UNCERTAINTY AND THE MEDICAL INTERVIEW

TOWARDS SELF-ASSESSMENT IN MACHINE LEARNING MODELS

Jakob Drachmann Havtorn

DTU Compute
Department of Applied Mathematics and Computer Science

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**

Overview
Thesis

Background

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.

Medical dialogue

Medical dialogue

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

Errors in medical dialogue

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

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 is a leading cause of medical error contributing to two out of three adverse events [51].
- Preventability: A considerable fraction of all hospital admissions have preventable adverse outcomes[a] [11].

[a] 9% to 16.6% in AU, NZ, UK, DK.

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 [27, 50, 52][a].
- Varying quality: Discharge summaries almost never meet *all* timeline, transmission, and content criteria. [23][b]

[a] Ambulatory care across four specialties in four states and tertiary care at an academic medical center.
[b] Outpatient visits, Yale-New Haven Hospital.

How might machine learning help?

- 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.

Reliability of machine learning systems

- Data: Quality, quantity, diversity, bias, privacy, ethics.

Reliability of machine learning systems

- 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).

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).

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.

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

Modular approach:

Building a decision-support system

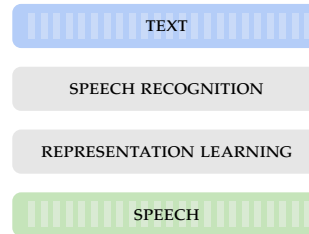
Modular approach:

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

Building a decision-support system

Modular approach:

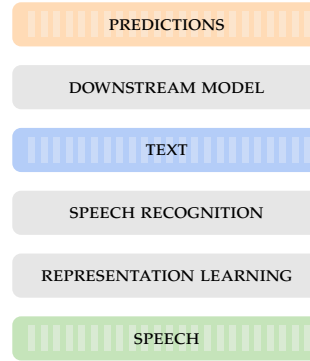
- Source data: Speech or text (potentially images, video, electronic health records, etc.).
- Foundation modelling: Converting the input into representations useful for downstream tasks.



Building a decision-support system

Modular approach:

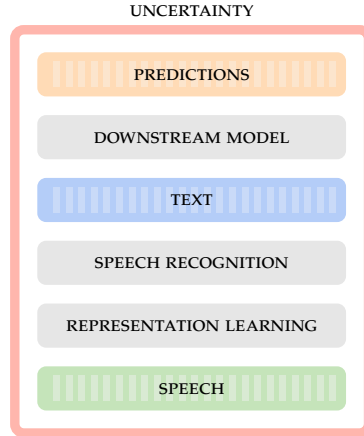
- Source data: Speech or text (potentially images, video, electronic health records, etc.).
- Foundation modelling: Converting the input into representations useful for downstream tasks.
- Downstream tasks: Suggesting, summarizing, classifying, transcribing, translating, etc.



Building a decision-support system

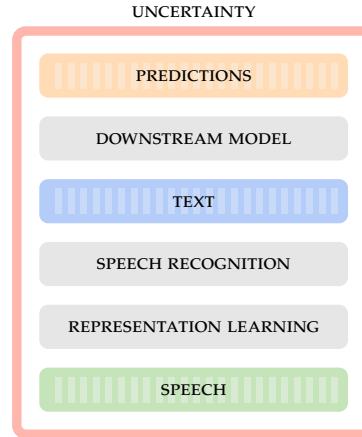
Modular approach:

- 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.

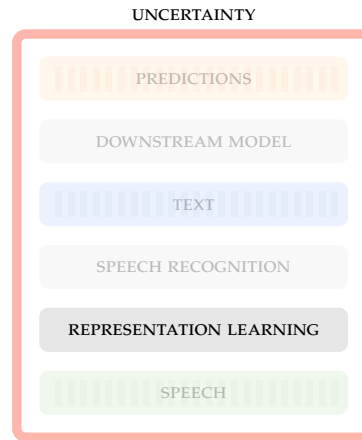


Thesis

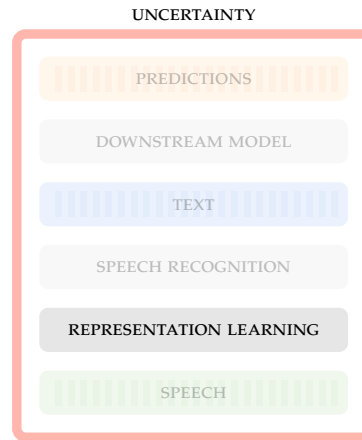
Chapter 1-3	**Introduction, research questions, and background**
Chapter 4	**Hierarchical VAEs know what they don't know**
Chapter 5	**Model-agnostic out-of-distribution detection using combined statistical tests**
Chapter 6	**A brief overview of unsupervised speech representation learning**
Chapter 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**



