

TOWARDS SELF-ASSESSMENT IN MACHINE LEARNING MODELS

Jakob Drachmann Havtorn



UNCERTAINTY AND THE MEDICAL INTERVIEW

UNCERTAINTY AND THE MEDICAL INTERVIEW

2024-03-03

- 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**



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

- Overview**
- Thesis**

Overview	
Overview	Thesis
<hr/>	
chapter 1.1	INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND
chapter 1.2	HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
chapter 1.3	MODEL-AGNOSTIC OUT-OF-DISTRIBUTION DETECTION
chapter 1.4	USING COMBINED STATISTICAL TESTS
chapter 1.5	A BRIEF OVERVIEW OF UNSUPERVISED SPEECH
chapter 1.6	REPRESENTATION LEARNING
chapter 1.7	BENCHMARKING LATENT VARIABLE MODELS FOR SPEECH
chapter 1.8	AUTOMATED MEDICAL CODING ON MIMIC-III AND
chapter 1.9	MIMIC-IV: A CRITICAL REVIEW AND REPLICABILITY STUDY
chapter 1.10	A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS
<hr/>	
chapter 1.11	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

2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

Overview

Thesis

Overview Thesis

introduction, research questions, and background

- introduction
- model-agnostic out-of-distribution detection
- 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

Conclusion 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

2024-03-03

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 What is healthcare?

Healthcare is the improvement of health via the prevention, diagnosis, treatment, amelioration or cure of disease, illness, injury, and other physical and mental impairments in people.

4



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

└ Introduction

└ What is healthcare?

1. Machine learning for healthcare touches on most of these aspects. Fx:
2.
 - **Prevention** by predicting risk of disease allowing for early intervention.
 - **Diagnosis** by detecting stroke in emergency calls.
 - **Treatment** by suggesting medication.

INTRODUCTION
What is healthcare?

Healthcare is the improvement of health via the prevention, diagnosis, treatment, amelioration or cure of disease, illness, injury, and other physical and mental impairments in people.

— Wikipedia

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



2024-03-03

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**:



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

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



INTRODUCTION

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 [11].



- 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 [11].



Documenting medical encounters

- Documentation is a central part of healthcare.
- E.g. patient records, insurance claims, billing, research, training, legal purposes.



UNCERTAINTY AND THE MEDICAL INTERVIEW

Introduction

Documenting medical encounters

- Documentation is a central part of healthcare.
- E.g. patient records, insurance claims, billing, research, training, legal purposes.



Documenting medical encounters

- Documentation is a central part of healthcare.
- E.g. patient records, insurance claims, billing, research, training, legal purposes.
- **Time-consuming:** Physicians spend 34-37% of their time on documentation [28, 53, 55]^a.
- **Varying quality:** Discharge summaries almost never meet *all* timeline, transmission, and content criteria. [24]^b

^aAmbulatory care across four specialties in four states and tertiary care at an academic medical center.

^bOutpatient visits, Yale-New Haven Hospital.



- Documentation is a central part of healthcare.
- E.g. patient records, insurance claims, billing, research, training, legal purposes.

- **Time-consuming:** Physicians spend 34-37% of their time on documentation [28, 53, 55]^a.
- **Varying quality:** Discharge summaries almost never meet all timeline, transmission, and content criteria. [24]^b

^aAmbulatory care across four specialties in four states and tertiary care at an academic medical center.

^bOutpatient visits, Yale-New Haven Hospital.



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.



- 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.



Reliability of machine learning systems



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

└ Introduction

└ Reliability of machine learning systems

2024-03-03

- 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.

2024-03-03

└ Introduction

└ 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.

Building a decision-support system



2024-03-03

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.).

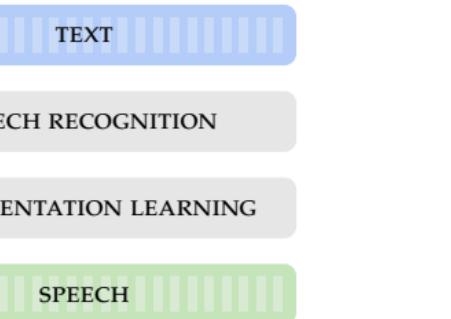


└ Introduction

└ Building a decision-support system

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.



UNCERTAINTY AND THE MEDICAL INTERVIEW

Introduction

Building a decision-support system

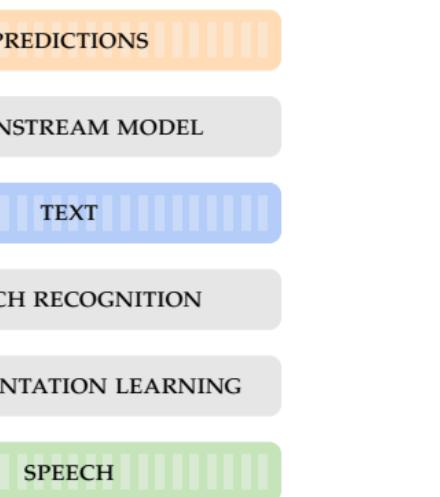
1. We will take a modular approach to building a decision-support system.
2. First we need source data
3. Then we need to convert it into representations useful for downstream tasks.
4. Then we can perform the downstream tasks.
5. Finally, we need to estimate the reliability of our data, representations,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



Introduction

Building a decision-support system

2024-03-03

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

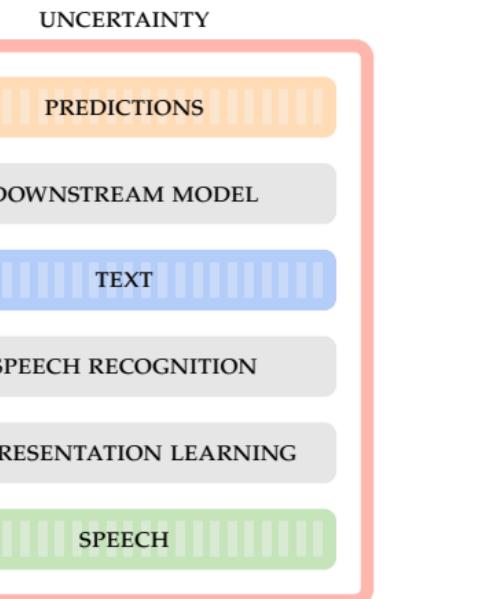
- **Foundation modelling:** Converting the input into representations useful for downstream tasks.

- **Downstream tasks:** Suggesting, summarizing, classifying, transcribing, 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, summarizing, classifying, transcribing, translating, etc.
- **Uncertainty:** Estimating the reliability of data, representations, predictions.



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

Introduction

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, summarizing, classifying, transcribing, translating, etc.
- **Uncertainty:** Estimating the reliability of data, representations, 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

Overview Thesis

2024-03-03

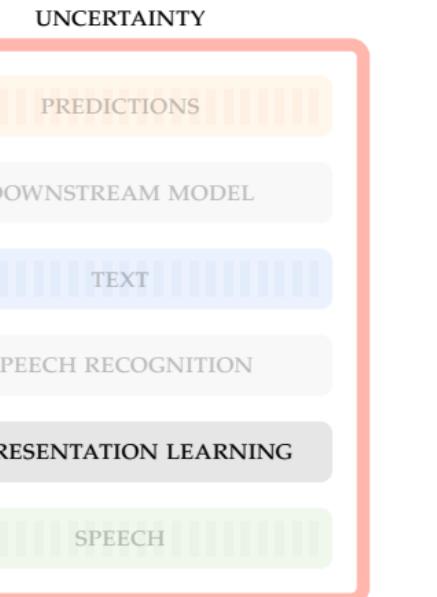
Overview Thesis

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

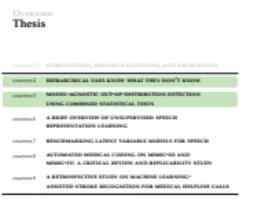


2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

- └ Overview
- └ Thesis

1. So how does the thesis tackle this problem of building such a system?
2. First we examine how a certain class of generative models can be used to detect data that is different from training data.
3. In our second paper we generalize this to be model-agnostic and phrased as a statistical test.



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

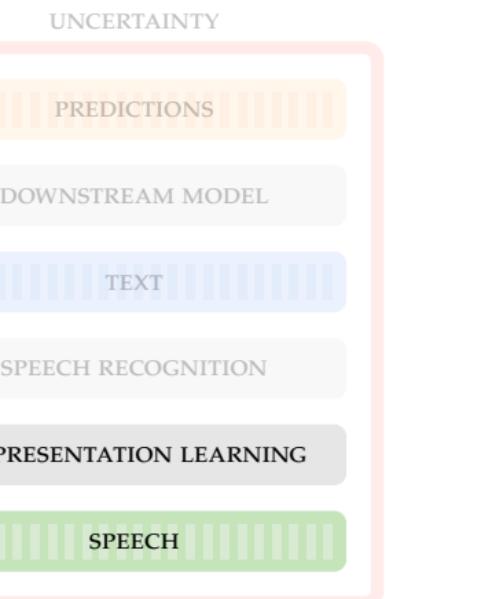
CHAPTER 10 DISCUSSION AND CONCLUSION



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

Overview
└ Thesis



1. So how does the thesis tackle this problem of building such a system?
2. First we examine how a certain class of generative models can be used to detect data that is different from training data.
3. In our second paper we generalize this to be model-agnostic and phrased as a statistical test.
4. We then take a step back and provide an overview of modern methods for unsupervised speech representation learning.



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

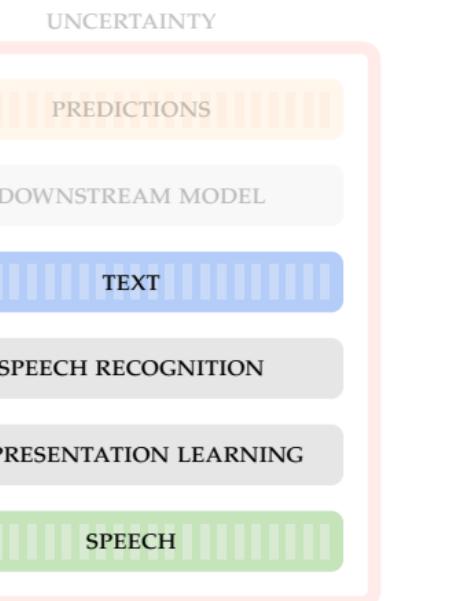
CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH
REPRESENTATION LEARNING

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

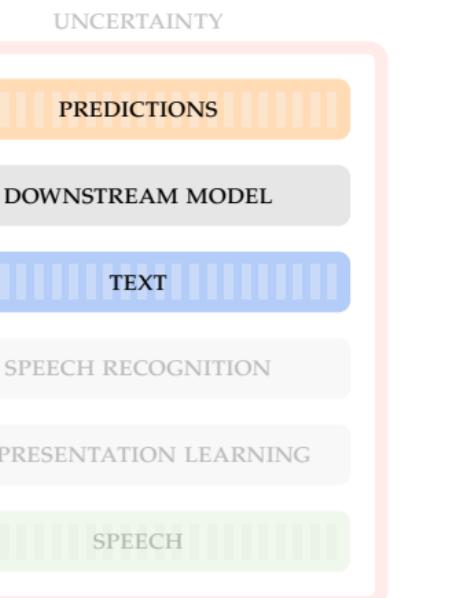
2024-03-03
Overview
Thesis



1. So how does the thesis tackle this problem of building such a system?
2. First we examine how a certain class of generative models can be used to detect data that is different from training data.
3. In our second paper we generalize this to be model-agnostic and phrased as a statistical test.
4. We then take a step back and provide an overview of modern methods for unsupervised speech representation learning.
5. We then benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.

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

2024-03-03

↳ Overview

↳ Thesis

OVERVIEW

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

PREDICTIONS

DOWNTREAM MODEL

TEXT

SPEECH RECOGNITION

REPRESENTATION LEARNING

SPEECH

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

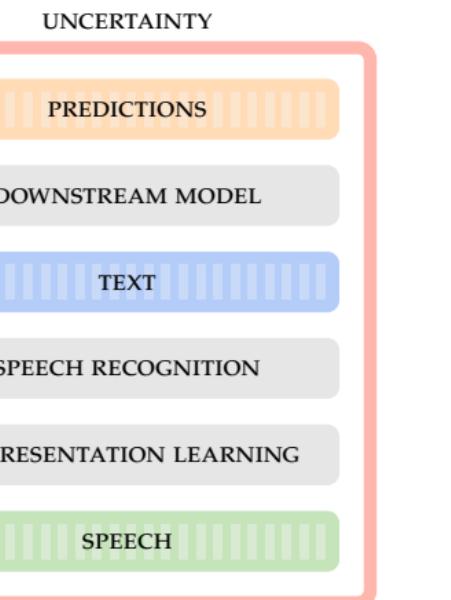
CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH
REPRESENTATION LEARNING

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



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

Overview

Thesis



1. So how does the thesis tackle this problem of building such a system?
2. First we examine how a certain class of generative models can be used to detect data that is different from training data.
3. In our second paper we generalize this to be model-agnostic and phrased as a statistical test.
4. We then take a step back and provide an overview of modern methods for unsupervised speech representation learning.
5. We then benchmark some of these methods, generative latent variable models, on likelihood and automatic speech recognition.
6. We then move to a different domain, medical coding, and provide a critical review of recent works.
7. Finally, we provide a retrospective study on a machine learning-assisted stroke recognition system for medical helpline calls.

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

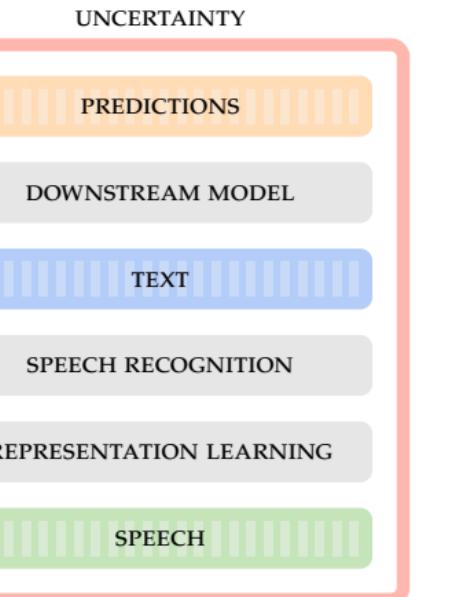
CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH
REPRESENTATION LEARNING

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



UNCERTAINTY AND THE MEDICAL INTERVIEW

2024-03-03

Overview

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

UNCERTAINTY

PREDICTIONS

DOWNSTREAM MODEL

TEXT

SPEECH RECOGNITION

REPRESENTATION LEARNING

SPEECH

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

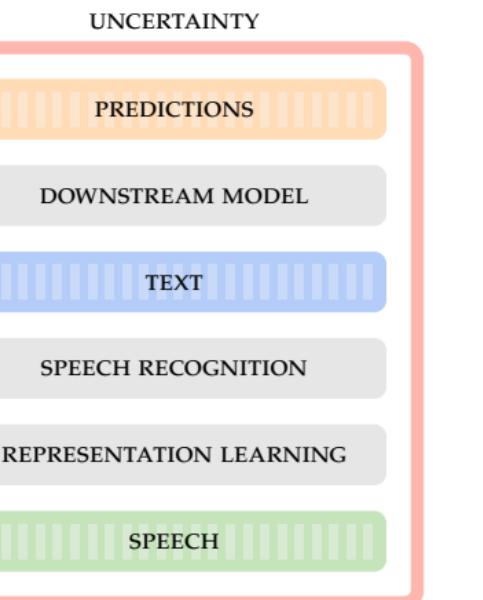
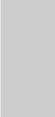
CHAPTER 6 A BRIEF OVERVIEW OF UNSUPERVISED SPEECH
REPRESENTATION LEARNING

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



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

Overview

Presentation

Overview

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

PREDICTIONS
DOWNSTREAM MODELS
TEXT
SPEECH RECOGNITION
REPRESENTATION LEARNING
SPEECH

OVERVIEW Presentation

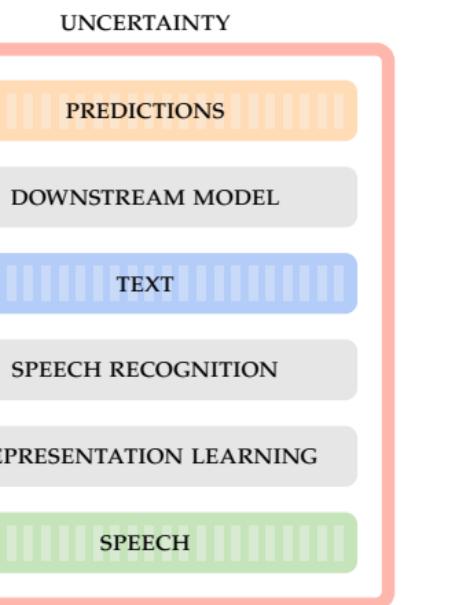
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

