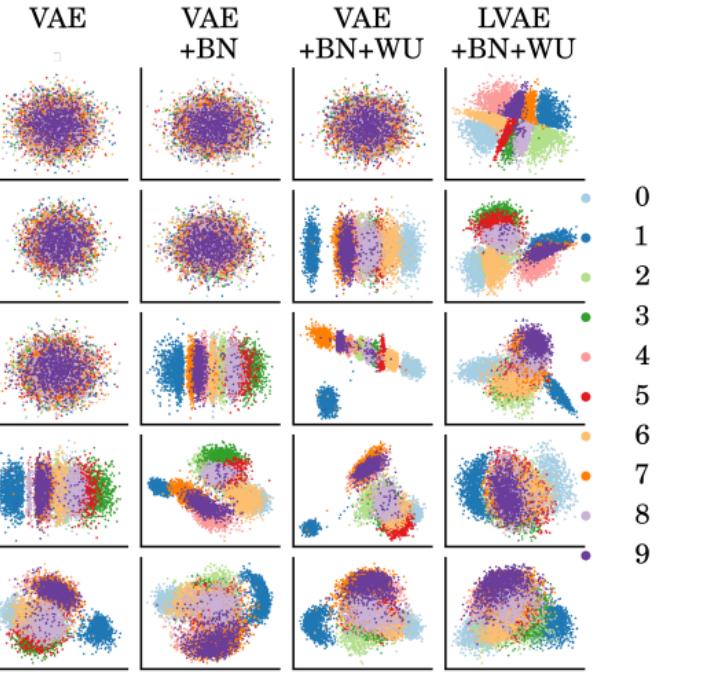


Figure 3: PCA of latent samples from  $q(z_i|z_{i-1})$  with  $z_0 \equiv x$ .

# UNCERTAINTY AND THE MEDICAL INTERVIEW

vae background and the ladder variational autoencoder (lvae)

Latent space representation

2024-03-05

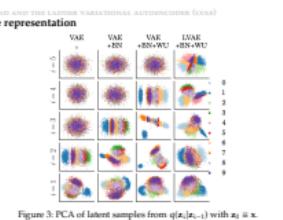


Figure 3: PCA of latent samples from  $q(z_i|z_{i-1})$  with  $z_0 \equiv x$ .

## Recent work



Figure 4: Samples from the generative model of [11] with more than 70 latents.

2024-03-05

# UNCERTAINTY AND THE MEDICAL INTERVIEW

vae background and the ladder variational autoencoder (lvae)

Recent work



Figure 4: Samples from the generative model of [11] with more than 70 latents.

## OVERVIEW

### Table of contents I

- project
- introduction
- hierarchical vaes know what they don't know
- overview
- a brief overview of unsupervised speech representation learning
- a retrospective study on machine learning-assisted stroke recognition for medical helpline calls
- discussion
- bibliography
- links
- introduction
- hierarchical vaes know what they don't know



2024-03-05

## UNCERTAINTY AND THE MEDICAL INTERVIEW

### Overview

### Table of contents

Overview	Table of contents I
• project	
• introduction	
• hierarchical vaes know what they don't know	
• overview	
• a brief overview of unsupervised speech representation learning	
• a retrospective study on machine learning-assisted stroke recognition for medical helpline calls	
• discussion	
• bibliography	
• links	
• introduction	
• hierarchical vaes know what they don't know	

## OVERVIEW

### Table of contents II

- a brief overview of unsupervised speech representation learning
- a retrospective study on machine learning-assisted stroke recognition for medical helpline calls
- vae background and the ladder variational autoencoder (lvae)



## UNCERTAINTY AND THE MEDICAL INTERVIEW

2024-03-05  
└ Overview

└ Table of contents

Overview  
Table of contents II  
• a brief overview of unsupervised speech representation learning  
• a retrospective study on machine learning-assisted stroke recognition for medical helpline calls  
• vae background and the ladder variational autoencoder (lvae)