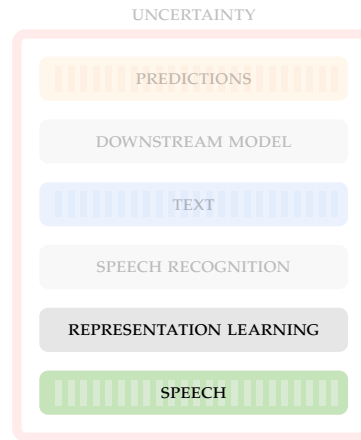
Thesis



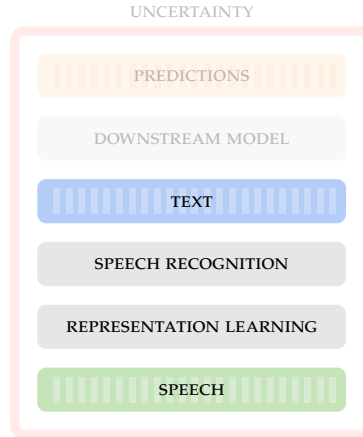
Thesis



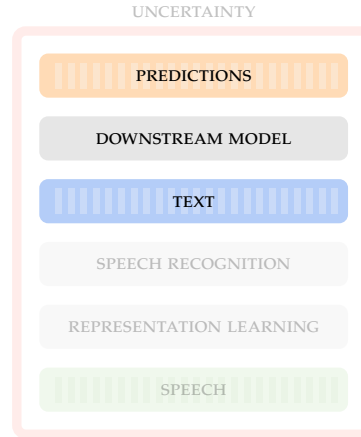
Thesis



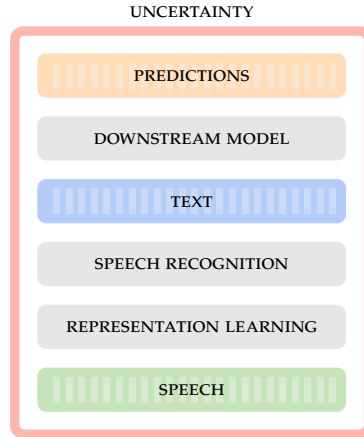
Thesis



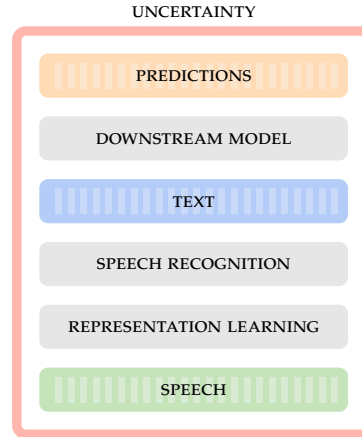
Thesis



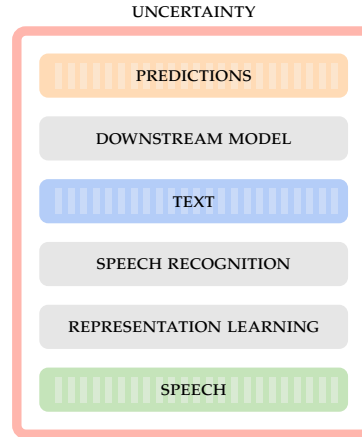
Thesis



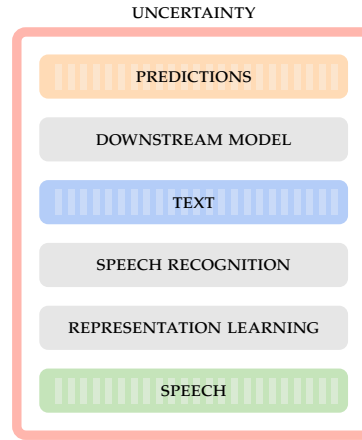
Presentation



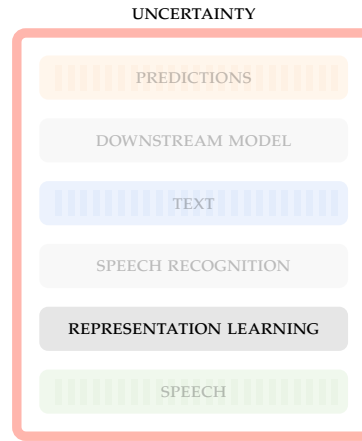
Presentation