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

2024-03-03

UNCERTAINTY

PREDICTIONS

DOWNTSTREAM MODEL

TEXT

SPEECH RECOGNITION

REPRESENTATION LEARNING

SPEECH

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

OVERVIEW Presentation

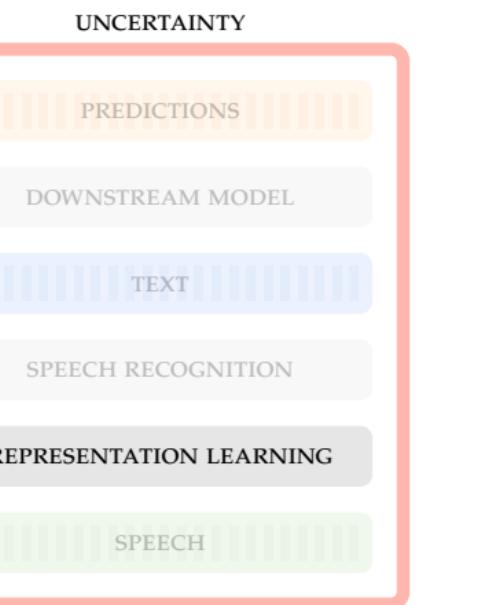
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

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

2024-03-03

Overview

Presentation

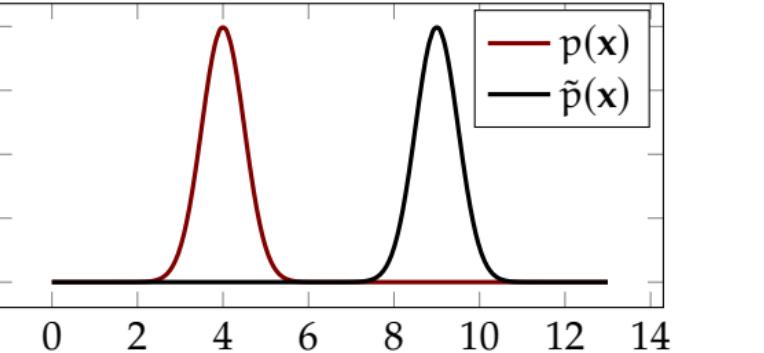
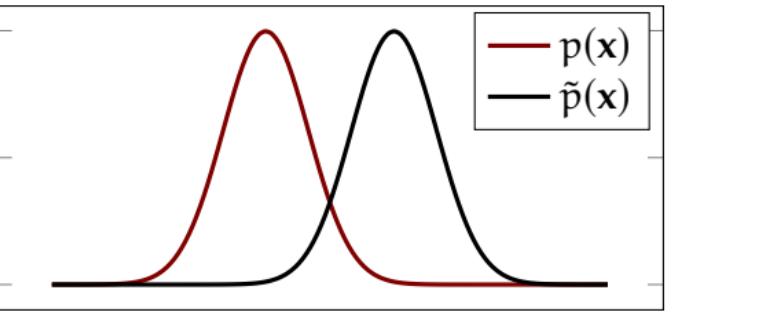


HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
Defining OOD detection

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

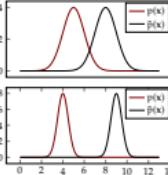
Defining OOD detection

2024-03-03

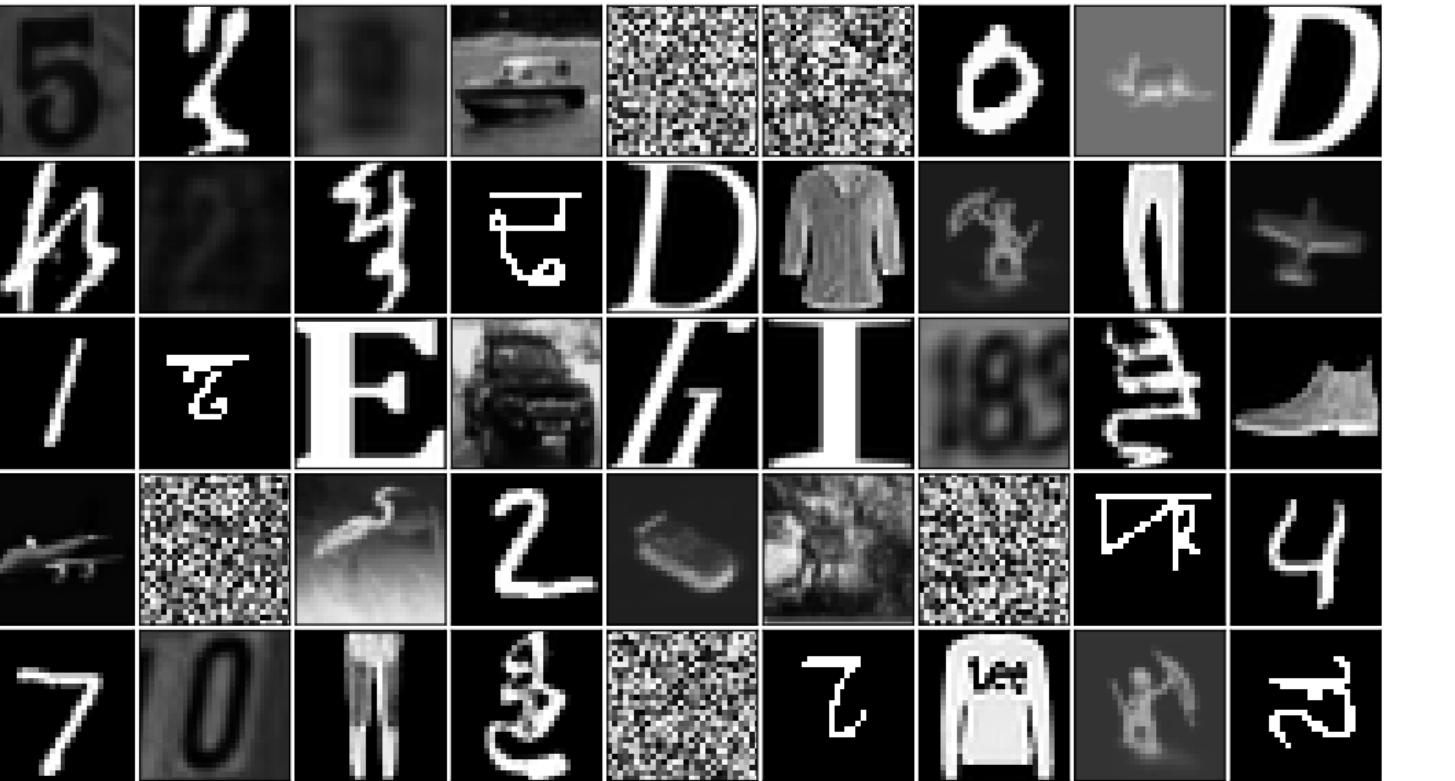
Enable models to distinguish the training data distribution $p(x)$ from any other distribution $\tilde{p}(x)$.

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

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



In distribution?

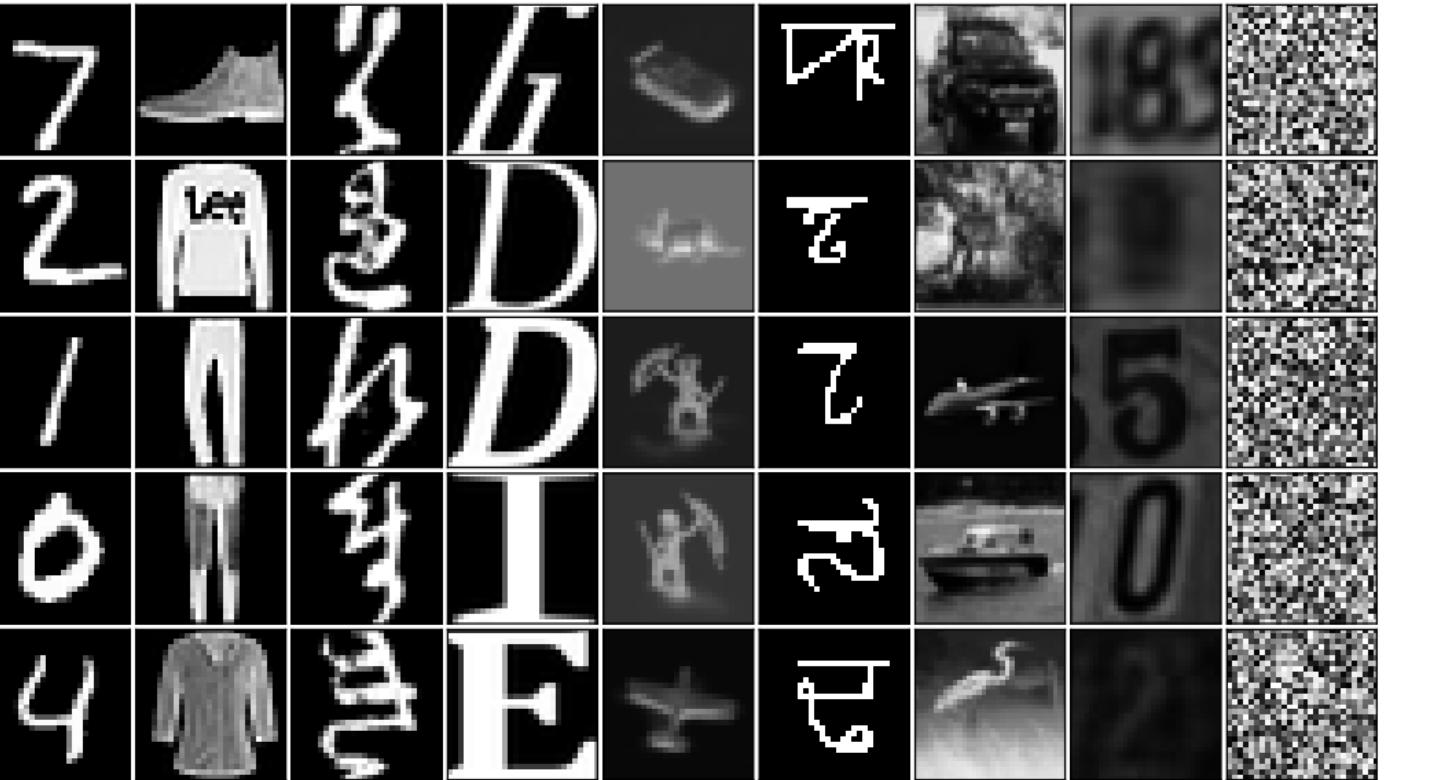


UNCERTAINTY AND THE MEDICAL INTERVIEW

↳ hierarchical vaes know what they don't know

↳ In distribution?



Out of distribution?**UNCERTAINTY AND THE MEDICAL INTERVIEW**

↳ hierarchical vaes know what they don't know

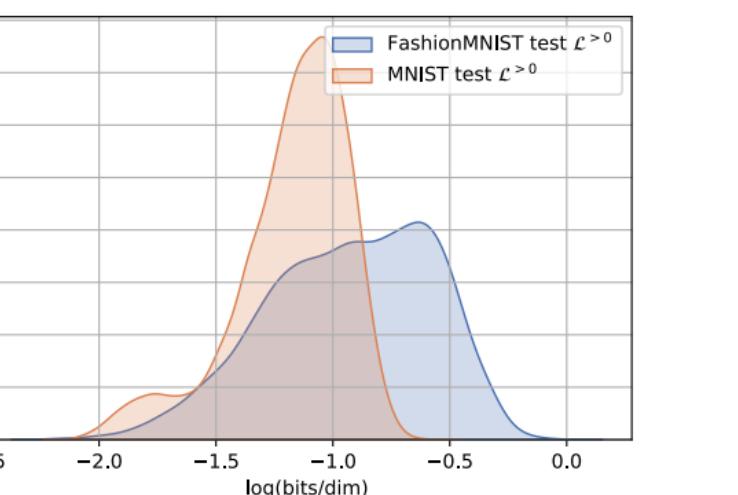
↳ Out of distribution?

1. Datasets can overlap quite a bit in their raw data space.



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 [6].



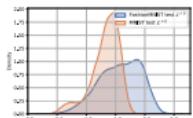
UNCERTAINTY AND THE MEDICAL INTERVIEW

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

↳ Out-of-distribution detection with generative models

2024-03-03

- 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 [6].

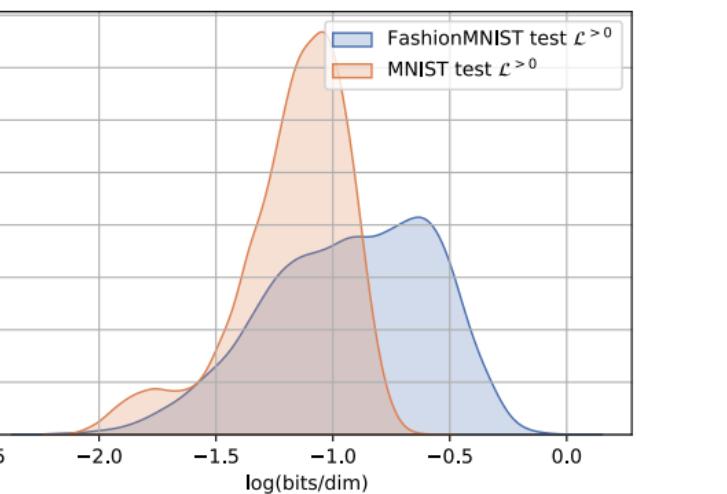


Out-of-distribution detection with generative models

- 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 [6].



HVAE trained on FashionMNIST evaluated on FashionMNIST and MNIST.

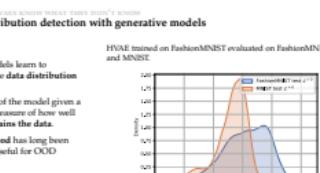


UNCERTAINTY AND THE MEDICAL INTERVIEW

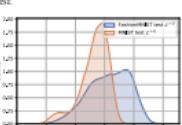
hierarchical vae's know what they don't know

Out-of-distribution detection with generative models

2024-03-03



- 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 [6].



HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
Hierarchical VAE

We choose the hierarchical VAE as our model [32, 50].

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

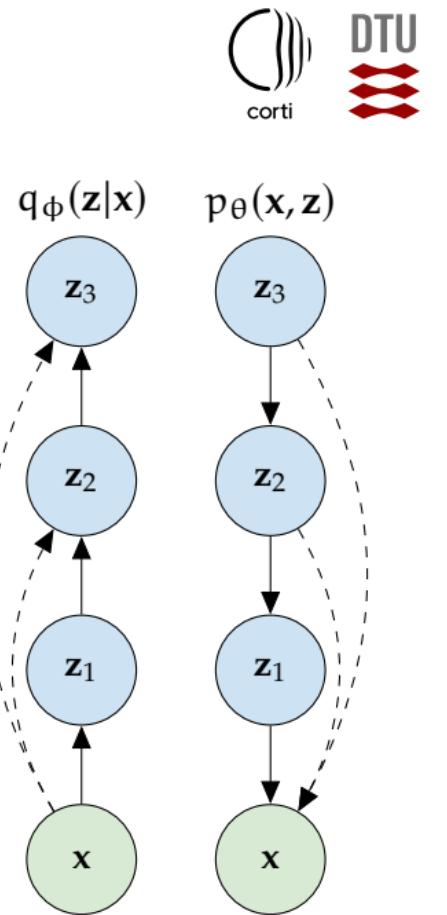
Specifically we use

- 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].



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

hierarchical vae know what they don't know

Hierarchical VAE

We choose the hierarchical VAE as our model [32, 50].

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

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].

q_φ(z|x) p_θ(x, z)

z₃ z₃
z₂ z₂
z₁ z₁
x x

What is wrong with the ELBO for OOD detection?

We can split the ELBO into two terms

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

The first term is high if the data is well-explained by z .

The second term we can rewrite as,

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

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



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

↳ hierarchical vaes know what they don't know

↳ What is wrong with the ELBO for OOD detection?

We can split the ELBO into two terms

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

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

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

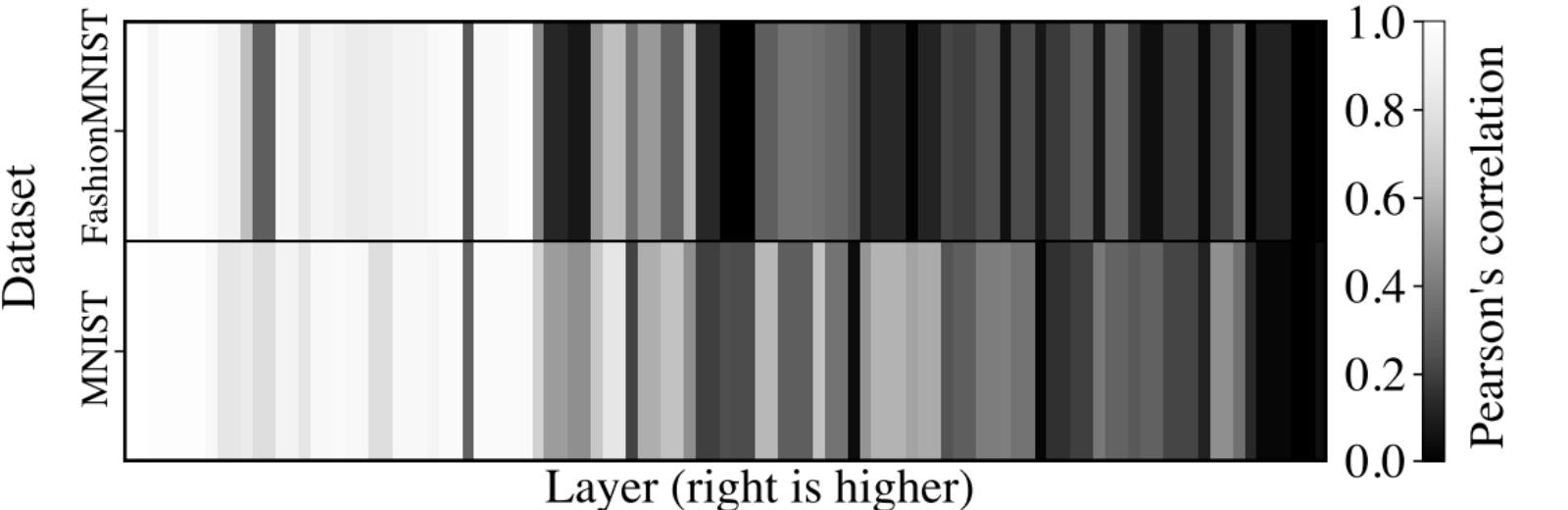
Since the individual terms are computed by summing over the dimensionality of z_i , the absolute log-ratios grow with $\text{dim}(z_i)$.

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW

What do the lowest latent variables code for?

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).

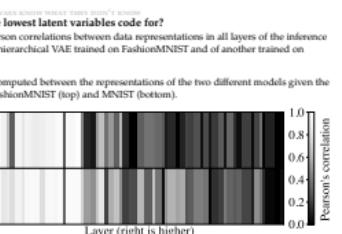


2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

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

↳ What do the lowest latent variables code for?



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]

$$\mathcal{L}^{>k}(x; \theta, \phi) = \mathbb{E}_{p_{\theta}(z_{\leq k}|z_{>k})q_{\phi}(z_{>k}|x)} \left[\log \frac{p_{\theta}(x|z)p_{\theta}(z_{>k})}{q_{\phi}(z_{>k}|x)} \right] \quad (3)$$

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.

2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

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]

$$\mathcal{L}^{>k}(x; \theta, \phi) = \mathbb{E}_{p_{\theta}(z_{\leq k}|z_{>k})q_{\phi}(z_{>k}|x)} \left[\log \frac{p_{\theta}(x|z)p_{\theta}(z_{>k})}{q_{\phi}(z_{>k}|x)} \right] \quad (3)$$

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.

Likelihood ratios

We can use our new bound to compute the score used in a standard likelihood ratio test [10].

$$\text{LLR}^{>k}(x) \equiv \mathcal{L}(x) - \mathcal{L}^{>k}(x). \quad (4)$$

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

$$\mathcal{L} = \log p_\theta(x) - D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x)), \quad (5)$$

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

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

$$\begin{aligned} \text{LLR}^{>k}(x) &= -D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x)) \\ &\quad + D_{\text{KL}}(p_\theta(z_{\leq k}|z_{>k})q_\phi(z_{>k}|x) \| p_\theta(z|x)). \end{aligned} \quad (6)$$

UNCERTAINTY AND THE MEDICAL INTERVIEW

hierarchical vaes know what they don't know

Likelihood ratios

2024-03-03

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 [10].
 $\text{LLR}^{>k}(x) \equiv \mathcal{L}(x) - \mathcal{L}^{>k}(x).$
In the likelihood ratio the reconstruction terms cancel out and only the KL-divergences from the approximate to the true posterior remain.
$$\begin{aligned} \mathcal{L} &= \log p_\theta(x) - D_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x)), \\ \mathcal{L}^{>k} &= \log p_\theta(x) - D_{\text{KL}}(p_\theta(z_{\leq k}|z_{>k})q_\phi(z_{>k}|x) \| p_\theta(z|x)). \end{aligned}$$

1. Write likelihood-ratio using the exact form of the bounds including intractable KL-divergence.

The importance weighted autoencoder (IWAE) bound is tight with the true likelihood in the limit of infinite samples, $S \rightarrow \infty$ [9],

$$\mathcal{L}_S = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\log \frac{1}{N} \sum_{s=1}^S \frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})} \right] \leq \log p_\theta(\mathbf{x}), \quad (7)$$

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 $\mathcal{L}^{>k}$.

$$LLR_S^{>k}(\mathbf{x}) \rightarrow D_{KL}(p(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})q(\mathbf{z}_{>k}|\mathbf{x}) \parallel p(\mathbf{z}|\mathbf{x})). \quad (8)$$

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

UNCERTAINTY AND THE MEDICAL INTERVIEW

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

2024-03-03

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, $S \rightarrow \infty$ [9].

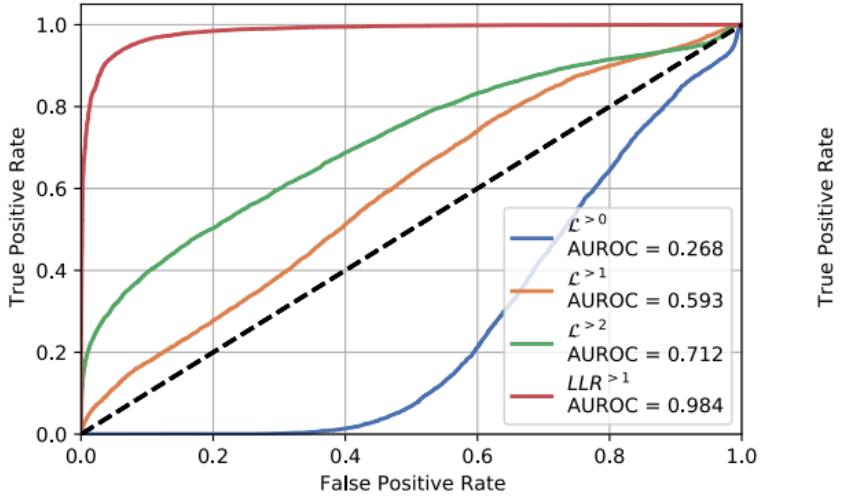
$$\mathcal{L}_S = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\log \frac{1}{N} \sum_{s=1}^S \frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})} \right] \leq \log p_\theta(\mathbf{x}), \quad (7)$$

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 $\mathcal{L}^{>k}$.

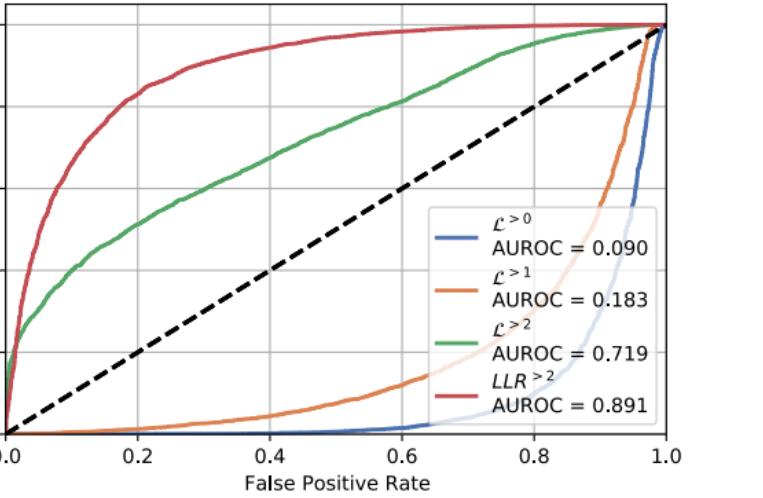
$$LLR_S^{>k}(\mathbf{x}) \rightarrow D_{KL}(p(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})q(\mathbf{z}_{>k}|\mathbf{x}) \parallel p(\mathbf{z}|\mathbf{x})). \quad (8)$$

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

HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
Results with $LLR^{>k}$



(a) FashionMNIST HVAE evaluated on MNIST



(b) CIFAR10 BIVA evaluated on SVHN

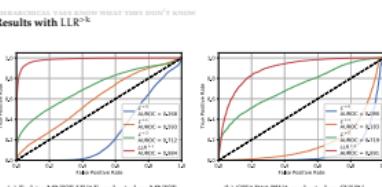


UNCERTAINTY AND THE MEDICAL INTERVIEW

hierarchical vaes know what they don't know

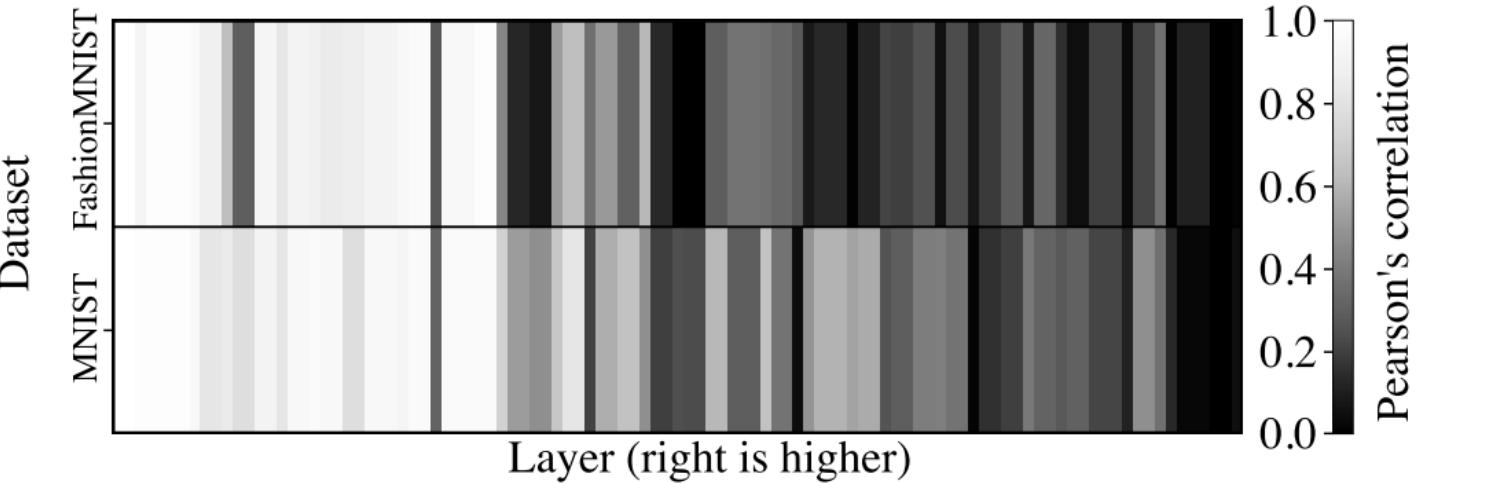
Results with $LLR^{>k}$

2024-03-03



Selecting the value of k

- Use validation OOD dataset(s).
- Compute $LLR^{>k}$ for different values of k and select the one that maximizes the AUROC.
- Compute feature correlations for different values of k and select k at the drop.

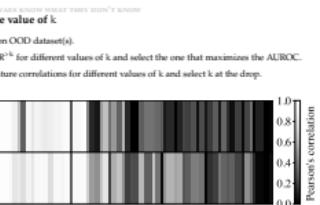


2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

hierarchical vae's know what they don't know

Selecting the value of k



HIERARCHICAL VAES KNOW WHAT THEY DON'T KNOW
Results on FashionMNIST/MNIST



Method	AUROC↑	AUPRC↑	FPR80↓
FashionMNIST (in) / MNIST (out)			
Use prior knowledge of OOD			
Backgr. contrast. LR (PixelCNN) [49]	0.994	0.993	0.001
Backgr. contrast. LR (VAE) [12]	0.924	-	-
Binary classifier [49]	0.455	0.505	0.886
$p(\hat{y} x)$ with OOD as noise class [49]	0.877	0.871	0.195
$p(\hat{y} x)$ with calibration on OOD [49]	0.904	0.895	0.139
Input complexity (S, Glow) [23]	0.998	-	-
Input complexity (S, PixelCNN++) [23]	0.967	-	-
Use in-distribution data labels y			
$p(\hat{y} x)$ [22, 49]	0.734	0.702	0.506
Entropy of $p(y x)$ [49]	0.746	0.726	0.448
ODIN [36, 49]	0.752	0.763	0.432
VIB [2, 12]	0.941	-	-
Mahalanobis distance, CNN [49]	0.942	0.928	0.088
Mahalanobis distance, DenseNet [35]	0.986	-	-
Ensemble, 20 classifiers [34, 49]	0.857	0.849	0.240
No OOD-specific assumptions			
- <i>Ensembles</i>			
WAIC, 5 models, VAE [12]	0.766	-	-
WAIC, 5 models, PixelCNN [49]	0.221	0.401	0.911
- <i>Not ensembles</i>			
Likelihood regret [59]	0.988	-	-
$\mathcal{L}^{>0}$ + HVAE (ours)	0.268	0.363	0.882
$\mathcal{L}^{>1}$ + HVAE (ours)	0.593	0.591	0.658
$\mathcal{L}^{>2}$ + HVAE (ours)	0.712	0.750	0.548
LLR $^{>1}$ + HVAE (ours)	0.964	0.961	0.036
LLR $^{>1}_{250}$ + HVAE (ours)	0.984	0.984	0.013

UNCERTAINTY AND THE MEDICAL INTERVIEW

hierarchical vaes know what they don't know
 Results on FashionMNIST/MNIST

2024-03-03

	OOD	OOD
HierarchicalVAE on MNIST/2023		
Baseline	0.993	0.993
Baseline + HVAE [12]	0.994	0.993
Binary classifier [49]	0.455	0.505
$p(\hat{y} x)$ with OOD as noise class [49]	0.877	0.871
$p(\hat{y} x)$ with calibration on OOD [49]	0.904	0.895
Input complexity (S, Glow) [23]	0.998	-
Input complexity (S, PixelCNN++) [23]	0.967	-
HierarchicalVAE on MNIST/2024		
Baseline	0.993	0.993
Baseline + HVAE [12]	0.994	0.993
Binary classifier [49]	0.455	0.505
$p(\hat{y} x)$ with OOD as noise class [49]	0.877	0.871
$p(\hat{y} x)$ with calibration on OOD [49]	0.904	0.895
Input complexity (S, Glow) [23]	0.998	-
Input complexity (S, PixelCNN++) [23]	0.967	-
Likelihood regret [59]	0.988	-
$\mathcal{L}^{>0}$ + HVAE (ours)	0.268	0.363
$\mathcal{L}^{>1}$ + HVAE (ours)	0.593	0.591
$\mathcal{L}^{>2}$ + HVAE (ours)	0.712	0.750
LLR $^{>1}$ + HVAE (ours)	0.964	0.961
LLR $^{>1}_{250}$ + HVAE (ours)	0.984	0.984

Results on CIFAR10/SVHN

Method	AUROC↑	AUPRC↑	FPR80↓
CIFAR10 (in) / SVHN (out)			
Use prior knowledge of OOD			
Backgr. contrast. LR (PixelCNN) [49]	0.930	0.881	0.066
Backgr. contrast. LR (VAE) [59]	0.265	-	-
Outlier exposure [23]	0.984	-	-
Input complexity (S, Glow) [52]	0.950	-	-
Input complexity (S, PixelCNN++) [52]	0.929	-	-
Input complexity (S, HVAE) (Ours) [52]	0.833	0.855	0.344
Use in-distribution data labels y			
Mahalanobis distance [35]	0.991	-	-
No OOD-specific assumptions			
- <i>Ensembles</i>			
WAIC, 5 models, Glow [12]	1.000	-	-
WAIC, 5 models, PixelCNN [49]	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)	0.891	0.875	0.172



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

hierarchical vaes know what they don't know

Results on CIFAR10/SVHN

Method	AUROC _{SVHN}	AUPRC _{SVHN}	FPR80 _{SVHN}
CIFAR10 (in) / SVHN (out)			
Use prior knowledge of OOD			
Backgr. contrast. LR (PixelCNN) [49]	0.930	0.881	0.066
Backgr. contrast. LR (VAE) [59]	0.265	-	-
Outlier exposure [23]	0.984	-	-
Input complexity (S, Glow) [52]	0.950	-	-
Input complexity (S, PixelCNN++) [52]	0.929	-	-
Input complexity (S, HVAE) (Ours) [52]	0.833	0.855	0.344
Use in-distribution data labels y			
Mahalanobis distance [35]	0.991	-	-
No OOD-specific assumptions			
- <i>Ensembles</i>			
WAIC, 5 models, Glow [12]	1.000	-	-
WAIC, 5 models, PixelCNN [49]	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)	0.891	0.875	0.172

- 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.
- Reconstructions of OOD data are good when using full approximate posterior.

- Proposed a likelihood-ratio score, $\text{LLR}^{>k}$, that uses the conditional prior for the top-most latent variables in the hierarchy.
- Showed that $\text{LLR}^{>k}$ outperforms the standard likelihood score across many different datasets and is competitive with state-of-the-art methods.

2024-03-03

- 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.
- Reconstructions of OOD data are good when using full approximate posterior.
- Proposed a likelihood-ratio score, $\text{LLR}^{>k}$, that uses the conditional prior for the top-most latent variables in the hierarchy.
- Showed that $\text{LLR}^{>k}$ outperforms the standard likelihood score across many different datasets and is competitive with state-of-the-art methods.