Presentation



Presentation

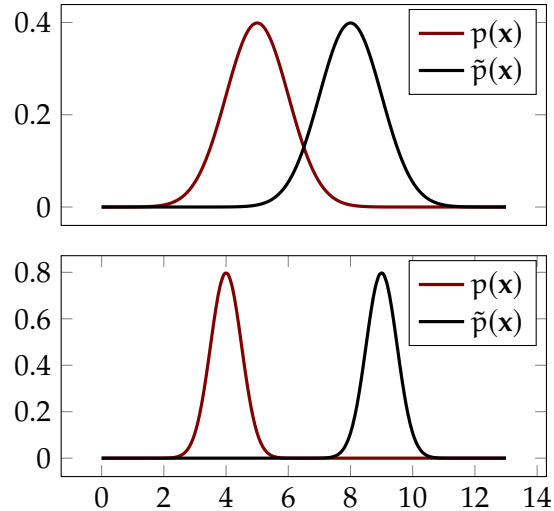


Defining OOD detection

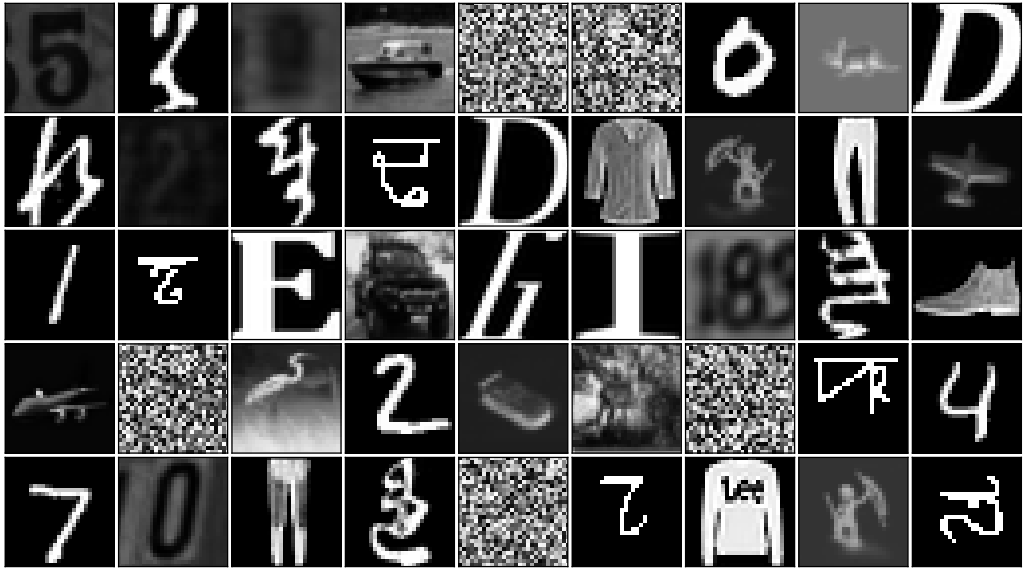
Enable models to distinguish the training data distribution $p(\mathbf{x})$ from any other distribution $\tilde{p}(\mathbf{x})$.

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

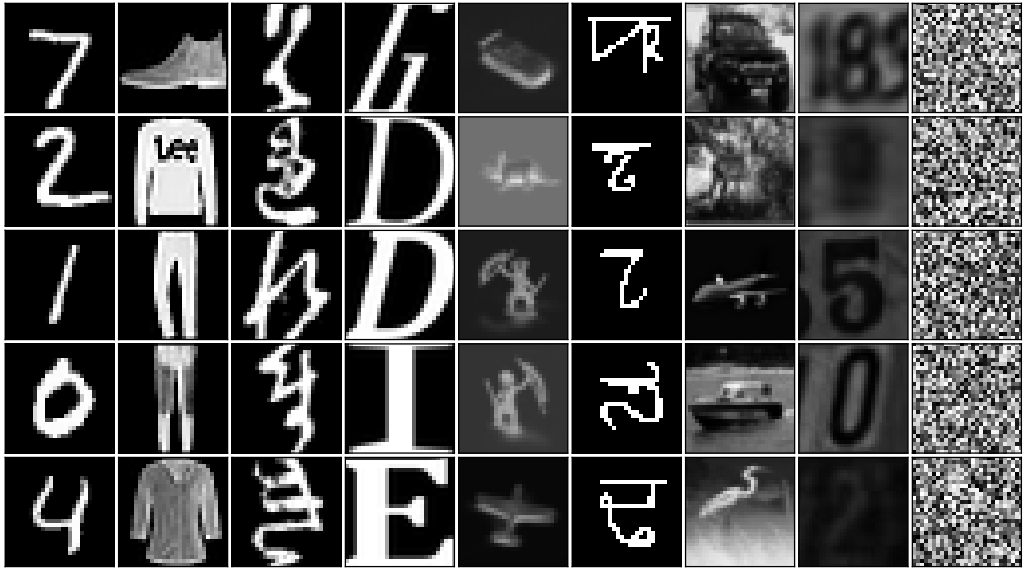
"Was $\mathbf{x}$ sampled from $p(\mathbf{x})$ or not?"