OVERVIEW Presentation

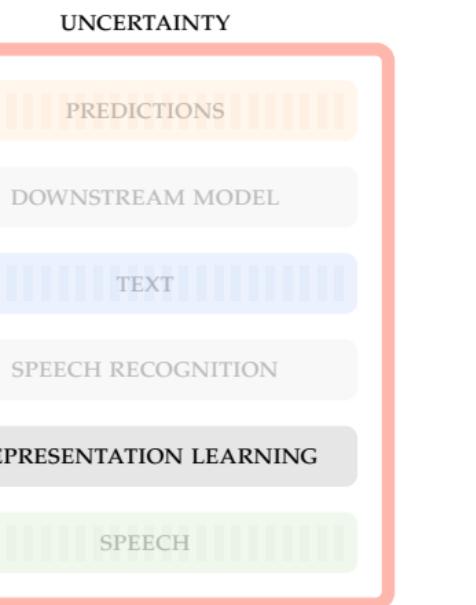
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

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



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

Overview

Presentation

1. We now move on to our review paper of speech representation learning.

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

OVERVIEW PRESENTATION

UNCERTAINTY

PREDICTIONS

HIERARCHICAL MODELS

TEXT

SPEECHrecognition

REPRESENTATION LEARNING

SPEECH

OVERVIEW Presentation

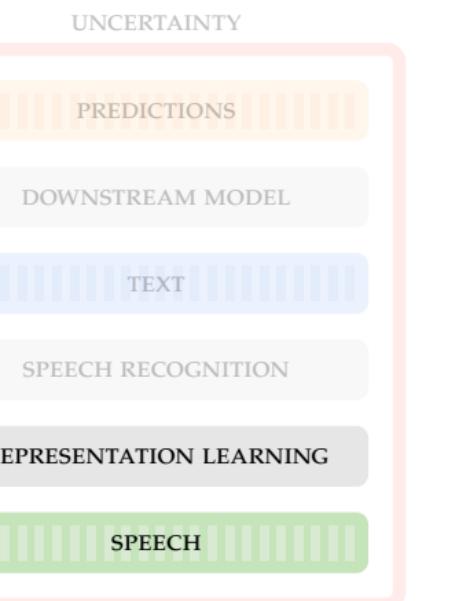
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

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

2024-03-03

Overview

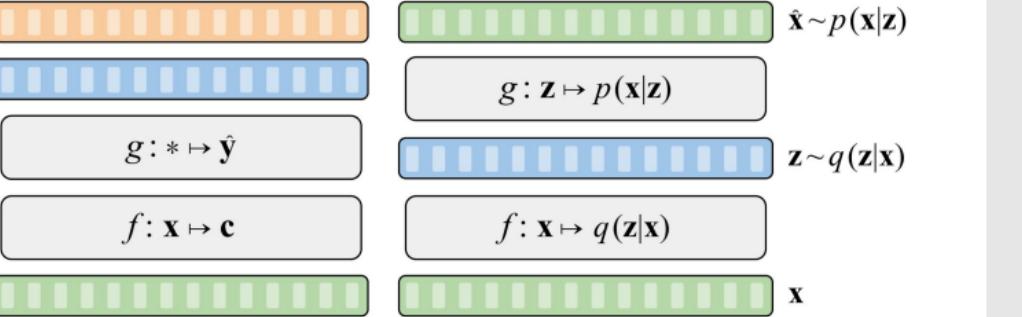
Presentation

Overview

Presentation



- Reviews 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 for selected self-supervised models.
it more focus on the self-supervised learning models.

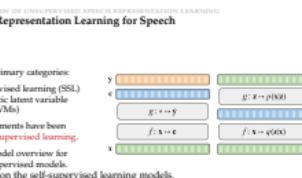


2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

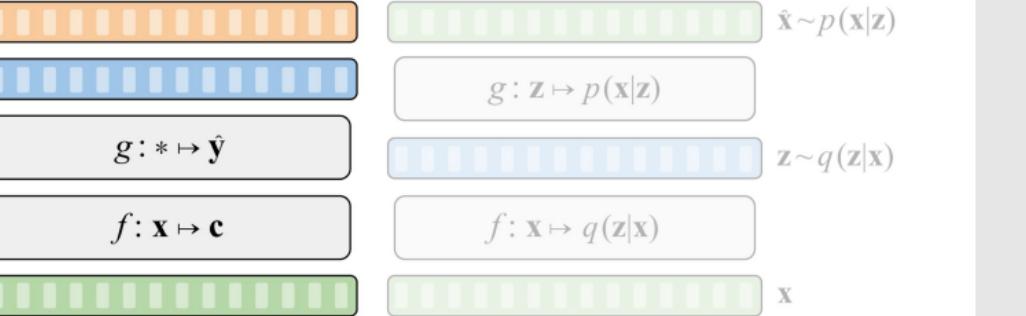
- └ a brief overview of unsupervised speech representation learning
- └ Overview: Representation Learning for Speech

1. B



Overview: Representation Learning for Speech

- Reviews 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 for selected self-supervised models.
it more focus on the self-supervised learning models.



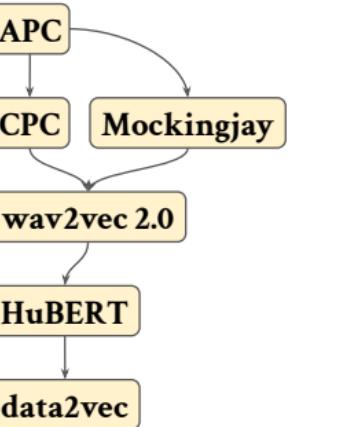
2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

- └ a brief overview of unsupervised speech representation learning
- └ Overview: Representation Learning for Speech

1. B





2024-03-03

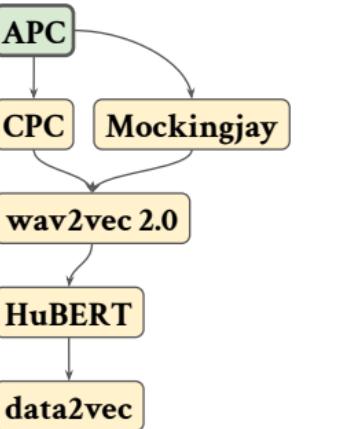
UNCERTAINTY AND THE MEDICAL INTERVIEW

- └ a brief overview of unsupervised speech representation learning
- └ Development of SSL for speech

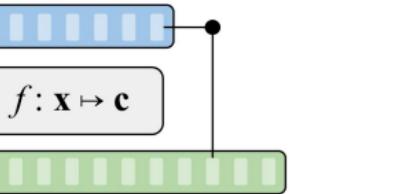
A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING
Development of SSL for speech

```
graph TD; APC[APC] --> CPC[CPC]; CPC --> Mockingjay[Mockingjay]; Mockingjay --> wav2vec2.0[wav2vec 2.0]; wav2vec2.0 --> HuBERT[HuBERT]; HuBERT --> data2vec[data2vec]
```

A small version of the flowchart from the left slide, showing the sequence from APC through data2vec.

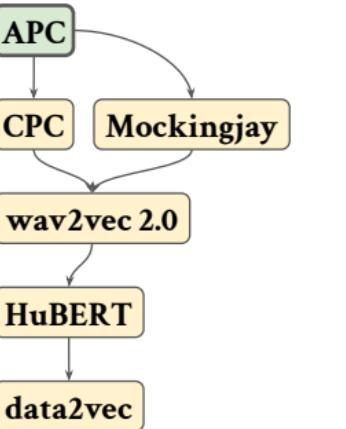


- **Task:** Predict future inputs.
- **Input/target:** Log-mel spectrogram.
- **Architecture:** RNN/Transformer decoder.
- **Slow features:** Predict k steps ahead.



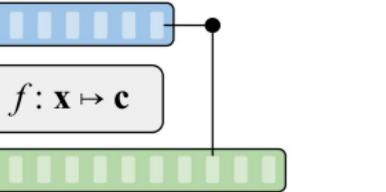
• Task: Predict future inputs.
• Input/target: Log-mel spectrogram.
• Architecture: RNN/Transformer decoder.
• Slow features: Predict k steps ahead.

$f: \mathbf{x} \mapsto \mathbf{c}$



- Challenges:

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

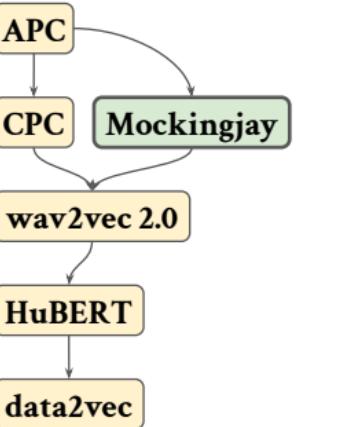


UNCERTAINTY AND THE MEDICAL INTERVIEW

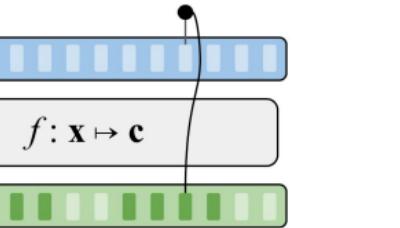
- └ a brief overview of unsupervised speech representation learning
 - └ Autoregressive Predictive Coding (APC)



1. Encoding only past inputs limits the model's ability to learn from future context.
2. Input as target limits the model's ability to learn abstract representations.

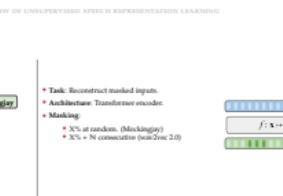


- **Task:** Reconstruct masked inputs.
- **Architecture:** Transformer encoder.
- **Masking:**
 - X% at random. (Mockingjay)
 - X% + N consecutive (wav2vec 2.0)

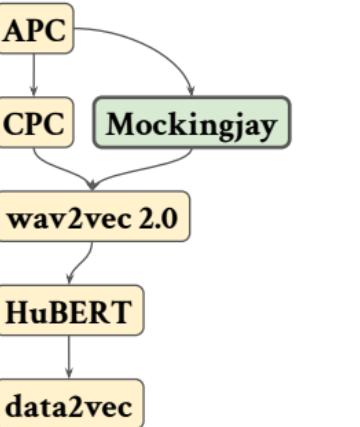


UNCERTAINTY AND THE MEDICAL INTERVIEW

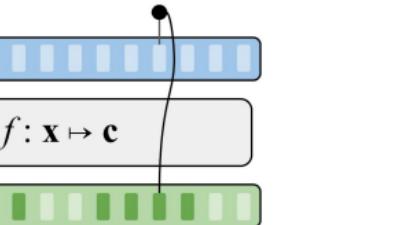
└ a brief overview of unsupervised speech representation learning
└ Mockingjay

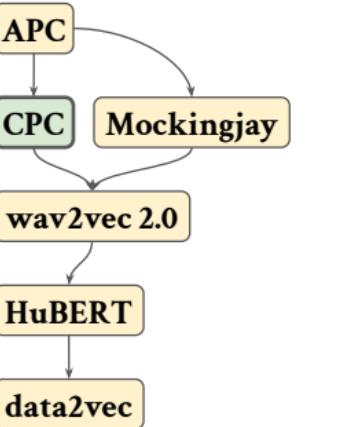


1. Replaces autoregressiveness with a masked reconstruction task.
2. Masking strategy to force learning long-term context.

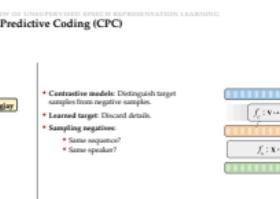
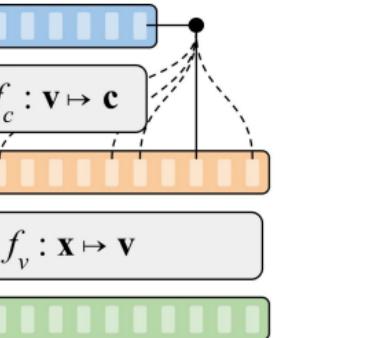


- Challenges:
 - Encodes the entire input ✓
 - Uses the input as target ✗





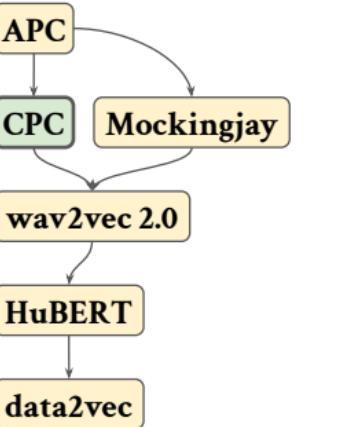
- **Contrastive models:** Distinguish target samples from negative samples.
- **Learned target:** Discard details.
- **Sampling negatives:**
 - Same sequence?
 - Same speaker?



1. Contrastive models allow using a learned target. Discard unimportant details of input.
2. Negative sampling is a challenge and requires careful design.
3. Sampling different speakers as negatives learns strong speaker-identities.

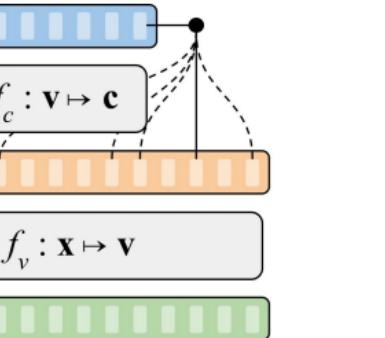
A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

Contrastive Predictive Coding (CPC)



- Challenges:

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

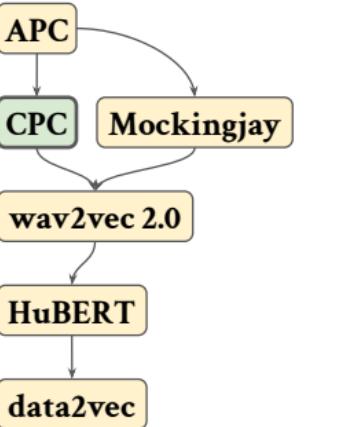


2024-03-03

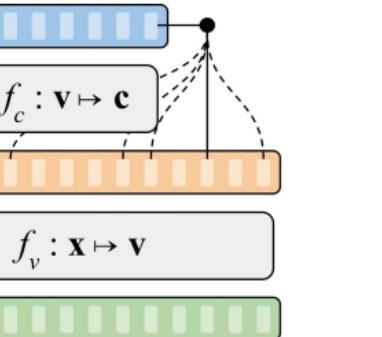
UNCERTAINTY AND THE MEDICAL INTERVIEW

- └ a brief overview of unsupervised speech representation learning
- └ Contrastive Predictive Coding (CPC)

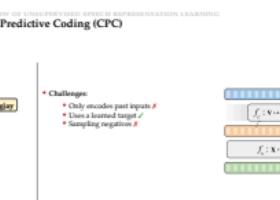


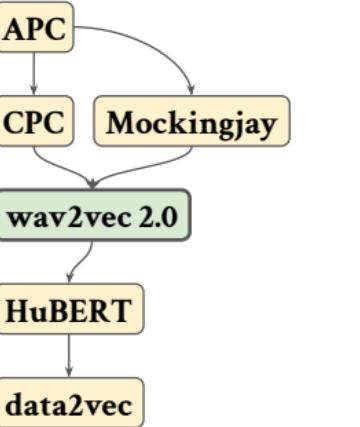


- Challenges:
- Only encodes past inputs ✗
- Uses a learned target ✓
- Sampling negatives ✗

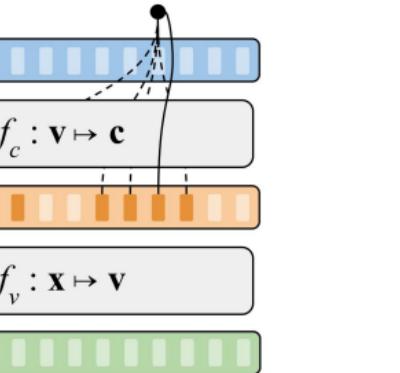


2024-03-03

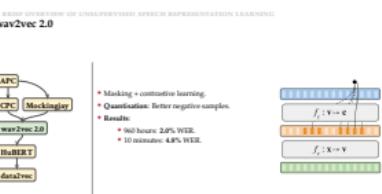




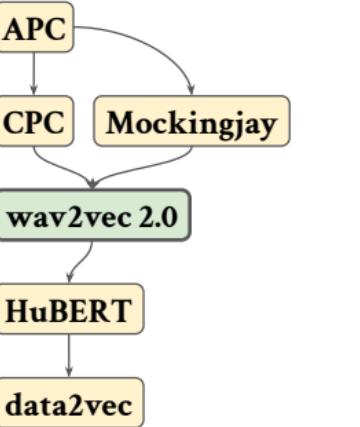
- Masking + contrastive learning.
- **Quantisation:** Better negative samples.
- **Results:**
 - 960 hours: **2.0%** WER.
 - 10 minutes: **4.8%** WER.



UNCERTAINTY AND THE MEDICAL INTERVIEW
a brief overview of unsupervised speech representation learning
2024-03-03
wav2vec 2.0

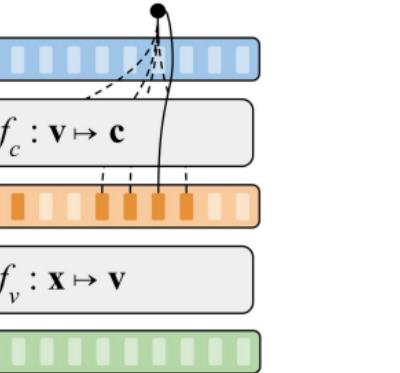


1. Training objective requires identifying the correct quantized latent audio representation in a set of distractors for each masked time step.
2. Quantisation improves negative sampling (requires approximation via Gumbel softmax).



- Challenges:

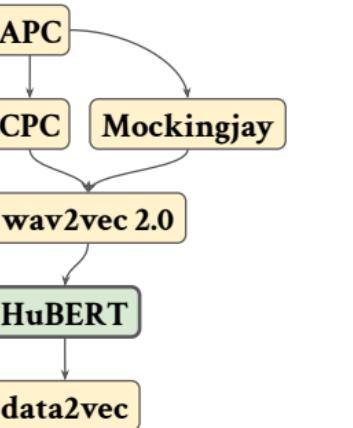
- Encodes the entire input ✓
- Uses a learned target ✓
- Sampling negatives ✗



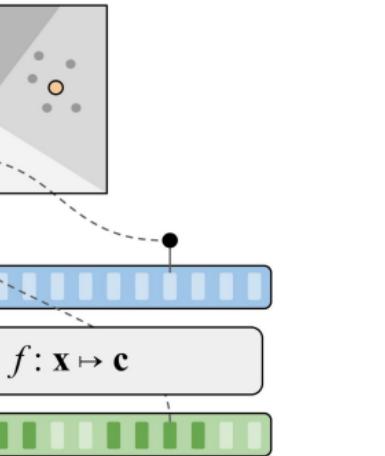
2024-03-03



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



- Target: K-means teacher.
- Training: Simple cross-entropy loss.
- 1st iteration: K-means on inputs.



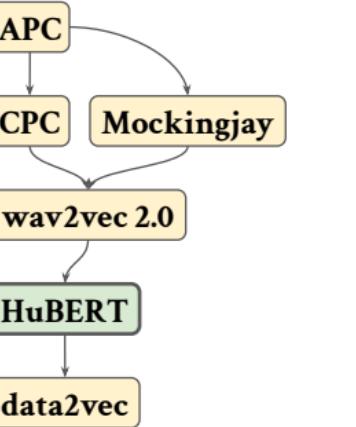
40

2024-03-03

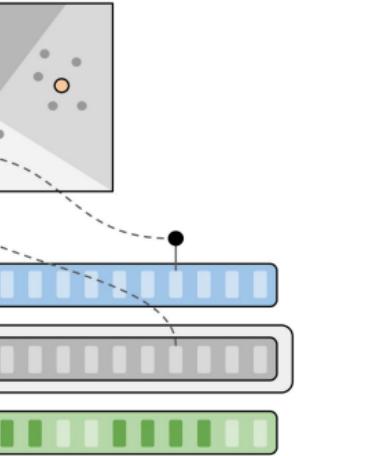
UNCERTAINTY AND THE MEDICAL INTERVIEW

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





- Target: K-means teacher.
- Training: Simple cross-entropy loss.
- 1st iteration: K-means on inputs.
- 2nd iteration: K-means on hidden layers.



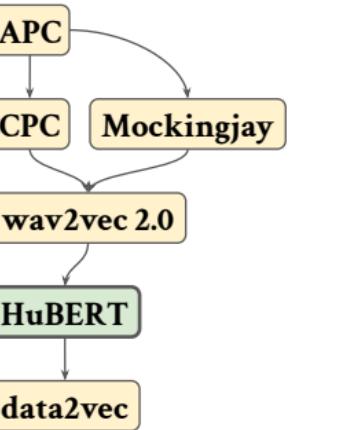
2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

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

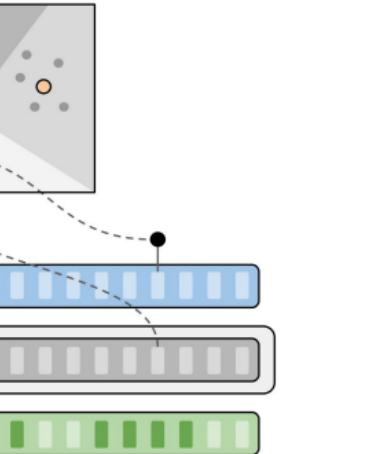
1. HuBERT approach predicts hidden cluster assignments of masked frames





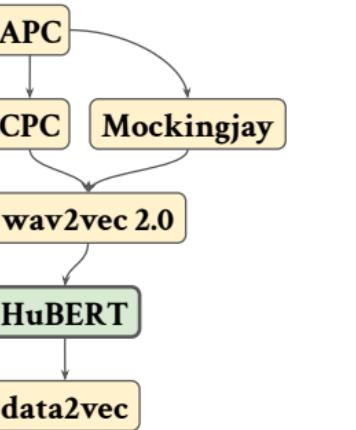
• Challenges:

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



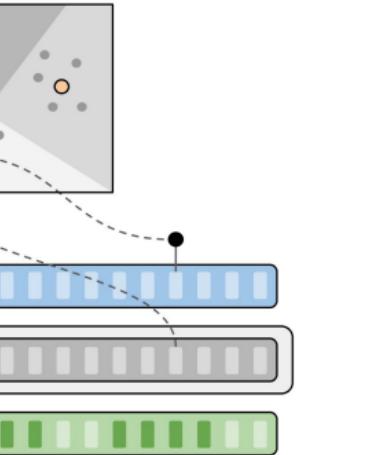
2024-03-03





• Challenges:

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

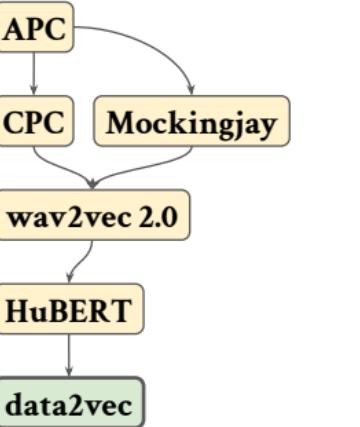


UNCERTAINTY AND THE MEDICAL INTERVIEW

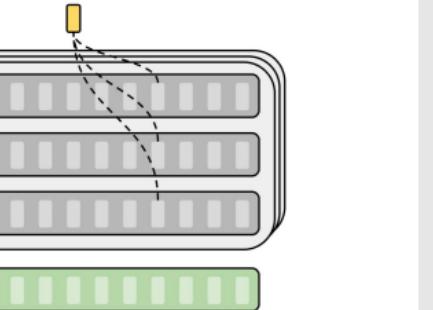
└ a brief overview of unsupervised speech representation learning
└ Hidden-unit BERT (HuBERT)



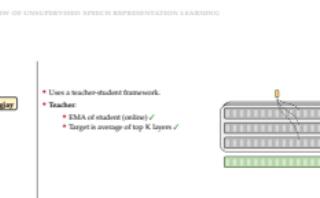
1. HuBERT approach predicts hidden cluster assignments of masked frames
2. Targets are still quantised although we no longer solve a contrastive sampling problem. Might reduce quality.

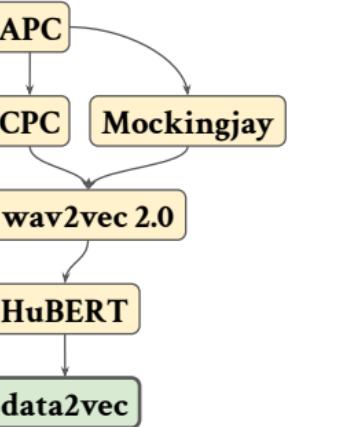


- Uses a teacher-student framework.
- Teacher:
 - EMA of student (online) ✓
 - Target is average of top K layers ✓

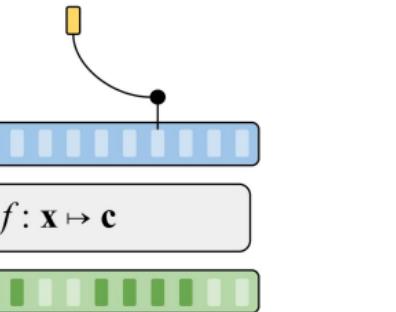


2024-03-03

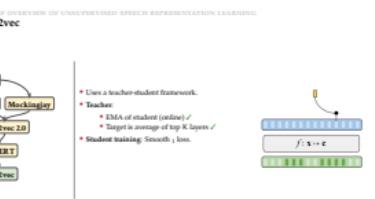


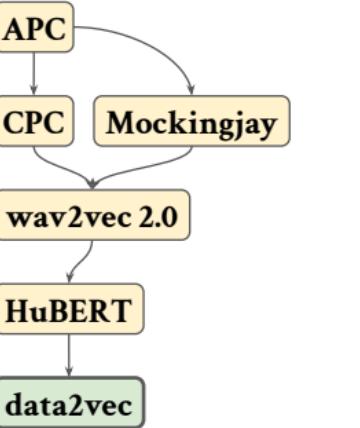


- Uses a teacher-student framework.
- **Teacher:**
 - EMA of student (online) ✓
 - Target is average of top K layers ✓
- **Student training:** Smooth $_1$ loss.



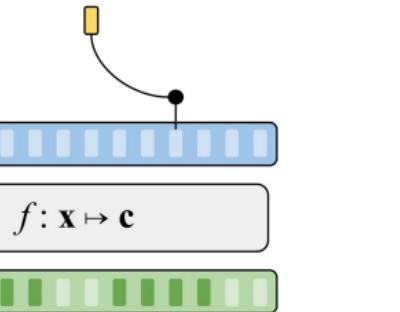
2024-03-03





• Challenges:

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



2024-03-03



- **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.

Conclusions

- **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

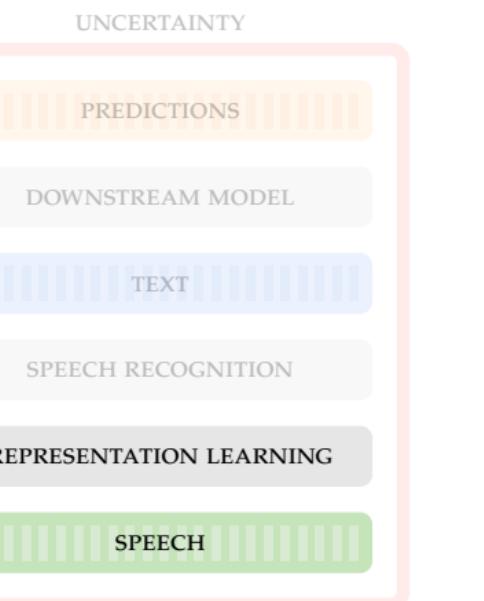
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

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



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

Overview

Presentation

OVERVIEW PRESENTATION

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

UNCERTAINTY PREDICTIONS DOWNSTREAM MODEL TEXT SPEECH RECOGNITION REPRESENTATION LEARNING SPEECH

OVERVIEW Presentation

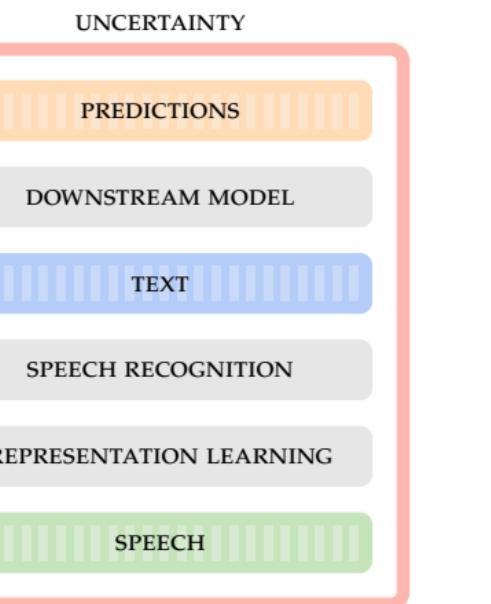
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

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

2024-03-03

Overview

Presentation

A hierarchical navigation structure for the presentation. The main title is 'UNCERTAINTY AND THE MEDICAL INTERVIEW'. Below it is a 'Overview' section, which further branches into 'Presentation'. The date '2024-03-03' is also displayed.

OVERVIEW PRESENTATION

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

UNCERTAINTY

PREDICTIONS

DOWNTSTREAM MODEL

TEXT

SPEECH RECOGNITION

REPRESENTATION LEARNING

SPEECH

- Stroke is the second leading cause of death (11.6%) and third leading cause of death and disability combined (5.7%) [19, 29, 33].
- Effective treatment is very **time-sensitive** [5, 56].
- The gateway to **ambulance transport and hospital admittance** is through **prehospital telehealth services**.
- **Mobile stroke units** has made it possible to deliver advanced treatment faster [21, 42].
- The effectiveness of mobile stroke units hinges on **call-taker recognition of stroke** [21, 42].
- Approximately half of all patients with stroke do not receive the correct triage for their condition from call-takers [8, 45, 58].

UNCERTAINTY AND THE MEDICAL INTERVIEW

- 2024-03-03
- └ 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%) [19, 29, 33].
- Effective treatment is very **time-sensitive** [5, 56].
- The gateway to **ambulance transport and hospital admittance** is through **prehospital telehealth services**.
- **Mobile stroke units** has made it possible to deliver advanced treatment faster [21, 42].
- The effectiveness of mobile stroke units hinges on **call-taker recognition of stroke** [21, 42].
- Approximately half of all patients with stroke do not receive the correct triage for their condition from call-takers [8, 45, 58].

The study

- Collaboration between **Corti** and the **Copenhagen Emergency Medical Services (CEMS)** ("Region Hovedstadens Akutberedskab").
- CEMS provides prehospital telehealth 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 general practitioner is unavailable).
- We wanted to investigate if a machine learning model could assist call-takers of 1813 in recognizing stroke.



UNCERTAINTY AND THE MEDICAL INTERVIEW

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

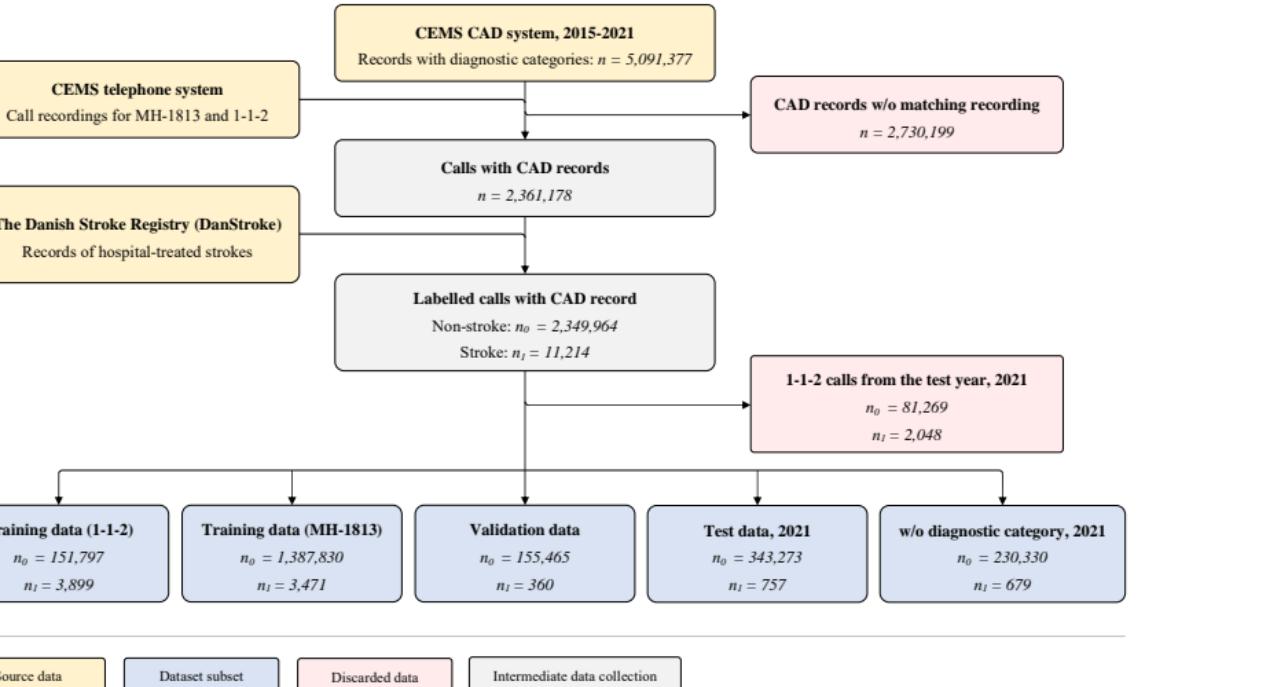
└ The study

2024-03-03

- Collaboration between **Corti** and the **Copenhagen Emergency Medical Services (CEMS)** ("Region Hovedstadens Akutberedskab").
- CEMS provides prehospital telehealth 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 general practitioner is unavailable).
- We wanted to investigate if a machine learning model could assist call-takers of 1813 in recognizing stroke.