In distribution?



Out of distribution?



Problem and Contributions

- Deep generative models often fail at OOD detection task when using their likelihood estimate as the score function [40] by, perhaps surprisingly, assigning higher likelihoods to OOD data.
- Contributions:
  - We provide evidence that out-of-distribution detection fails due to learned low-level features that generalize across datasets.
  - We present a new score for OOD detection with hierarchical VAEs that alleviates this issue.

Hierarchical VAE

We choose the hierarchical VAE as our model [31, 47].

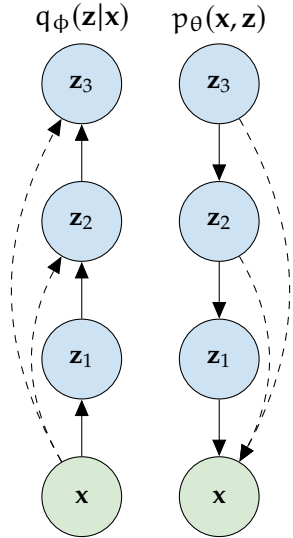
$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z}$$

Specifically we use

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

$$\text{Generative model:} \quad p_\theta(\mathbf{x}|\mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z}_1) p_\theta(\mathbf{z}_1|\mathbf{z}_2) p(\mathbf{z}_3),$$
$$\text{Inference model:} \quad q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}_1|\mathbf{x}) q_\phi(\mathbf{z}_2|\mathbf{z}_1) q_\phi(\mathbf{z}_3|\mathbf{z}_2).$$

2. a ten-layered layered Bidirectional-Inference Variational Autoencoder (BIVA) [39].

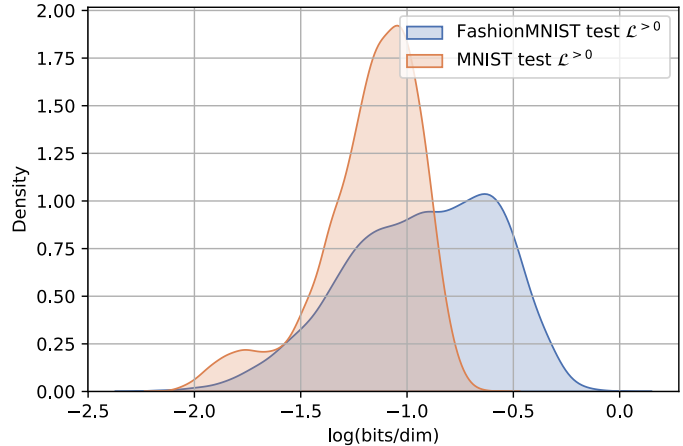


Out-of-distribution detection with hierarchical VAEs

- Generative models learn to approximate the **data distribution** $p(\mathbf{x})$.
- The likelihood of the model given a sample $\mathbf{x}$ 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 hierarchical VAEs

- 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].



What is wrong with the ELBO for OOD detection?

We can split the ELBO into two terms

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

The first term is high if the data is well-explained by $\mathbf{z}$.