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

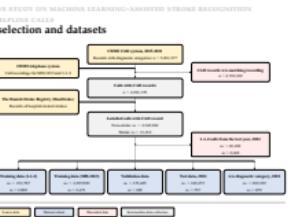
Population selection and datasets



UNCERTAINTY AND THE MEDICAL INTERVIEW

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

Population selection and datasets



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

Population characteristics



	Training (112)	Training (MH-1813)	Validation	Test	2021 w/o category
All calls	Num. calls 155,696	1,391,301	155,825	344,030	231,009
	Female 74,640 (47.94%)	792,783 (56.98%)	86,959 (55.81%)	190,974 (55.51%)	134,324 (58.14%)
	Male 79,564 (51.10%)	596,760 (42.89%)	68,866 (44.19%)	153,050 (44.49%)	96,258 (41.67%)
	65+ years 72,930 (46.84%)	335,146 (24.09%)	30,313 (19.45%)	65,652 (19.08%)	81,488 (35.27%)
	Age (mean ± std.) 59.47 ± 21.24	47.12 ± 21.38	44.63 ± 20.08	44.31 ± 20.10	50.36 ± 22.77
Stroke calls	Num. calls 3,899	3,471	360	757	679
	Female 1,784 (45.76%)	1,654 (47.65%)	161 (44.72%)	349 (46.10%)	366 (53.90%)
	Male 2,115 (54.24%)	1,815 (52.29%)	199 (55.28%)	408 (53.90%)	313 (46.10%)
	65+ years 2,968 (76.12%)	2,421 (69.75%)	250 (69.44%)	555 (73.32%)	567 (83.51%)
	Age (mean ± std.) 72.91 ± 12.77	70.68 ± 13.85	70.93 ± 13.83	71.51 ± 13.41	73.41 ± 14.11
Non-stroke	Num. calls 151,797	1,387,830	155,465	343,273	230,330
	Female 72,856 (48.00%)	791,129 (57.00%)	86,798 (55.83%)	190,625 (55.53%)	133,958 (58.16%)
	Male 77,449 (51.02%)	594,945 (42.87%)	68,667 (44.17%)	152,642 (44.47%)	95,945 (41.66%)
	65+ years 69,962 (46.09%)	332,725 (23.97%)	30,063 (19.34%)	65,097 (18.96%)	80,921 (35.13%)
	Age (mean ± std.) 59.12 ± 21.30	47.06 ± 21.36	44.57 ± 20.05	44.25 ± 20.08	50.29 ± 22.76

UNCERTAINTY AND THE MEDICAL INTERVIEW

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

Population characteristics

2024-03-03

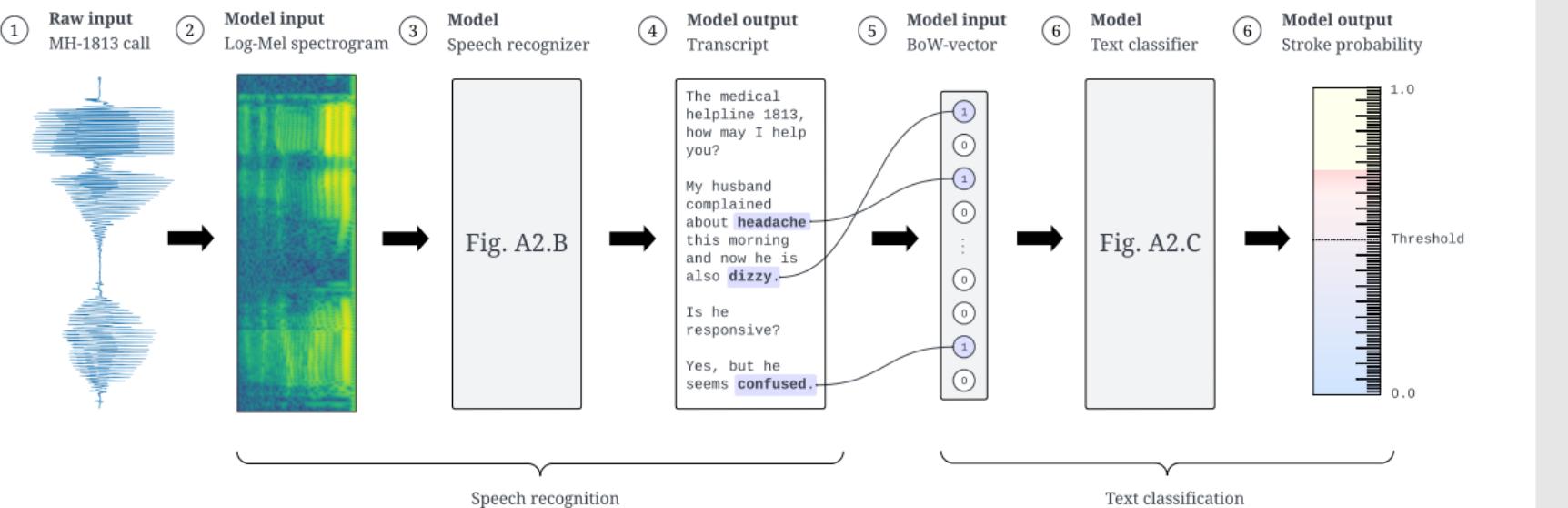
	Training (112)	Training (MH-1813)	Validation	Test	2021 w/o category
Population characteristics					
Num. calls	3,899	3,471	360	757	679
Female	1,784 (45.76%)	1,654 (47.65%)	161 (44.72%)	349 (46.10%)	366 (53.90%)
Male	2,115 (54.24%)	1,815 (52.29%)	199 (55.28%)	408 (53.90%)	313 (46.10%)
65+ years	2,968 (76.12%)	2,421 (69.75%)	250 (69.44%)	555 (73.32%)	567 (83.51%)
Age (mean ± std.)	72.91 ± 12.77	70.68 ± 13.85	70.93 ± 13.83	71.51 ± 13.41	73.41 ± 14.11
Num. calls	151,797	1,387,830	155,465	343,273	230,330
Female	72,856 (48.00%)	791,129 (57.00%)	86,798 (55.83%)	190,625 (55.53%)	133,958 (58.16%)
Male	77,449 (51.02%)	594,945 (42.87%)	68,667 (44.17%)	152,642 (44.47%)	95,945 (41.66%)
65+ years	69,962 (46.09%)	332,725 (23.97%)	30,063 (19.34%)	65,097 (18.96%)	80,921 (35.13%)
Age (mean ± std.)	59.12 ± 21.30	47.06 ± 21.36	44.57 ± 20.05	44.25 ± 20.08	50.29 ± 22.76

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

Model design



A. Schematic Overview of Stroke Classification Pipeline

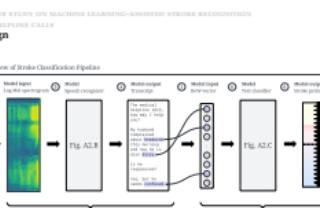


UNCERTAINTY AND THE MEDICAL INTERVIEW

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

Model design

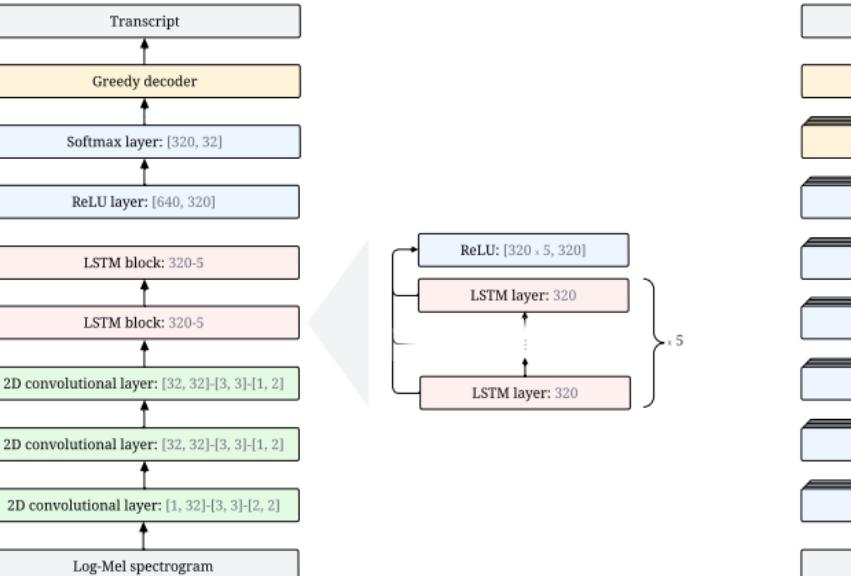
2024-03-03



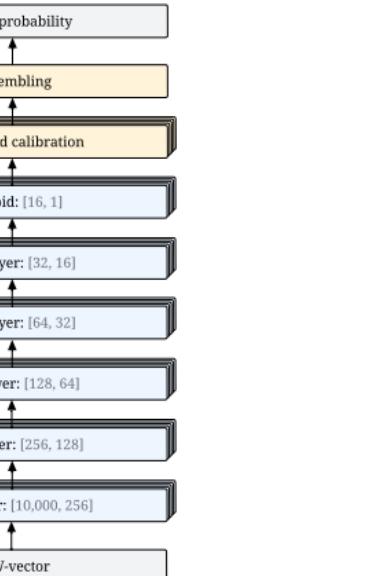
A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS

Model design

B. Speech Recognition Model



C. Text Classification Model

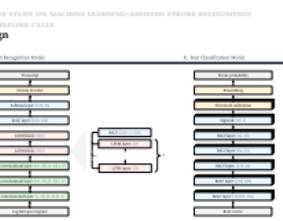


2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

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

Model design



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	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)
	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)	0.419 (0.413-0.426)
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)	0.353 (0.331-0.375)
	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)	0.231 (0.226-0.236)
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)	1.467 (1.373-1.560)
	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)	1.224 (1.198-1.249)
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)	0.542 (0.506-0.580)
	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)	0.435 (0.425-0.445)
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)	0.582 (0.547-0.616)
	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)	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

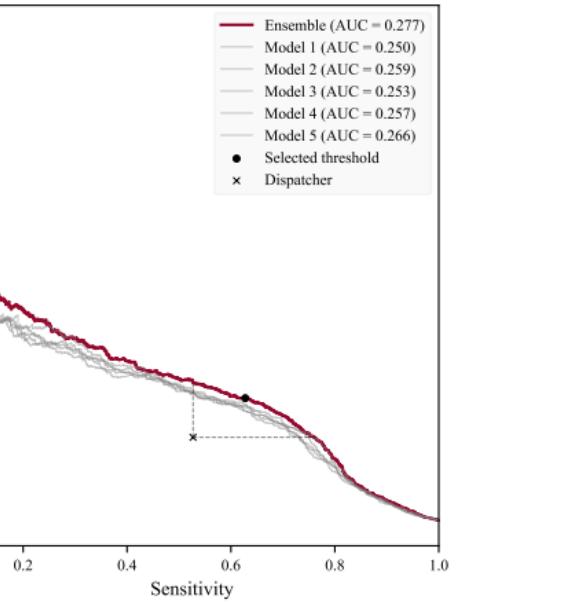
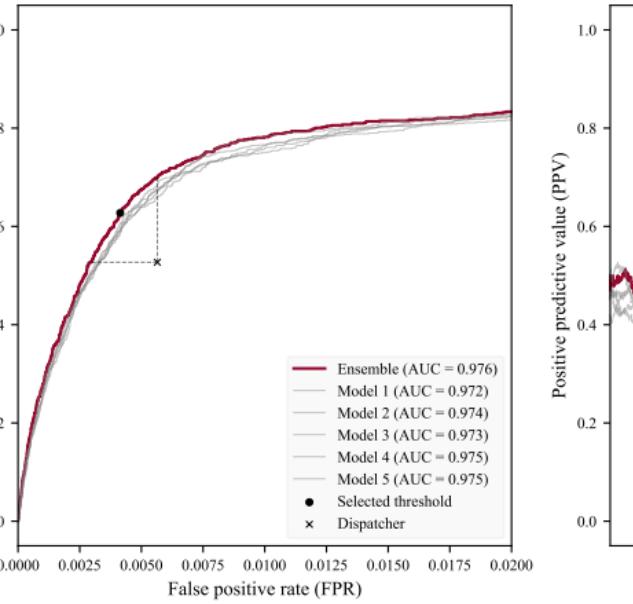
2024-03-03

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	52.7	17.1	0.105
Overall	Model	35.7	63.0	24.9	0.082
18-64 years	Call-takers	15.9	50.5	9.40	0.036
18-64 years	Model	22.9	54.1	14.5	0.033
65+ years	Call-takers	32.9	53.5	23.0	0.401
65+ years	Model	42.8	66.3	31.6	0.290
Male	Call-takers	30.2	53.9	21.0	0.124
Male	Model	39.0	63.7	28.1	0.097
Female	Call-takers	21.9	51.3	13.9	0.090
Female	Model	32.4	62.3	21.9	0.069

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.

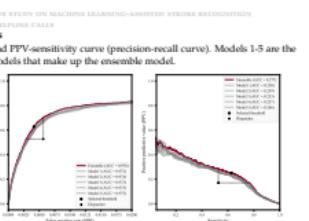


2024-03-03

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

Which features are important?

Let $z^{(n,d,w)}$ be the logit output of model n in the ensemble for transcript d when the word w is occluded. For transcript d , we computed the word impact score $i^{(d,w)}$ as the mean difference between the logit before and after occlusion.

$$i^{(d,w)} = \frac{1}{N_d} \sum_{n=1}^{N_d} (z^{(n,d)} - z^{(n,d,w)}) . \quad (9)$$

To select words for inspection, we computed a word-rank score, $r^{(w)}$, as the sum of the signed squares of the impact:

$$r^{(w)} = \sum_{d=1}^N \text{sign}(i^{(d,w)}) (i^{(d,w)})^2 . \quad (10)$$

Squaring $i^{(d,w)}$ favors rare features with a high impact over common features with a low impact.

UNCERTAINTY AND THE MEDICAL INTERVIEW

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

Which features are important?

2024-03-03

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

Which features are important?

Let $z^{(n,d,w)}$ be the logit output of model n in the ensemble for transcript d when the word w is occluded. For transcript d , we computed the word impact score $i^{(d,w)}$ as the mean difference between the logit before and after occlusion.

$$i^{(d,w)} = \frac{1}{N_d} \sum_{n=1}^{N_d} (z^{(n,d)} - z^{(n,d,w)}) . \quad (9)$$

To select words for inspection, we computed a word-rank score, $r^{(w)}$, as the sum of the signed squares of the impact:

$$r^{(w)} = \sum_{d=1}^N \text{sign}(i^{(d,w)}) (i^{(d,w)})^2 . \quad (10)$$

Squaring $i^{(d,w)}$ favors rare features with a high impact over common features with a low impact.

Which features are important?

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

Which features are important?

2024-03-03

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.	The ambulance	521
4.	Right	1,050	19.	Hand	297
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]

Which features are important?



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

Which features are important?

2024-03-03

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS			
Which features are important?			
<small>Features with positive ranking score ($r^{(w)} > 0$) computed on stroke positive predictions ($D = 1,897$)</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]

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]

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

Which features are important?



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

↳ Which features are important?

2024-03-03

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS			
Which features are important?			
Features with positive ranking score ($r^{(w)} > 0$) computed on stroke positive predictions ($D = 1,897$)			
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.
Not making sense			

[Recognition, Symptom, Urgency/Time]

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

Which features are important?



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

└ Which features are important?

2024-03-03

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS			
Which features are important?			
<small>Features with positive ranking score ($r^{(w)} > 0$) computed on stroke positive predictions ($D = 1,897$)</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.
<small>[Recognition, Symptom, Urgency/Time]</small>			

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS			
Which features are important?			
<small>Features with positive ranking score ($r^{(w)} > 0$) computed on stroke positive predictions ($D = 1,897$)</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.
<small>[Recognition, Symptom, Urgency/Time]</small>			

[Recognition, Symptom, Urgency/Time]

58

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

Which features are important?



Features with negative ranking score ($r^{(w)} < 0$) computed on stroke-negative predictions ($D = 342,133$)					
Rank	Word, w (translated)	Occurrences, D ^(w)	Rank	Word, w (translated)	Occurrences, D ^(w)
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

Which features are important?

2024-03-03

Features with negative ranking score ($r^{(w)} < 0$) computed on stroke-negative predictions ($D = 342,133$)		
Rank	Word, w (translated)	Occurrences, D ^(w)
1.	Tetanus	8,078
2.	Pregnant	8,749
3.	Cut	7,592
4.	Bandage	4,561
5.	Over the counter (OTC)	23,776
6.	O'clock	94,436
7.	The emergency room	42,809
8.	The ribs	3,928
9.	Swollen	60,559
10.	Over the counter (OTC)	4,641
11.	The neck	30,151
12.	Fever	112,586
13.	Prescription	5,450
14.	Centimeter	12,026
15.	The knee	8,875

[Recognition, Symptom, Urgency/Time]

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

Which features are important?



Features with negative ranking score ($r^{(w)} < 0$) computed on stroke-negative predictions ($D = 342,133$)					
Rank	Word, w (translated)	Occurrences, D ^(w)	Rank	Word, w (translated)	Occurrences, D ^(w)
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

Which features are important?

2024-03-03

A RETROSPECTIVE STUDY ON MACHINE LEARNING-ASSISTED STROKE RECOGNITION FOR MEDICAL HELPLINE CALLS			
Which features are important?			
<small>Features with negative ranking score ($r^{(w)} < 0$) computed on stroke-negative predictions ($D = 342,133$)</small>			
Rank	Word, w (translated)	Occurrences, D ^(w)	Rank
1.	Tetanus	8,078	16.
2.	Pregnant	8,749	17.
3.	Cut	7,592	18.
4.	Bandage	4,561	19.
5.	Amager (a location)	23,776	20.
6.	O'clock	94,436	21.
7.	The emergency room	42,809	22.
8.	The ribs	3,928	23.
9.	Bleed	10,501	24.
10.	Over the counter (OTC)	4,641	25.
11.	The neck	30,151	26.
12.	Fever	112,586	27.
13.	Prescription	5,450	28.
14.	Centimeter	12,026	29.
15.	The knee	8,875	30.

[Recognition, Symptom, Urgency/Time]

Which features are important?

Features with negative ranking score ($r^{(w)} < 0$) computed on stroke-negative predictions (D = 342,133)					
Rank	Word, w (translated)	Occurrences, D ^(w)	Rank	Word, w (translated)	Occurrences, D ^(w)
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

Which features are important?

2024-03-03

Features with negative ranking score ($r^{(w)} < 0$) computed on stroke-negative predictions (D = 342,133)		
Rank	Word, w (translated)	Occurrences, D ^(w)
1.	Tetanus	8,079
2.	Pregnant	8,749
3.	Cut	7,592
4.	Bandage	4,561
5.	Amager (a location)	23,776
6.	O'clock	94,436
7.	The emergency room	42,809
8.	The ribs	3,928
9.	Swollen	60,559
10.	Over the counter (OTC)	4,641
11.	The neck	30,151
12.	Fever	112,586
13.	Prescription	5,450
14.	Centimeter	12,026
15.	The knee	8,875
16.	The pharmacy	42,105
17.	The stomach	3,688
18.	Psychiatric	7,597
19.	Pneumonia	10,551
20.	Stomach pain	19,155
21.	Stool	24,313
22.	The ribs	2,941
23.	Bleed	19,415
24.	Bleeding	8,127
25.	Ribs	78,558
26.	Broken	12,026
27.	Inflammation	17,762
28.	Common cold	76,948
29.	Morning or tomorrow	42,105
30.	Swelling	27,742

[Recognition, Symptom, Urgency/Time]

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

Which features are important?



Features with negative ranking score ($r^{(w)} < 0$) computed on stroke-negative predictions ($D = 342,133$)					
Rank	Word, w (translated)	Occurrences, D ^(w)	Rank	Word, w (translated)	Occurrences, D ^(w)
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

Which features are important?

2024-03-03

Features with negative ranking score ($r^{(w)} < 0$) computed on stroke-negative predictions ($D = 342,133$)					
Rank	Word, w (translated)	Occurrences, D ^(w)	Rank	Word, w (translated)	Occurrences, D ^(w)
1.	Stroke	8,079	16.	The pharmacy	42,105
2.	Pregnant	8,749	17.	The stomach	42,105
3.	Bandage	4,561	18.	Psychiatric	3,688
4.	Amager (a location)	23,776	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.	Stomach pain	10,551
9.	Bleeding	24,313	24.	Stool	19,155
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]

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 [7].

UNCERTAINTY AND THE MEDICAL INTERVIEW

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

Simulated prospective study

2024-03-03

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 [7].

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)	

UNCERTAINTY AND THE MEDICAL INTERVIEW

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

Simulated prospective study

2024-03-03

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	After call	During call	After call	During call	After 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, 47].
- 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)

UNCERTAINTY AND THE MEDICAL INTERVIEW

2024-03-03

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, 47].
- Might BERT be useful for recognizing stroke?

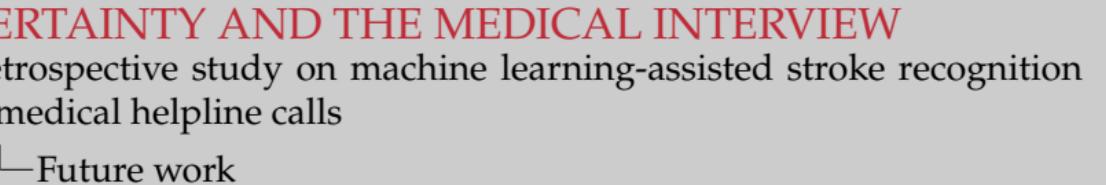
Subset	Predictor	F1-score [%] ↑	Sensitivity [%] ↑	PPV [%] ↑	FOR [%] ↓ (1 - NPV)	FPR [%] ↓ (1 - specificity)
Overall	Call-takers	25.8 (23.7-27.9)	52.7 (49.2-56.4)	17.1 (15.5-18.6)	0.105 (0.094-0.116)	0.565 (0.539-0.590)
	MLP	35.7 (35.0-36.4)	63.0 (62.0-64.1)	24.9 (24.3-25.5)	0.082 (0.079-0.085)	0.419 (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)

Future work

- Machine learning
 - Learning to predict directly from audio data (SSL).
 - Investigate learning to defer to predict methods [57].



2024-03-03



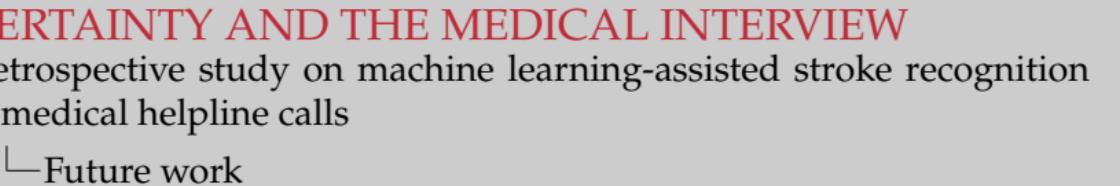
- Machine learning
 - Learning to predict directly from audio data (SSL).
 - Investigate learning to defer to predict methods [57].

Future work

- Machine learning
 - Learning to predict directly from audio data (SSL).
 - Investigate 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.



2024-03-03



- Machine learning
 - Learning to predict directly from audio data (SSL).
 - Investigate 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

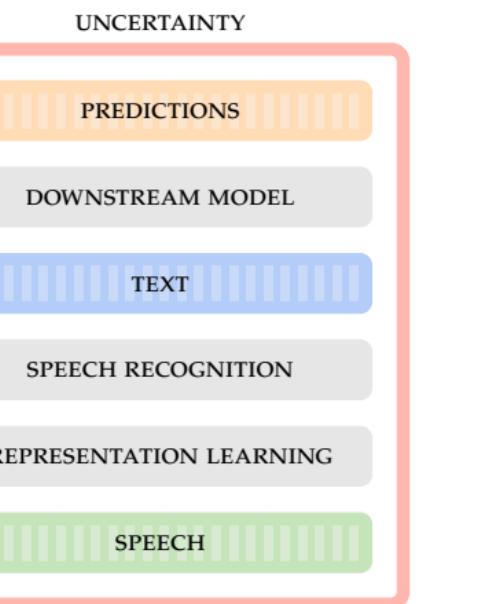
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

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



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

Overview

Presentation

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

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

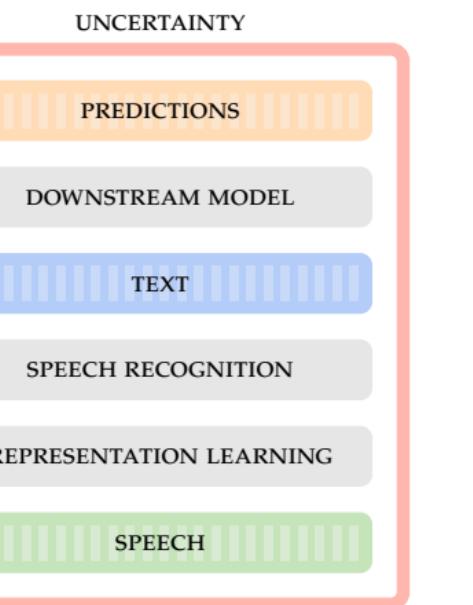
CHAPTER 1-3 INTRODUCTION, RESEARCH QUESTIONS, AND BACKGROUND

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



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

Overview

Presentation

UNCERTAINTY

- PREDICTIONS
- DOWNSTREAM MODEL
- TEXT
- SPEECH RECOGNITION
- REPRESENTATION LEARNING
- SPEECH

The broad picture: The field of AI since 2020



2020 Project start

- Out-of-distribution detection with generative models: Mysterious new topic.
- Speech recognition: Inflection point between supervised methods and dawning self-supervised approaches.
- Large language models such as BERT and GPT were becoming popular for fine-tuning on specific tasks.

└ discussion

└ The broad picture: The field of AI since 2020

2024-03-03

- Out-of-distribution detection with generative models: Mysterious new topic.
- Speech recognition: Inflection point between supervised methods and dawning self-supervised approaches.
- Large language models such as BERT and GPT were becoming popular for fine-tuning on specific tasks.

DISCUSSION

The broad picture: The field of AI since 2020



2020 Project start

- Out-of-distribution detection with generative models: Mysterious new topic.
- Speech recognition: Inflection point between supervised methods and dawning self-supervised approaches.
- Large language models such as BERT and GPT were becoming popular for fine-tuning on specific tasks.

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.
- Large language models are a commodity and the paradigm of fine-tuning for instruction (RLFH/DPO [46, 48]) has fundamentally changed how they are used.

UNCERTAINTY AND THE MEDICAL INTERVIEW

↳ discussion

2024-03-03

↳ The broad picture: The field of AI since 2020

- Out-of-distribution detection with generative models: Mysterious new topic.
- Speech recognition: Inflection point between supervised methods and dawning self-supervised approaches.
- Large language models such as BERT and GPT were becoming popular for fine-tuning on specific tasks.

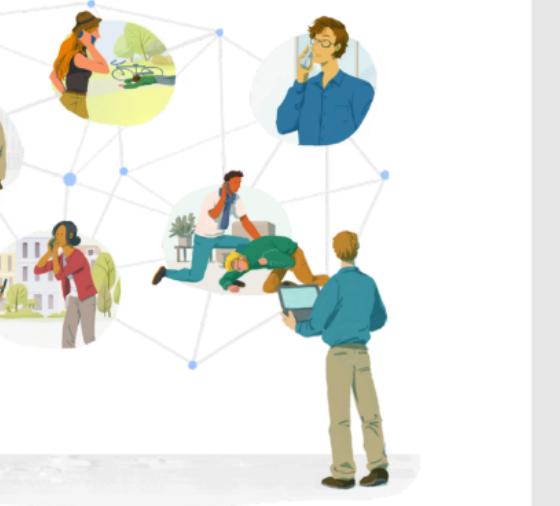
- Out-of-distribution detection is a mature field with a wide range of methods.
- Self-supervised learning is the dominant paradigm in speech recognition.
- Large language models are a commodity and the paradigm of fine-tuning for instruction (RLFH/DPO [46, 48]) has fundamentally changed how they are used.

DISCUSSION

The role of uncertainty in an operational decision support system

- Do we need true uncertainty estimates? Pragmatism versus idealism.
- Role of explainability compared to uncertainty estimates.
- European Parliamentary Research Services [17]:

"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"



UNCERTAINTY AND THE MEDICAL INTERVIEW

discussion

The role of uncertainty in an operational decision support system

2024-03-03

- Do we need true uncertainty estimates? Pragmatism versus idealism.
- Role of explainability compared to uncertainty estimates.
- European Parliamentary Research Services [17]:
 - "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"



What lies ahead

Self-supervised learning in the wild:

- A single emergency service can store hundreds of thousands of calls with rich metadata every year. Does the recent progress on academic datasets translate to this real-world setting?



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

└ discussion

└ What lies ahead

DISCUSSION

What lies ahead

Self-supervised learning in the wild:
• A single emergency service can store hundreds of thousands of calls with rich metadata every year. Does the recent progress on academic datasets translate to this real-world setting?

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 116, 117).
- [2] Alemi, A. A., Fischer, I., Dillon, J. V., *Uncertainty in the Variational Information Bottleneck*. 2018. arXiv: 1807.00906 (cited on page 44).
- [3] 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 117).
- [4] 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 117).
- [5] 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 71).
- [6] 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 34, 35).
- [7] 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 89).



2024-03-03

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 116, 117).
- [2] Alemi, A. A., Fischer, I., Dillon, J. V., *Uncertainty in the Variational Information Bottleneck*. 2018. arXiv: 1807.00906 (cited on page 44).
- [3] 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 117).
- [4] 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 117).
- [5] 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 71).
- [6] 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 34, 35).
- [7] 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 89).

Bibliography II

- [8] 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 71).
- [9] 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 41).
- [10] Buse, A. "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note". In: *The American Statistician* 36 (3a 1982) (cited on page 40).
- [11] Carver, N., Gupta, V., Hipskind, J. E., "Medical Errors". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024. pmid: 28613514 (cited on pages 8, 9).
- [12] Choi, H., Jang, E., Alemi, A. A., WAIC, but Why? *Generative Ensembles for Robust Anomaly Detection*. 2019. arXiv: 1810.01392 (cited on pages 44, 45).
- [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 117).
- [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 116, 117).
- [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 91).



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

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

- BIBLIOGRAPHY
Bibliography II
- [8] 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 71).
 - [9] 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 41).
 - [10] Buse, A. "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note". In: *The American Statistician* 36 (3a 1982) (cited on page 40).
 - [11] Carver, N., Gupta, V., Hipskind, J. E., "Medical Errors". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024. pmid: 28613514 (cited on pages 8, 9).
 - [12] Choi, H., Jang, E., Alemi, A. A., WAIC, but Why? *Generative Ensembles for Robust Anomaly Detection*. 2019. arXiv: 1810.01392 (cited on pages 44, 45).
 - [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 117).
 - [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 116, 117).
 - [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 91).

Bibliography III

- [16] 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 116).
- [17] 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 98).
- [18] 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 116, 117).
- [19] 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 71).
- [20] 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 116).
- [21] 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 71).



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

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

- BIBLIOGRAPHY
Bibliography III
- [14] 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 116).
 - [15] 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 98).
 - [16] 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 116, 117).
 - [17] 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 71).
 - [18] 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 116).
 - [19] 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 71).

Bibliography IV



- [22] Hendrycks, D., Gimpel, K., "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks". In: *Proceedings of the 5th International Conference on Learning Representations (ICRL)*. Toulon, France, 2017 (cited on page 44).
- [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 pages 44, 45).
- [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 10, 11).
- [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 117).
- [26] Hsu, W.-N., Zhang, Y., Glass, J., *Learning Latent Representations for Speech Generation and Transformation*. 2017. arXiv: 1704.04222 (cited on pages 116, 117).
- [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 116, 117).

bibliography

Bibliography

2024-03-03

- BIBLIOGRAPHY**
- Bibliography IV**
- [22] Hendrycks, D., Gimpel, K., "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks". In: *Proceedings of the 5th International Conference on Learning Representations (ICRL)*. Toulon, France, 2017 (cited on page 44).
 - [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 pages 44, 45).
 - [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 10, 11).
 - [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 117).
 - [26] Hsu, W.-N., Zhang, Y., Glass, J., *Learning Latent Representations for Speech Generation and Transformation*. 2017. arXiv: 1704.04222 (cited on pages 116, 117).
 - [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 116, 117).

Bibliography V



- [28] 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 10, 11).
- [29] Katan, M., Luft, A., "Global Burden of Stroke". In: *Seminars in Neurology*. Volume 38. 02. Thieme Medical Publishers, 2018 (cited on page 71).
- [30] 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 116, 117).
- [31] 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 116, 117).
- [32] 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 page 36).
- [33] 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 71).

bibliography

Bibliography

2024-03-03

- [28] 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 10, 11).
- [29] Katan, M., Luft, A., "Global Burden of Stroke". In: *Seminars in Neurology*. Volume 38. 02. Thieme Medical Publishers, 2018 (cited on page 71).
- [30] 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 116, 117).
- [31] 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 116, 117).
- [32] 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 page 36).
- [33] 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 71).

Bibliography VI



- [34] Lakshminarayanan, B., Pritzel, A., Blundell, C., "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles". In: *In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA, 2017 (cited on page 44).
- [35] 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, Quebec, Canada, 2018 (cited on pages 44, 45).
- [36] Liang, S., Li, Y., Srikant, R., "Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks". In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. Vancouver, Canada, 2018 (cited on page 44).
- [37] Ling, S., Liu, Y., "DeCoAR 2.0: Deep Contextualized Acoustic Representations with Vector Quantization". 2020. arXiv: 2012.06659 (cited on page 117).
- [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 117).
- [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 117).

bibliography

Bibliography

2024-03-03

- BIBLIOGRAPHY**
- Bibliography VI**
- [34] Lakshminarayanan, B., Pritzel, A., Blundell, C., "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles". In: *In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA, 2017 (cited on page 44).
 - [35] 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, Quebec, Canada, 2018 (cited on pages 44, 45).
 - [36] Ling, S., Li, Y., Srikant, R., "Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks". In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. Vancouver, Canada, 2018 (cited on page 44).
 - [37] Ling, S., Liu, Y., "DeCoAR 2.0: Deep Contextualized Acoustic Representations with Vector Quantization". 2020. arXiv: 2012.06659 (cited on page 117).
 - [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 117).
 - [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 117).