The second term we can rewrite as,

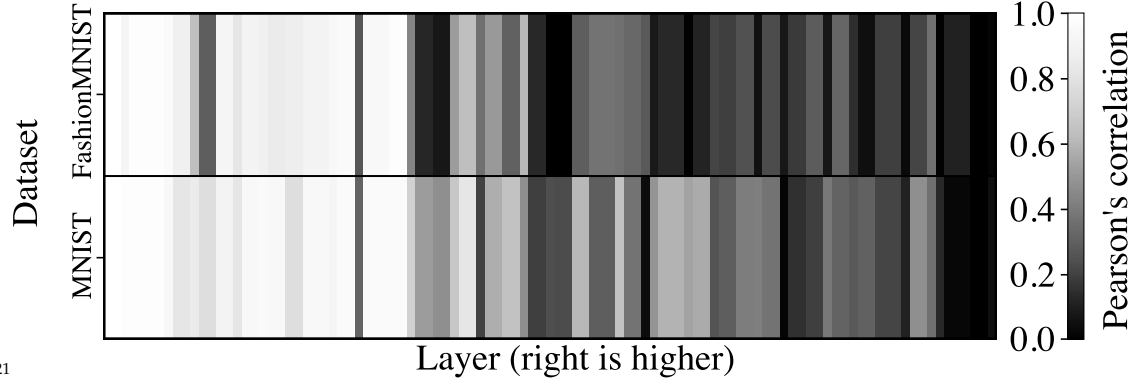
$$D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{i=1}^{L-1} \log \frac{p_\theta(\mathbf{z}_i|\mathbf{z}_{i+1})}{q_\phi(\mathbf{z}_i|\mathbf{z}_{i-1})} + \log \frac{p_\theta(\mathbf{z}_L)}{q_\phi(\mathbf{z}_L|\mathbf{z}_{L-1})}\right] . \quad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

The absolute log-ratios grow with $\dim(\mathbf{z}_i)$ since the log probability terms are computed by summing over the dimensionality of $\mathbf{z}_i$.

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).



An alternative likelihood bound, $\mathcal{L}^{>k}$

An alternative version of the ELBO that only partially uses the approximate posterior can be written as [39]

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

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

$$p_\theta(\mathbf{z}_{\leqslant k}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})$$

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].

$$\mathrm{LLR}^{>k}(\mathbf{x}) \equiv \mathcal{L}(\mathbf{x}) - \mathcal{L}^{>k}(\mathbf{x}) \,. \tag{4}$$

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

$$\begin{aligned} \mathcal{L} &= \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right), \\ \mathcal{L}^{>k} &= \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(p_\theta(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right) \,. \end{aligned} \tag{5}$$

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

$$\begin{aligned} \mathrm{LLR}^{>k}(\mathbf{x}) = &-D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right) \\ &+ D_{\mathrm{KL}}\left(p_\theta(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right) \,. \end{aligned} \tag{6}$$

Importance sampling the ELBO

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

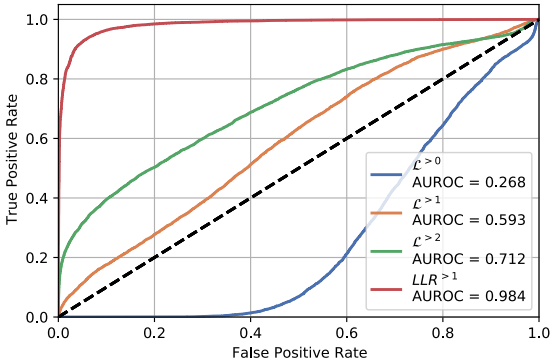
$$\mathcal{L}_S = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log \frac{1}{N}\sum_{s=1}^{S}\frac{p(\mathbf{x},\mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})}\right] \leqslant \log p_\theta(\mathbf{x})\,, \tag{7}$$

Consequently, by importance sampling the ELBO, the associated KL-divergence vanishes and our likelihood ratio reduces to the KL-divergence of $\mathcal{L}^{>k}$.

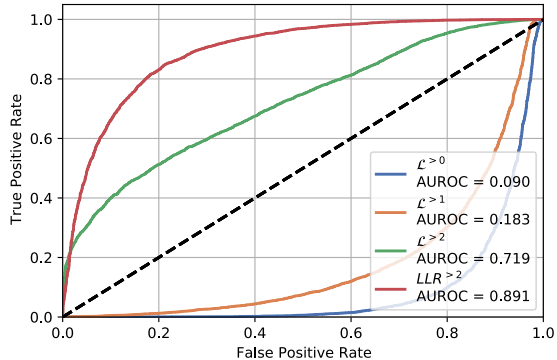
$$\mathrm{LLR}_S^{>k}(\mathbf{x}) \to \mathrm{D}_{\mathrm{KL}}(p(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k})q(\mathbf{z}_{>k}|\mathbf{x})||p(\mathbf{z}|\mathbf{x}))\,. \tag{8}$$