Bibliography VII

- [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 36, 39).
- [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 46).
- [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 71).
- [43] Oord, A., Li, Y., Vinyals, O., *Representation Learning with Contrastive Predictive Coding*. 2018. arXiv: 1807.03748 (cited on page 117).
- [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 116, 117).
- [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 71).
- [46] Ouyang, L. "Training Language Models to Follow Instructions with Human Feedback". In: *Advances in Neural Information Processing Systems*. 2022 (cited on pages 96, 97).



bibliography

Bibliography

2024-03-03

- [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 36, 39).
- [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 46).
- [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 71).
- [43] Oord, A., Li, Y., Vinyals, O., *Representation Learning with Contrastive Predictive Coding*. 2018. arXiv: 1807.03748 (cited on page 117).
- [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 116, 117).
- [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 71).
- [46] Ouyang, L. "Training Language Models to Follow Instructions with Human Feedback". In: *Advances in Neural Information Processing Systems*. 2022 (cited on pages 96, 97).

BIBLIOGRAPHY

Bibliography VIII

- [47] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., "Improving Language Understanding by Generative Pre-Training". In: (2018) (cited on page 91).
- [48] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., Finn, C., "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model". In: *Advances in Neural Information Processing Systems*. Volume 36. 2024 (cited on pages 96, 97).
- [49] 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 pages 44, 45).
- [50] 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 36).
- [51] 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 117).
- [52] 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 45).



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

- ### bibliography

 - ### Bibliography

- BIBLIOGRAPHY
- Bibliography VIII
- [47] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., "Improving Language Understanding by Generative Pre-Training". In: (2018) (cited on page 91).
- [48] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., Finn, C., "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model". In: *Advances in Neural Information Processing Systems*. Volume 36. 2024 (cited on pages 96, 97).
- [49] 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 pages 44, 45).
- [50] 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 36).
- [51] 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 117).
- [52] 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 45).

Bibliography IX



- [53] 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 10, 11).
- [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 8, 9).
- [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 10, 11).
- [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 71).
- [57] Verma, R., Nalisnick, E., "Calibrated Learning to Defer with One-vs-All Classifiers". In: *International Conference on Machine Learning*. PMLR, 2022 (cited on pages 92, 93).
- [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 71).

bibliography

Bibliography

2024-03-03

- BIBLIOGRAPHY**
- Bibliography IX**
- [53] 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 10, 11).
 - [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 8, 9).
 - [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 10, 11).
 - [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 71).
 - [57] Verma, R., Nalisnick, E., "Calibrated Learning to Defer with One-vs-All Classifiers". In: *International Conference on Machine Learning*. PMLR, 2022 (cited on pages 92, 93).
 - [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 71).

Bibliography X

- [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 pages 44, 45).



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

- └ bibliography
- └ Bibliography

Results on diverse datasets

OOD dataset	Metric	AUROC↑	AUPRC↑	FPR80↓
Trained on CIFAR10				
SVHN	LLR ^{>2}	0.811	0.837	0.394
CIFAR10	LLR ^{>1}	0.469	0.479	0.835
Trained on SVHN				
CIFAR10	LLR ^{>1}	0.939	0.950	0.052
SVHN	LLR ^{>1}	0.489	0.484	0.799



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

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

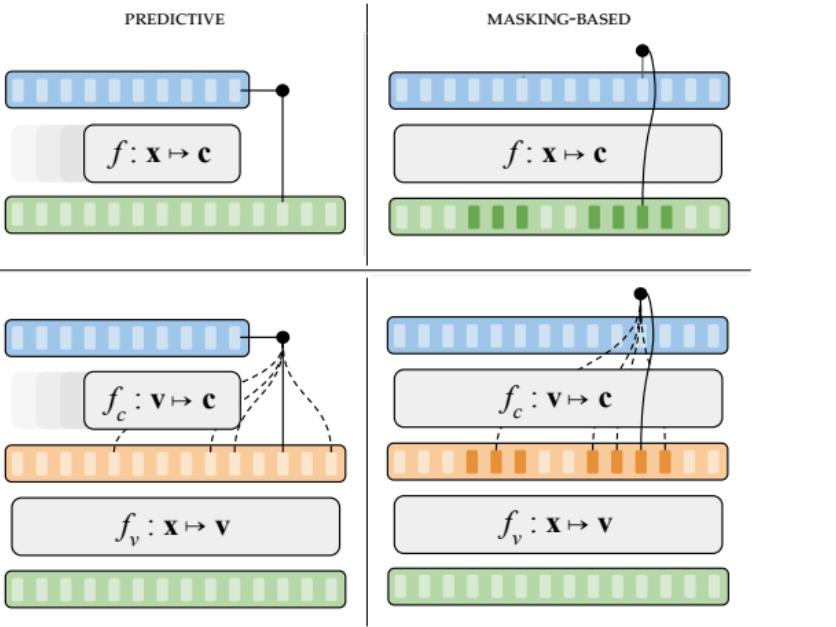
↳ Results on diverse datasets

OOD dataset	Metric	Trained on FashionMNIST		
		AUROC↑	AUPRC↑	FPR80↓
MNIST	LLR ^{>1}	0.986	0.987	0.011
notMNIST	LLR ^{>1}	0.998	0.998	0.000
KMNIST	LLR ^{>1}	0.974	0.977	0.017
Omniglot28x28	LLR ^{>2}	1.000	1.000	0.000
Omniglot28x28Inverted	LLR ^{>1}	0.954	0.954	0.050
SmallNORB28x28	LLR ^{>2}	0.999	0.999	0.002
SmallNORB28x28Inverted	LLR ^{>2}	0.941	0.946	0.069
FashionMNIST	LLR ^{>1}	0.488	0.496	0.811

OOD dataset	Metric	Trained on MNIST		
		AUROC↑	AUPRC↑	FPR80↓
FashionMNIST	LLR ^{>1}	0.999	0.999	0.000
notMNIST	LLR ^{>1}	1.000	0.999	0.000
KMNIST	LLR ^{>1}	0.999	0.999	0.000
Omniglot28x28	LLR ^{>1}	1.000	1.000	0.000
Omniglot28x28Inverted	LLR ^{>1}	0.944	0.953	0.057
SmallNORB28x28	LLR ^{>1}	1.000	1.000	0.000
SmallNORB28x28Inverted	LLR ^{>1}	0.985	0.987	0.000
MNIST	LLR ^{>2}	0.515	0.507	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.

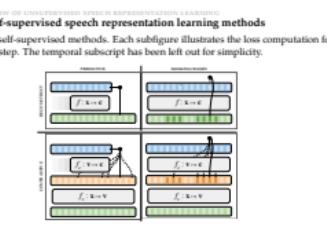


UNCERTAINTY AND THE MEDICAL INTERVIEW

└ a brief overview of unsupervised speech representation learning

└ Types of self-supervised speech representation learning methods

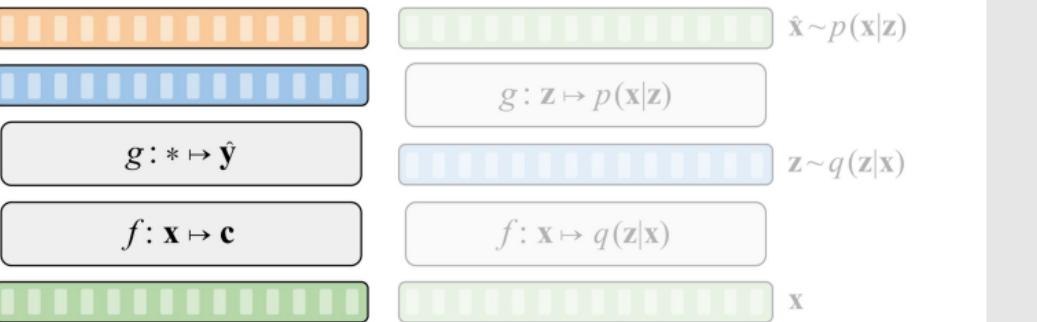
2024-03-03



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.



2024-03-03

UNCERTAINTY AND THE MEDICAL INTERVIEW

a brief overview of unsupervised speech representation learning

Overview: Representation Learning for Speech

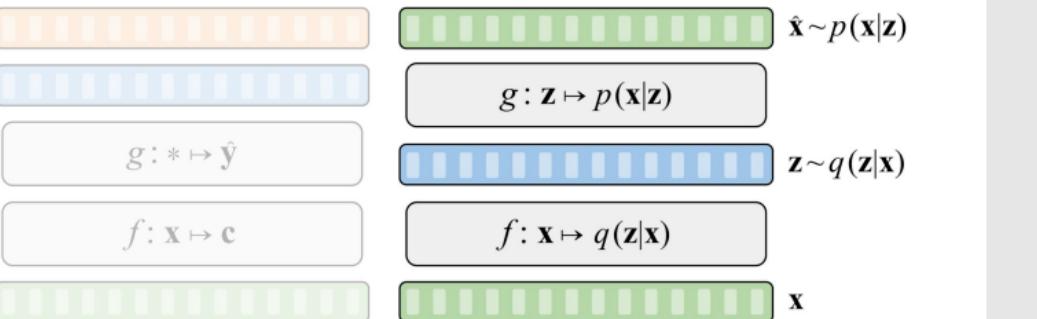
A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING
Overview: Representation Learning for Speech

- 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.

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

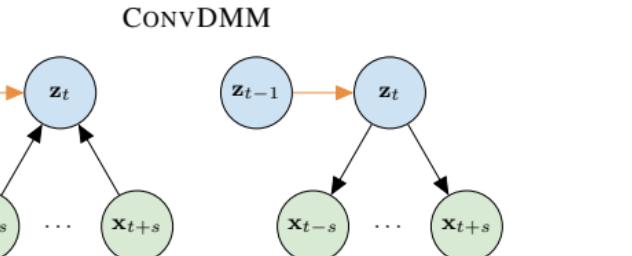
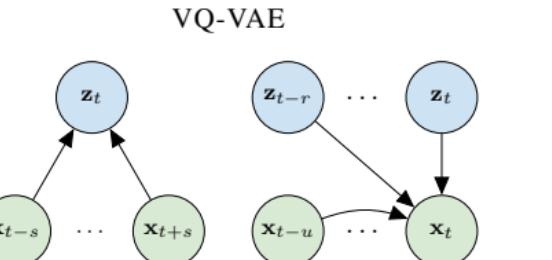
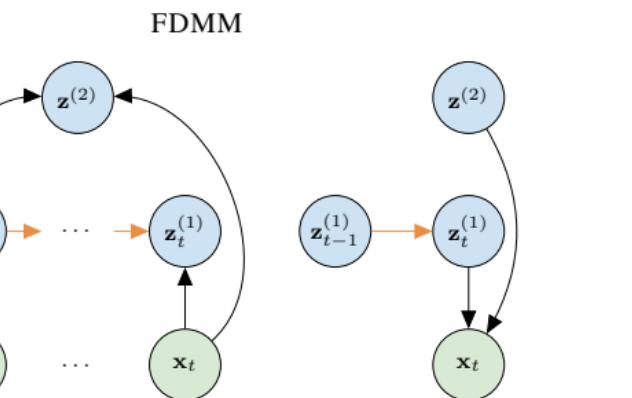
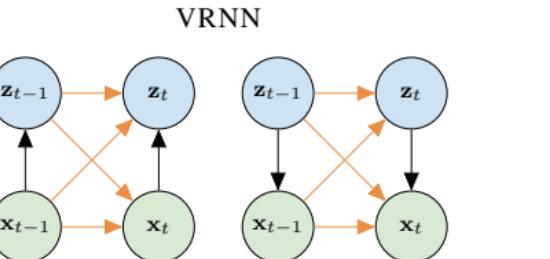
- └ a brief overview of unsupervised speech representation learning
- └ Overview: Representation Learning for Speech

2024-03-03

- 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



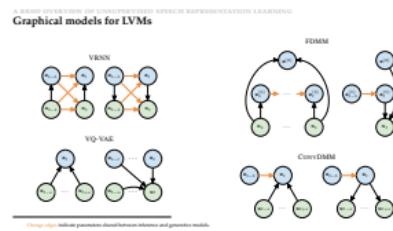
Orange edges indicate parameters shared between inference and generative models.

UNCERTAINTY AND THE MEDICAL INTERVIEW

a brief overview of unsupervised speech representation learning

Graphical models for LVMs

2024-03-03



A BRIEF OVERVIEW OF UNSUPERVISED SPEECH REPRESENTATION LEARNING

Overview of LVM probabilistic components



TYPE	FORM
OBSERVATION MODEL	
ARX	Autoregressive on x_t $p(x_t x_{1:t-1})$
LOC	Local latent variable $p(x_t z_{1:t})$
GLB	Global latent variable $p(x_t z)$
PRIOR	
ARX	Autoregressive on x_t $p(z_t x_{1:t-1})$
ARZ	Autoregressive on z_t $p(z_t z_{1:t-1})$
IND	Locally independent z_t $p(z_t)$
GLB	Global latent variable $p(z)$
INFERENCE MODEL	
ARZ	Autoregressive on z_t $q(z_t z_{1:t-1})$
FLT	Filtering $q(z_t x_{1:t})$
LSM	Local smoothing $q(z_t x_{t-r:t+r})$
GSM	Global smoothing $q(z_t x_{1:T})$
GLB	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	
ARX	Autoregressive on x_t $p(x_t x_{1:t-1})$
LOC	Local latent variable $p(x_t z_{1:t})$
GLB	Global latent variable $p(x_t z)$
Prior	
ARX	Autoregressive on x_t $p(z_t x_{1:t-1})$
ARZ	Autoregressive on z_t $p(z_t z_{1:t-1})$
IND	Locally independent z_t $p(z_t)$
GLB	Global latent variable $p(z)$
Inference model	
ARZ	Autoregressive on z_t $q(z_t z_{1:t-1})$
FLT	Filtering $q(z_t x_{1:t})$
LSM	Local smoothing $q(z_t x_{t-r:t+r})$
GSM	Global smoothing $q(z_t x_{1:T})$
GLB	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 [18]	✓	✓	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗
HMM-VAE [16]	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✓
ConvVAE [26]	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗
FHVAE [27]	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓	✓	✓	✓
VQ-VAE [44]	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗
BHMM-VAE [20]	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗
STCN [1]	✗	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓
FDMM [30]	✗	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✓	✓	✓
ConvDMM [31]	✗	✓	✗	✗	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗

UNCERTAINTY AND THE MEDICAL INTERVIEW

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

2024-03-03

MODEL	OBSERVATION				PRIOR				INFERENCE					
	ARX	LOC	GLB	ARX IND GLB	ARZ	IND	GLB	ARZ	FLT	LSM	GSM	GLB	HIE	
VRNN [14]	✓	✓	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗
SRNN [18]	✓	✓	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗
HMM-VAE [16]	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✓
ConvVAE [26]	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗
FHVAE [27]	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓	✓	✓	✓
VQ-VAE [44]	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗
BHMM-VAE [20]	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗
STCN [1]	✗	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓
FDMM [30]	✗	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✓	✓	✓
ConvDMM [31]	✗	✓	✗	✗	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗

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
SELF-SUPERVISED MODELS											
CPC [43]	✗	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗
APC [13]	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗
wav2vec [51]	✗	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗
Mockingjay [39]	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✓
wav2vec 2.0 [4]	✓	✗	✓	✗	✓	✗	✓	✗	✗	✗	✓
NPC [38]	✓	✗	✗	✓	✓	✗	✓	✗	✗	✓	✗
DeCoAR 2.0 [37]	✓	✗	✗	✓	✓	✗	✓	✗	✗	✓	✗
HuBERT [25]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✓
data2vec [3]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓
LATENT VARIABLE MODELS											
VRNN [14]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
SRNN [18]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
ConvVAE [26]	✗	✗	✗	✓	✗	✓	✗	✓	✗	✓	✗
FHVAE [27]	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✗
VQ-VAE [44]	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗
STCN [1]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
FDMM [30]	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✗
ConvDMM [31]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗

UNCERTAINTY AND THE MEDICAL INTERVIEW

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

2024-03-03

Method	Model and Task Design			Resolution	Usage
	MSK	PRD	CON		
CPC [43]	✗	✓	✓	✓	✓
APC [13]	✗	✓	✗	✓	✓
Mockingjay [39]	✓	✗	✗	✓	✓
data2vec [3]	✓	✗	✗	✓	✓
VRNN [14]	✗	✗	✗	✓	✓
SRNN [18]	✗	✗	✗	✓	✓
ConvVAE [26]	✗	✗	✗	✓	✓
FHVAE [27]	✗	✗	✗	✓	✓
VQ-VAE [44]	✗	✗	✗	✓	✓
STCN [1]	✗	✗	✗	✓	✓
FDMM [30]	✗	✗	✗	✓	✓
ConvDMM [31]	✗	✗	✗	✓	✓