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

Results with $\text{LLR}^{>k}$



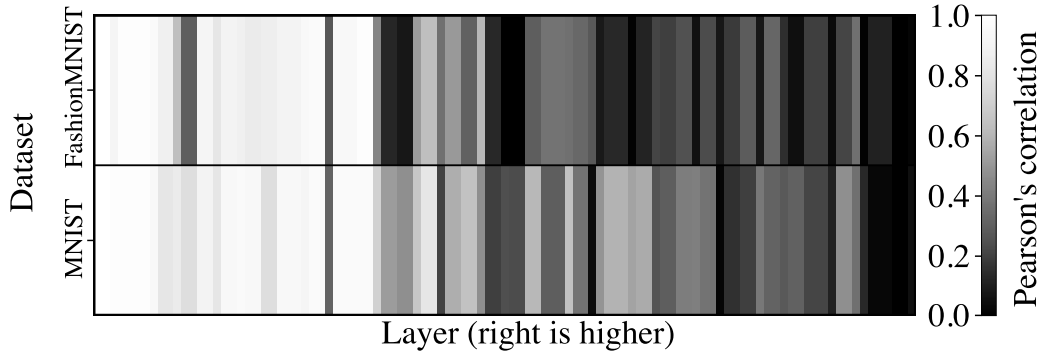
(a) FashionMNIST HVAE evaluated on MNIST



(b) CIFAR10 BIVA evaluated on SVHN

Selecting the value of k

- Use validation OOD dataset(s).
- Compute $\text{LLR}^{>k}$ 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.



Results on FashionMNIST/MNIST

Method	AUROC↑	AUPRC↑	FPR80↓
FashionMNIST (in) / MNIST (out)			
Use prior knowledge of OOD			
Backgr. contrast. LR (PixelCNN) [46]	0.994	0.993	0.001
Backgr. contrast. LR (VAE) [12]	0.924	-	-
Binary classifier [46]	0.455	0.505	0.886
$p(\hat{y}\|\mathbf{x})$ with OOD as noise class [46]	0.877	0.871	0.195
$p(\hat{y}\|\mathbf{x})$ with calibration on OOD [46]	0.904	0.895	0.139
Input complexity (S, Glow) [22]	0.998	-	-
Input complexity (S, PixelCNN++) [22]	0.967	-	-
Use in-distribution data labels $\mathbf{y}$			
$p(\hat{y}\|\mathbf{x})$ [21, 46]	0.734	0.702	0.506
Entropy of $p(\mathbf{y}\|\mathbf{x})$ [46]	0.746	0.726	0.448
ODIN [35, 46]	0.752	0.763	0.432
VIB [2, 12]	0.941	-	-
Mahalanobis distance, CNN [46]	0.942	0.928	0.088
Mahalanobis distance, DenseNet [34]	0.986	-	-
Ensemble, 20 classifiers [33, 46]	0.857	0.849	0.240
No OOD-specific assumptions			
- Ensembles			
WAIC, 5 models, VAE [12]	0.766	-	-
WAIC, 5 models, PixelCNN [46]	0.221	0.401	0.911
- Not ensembles			
Likelihood regret [56]	**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
$\text{LLR}^{>1}$ + HVAE (ours)	0.964	0.961	0.036
$\text{LLR}^{>1}_{250}$ + HVAE (ours)	0.984	**0.984**	**0.013**

Results on CIFAR10/SVHN

Method	AUROC↑	AUPRC↑	FPR80↓
CIFAR10 (in) / SVHN (out)			
Use prior knowledge of OOD			
Backgr. contrast. LR (PixelCNN) [46]	0.930	0.881	0.066
Backgr. contrast. LR (VAE) [56]	0.265	-	-
Outlier exposure [22]	0.984	-	-
Input complexity (S, Glow) [49]	0.950	-	-
Input complexity (S, PixelCNN++) [49]	0.929	-	-
Input complexity (S, HVAE) (Ours) [49]	0.833	0.855	0.344
Use in-distribution data labels y			
Mahalanobis distance [34]	0.991	-	-
No OOD-specific assumptions			
- Ensembles			
WAIC, 5 models, Glow [12]	1.000	-	-
WAIC, 5 models, PixelCNN [46]	0.628	0.616	0.657
- Not ensembles			
Likelihood regret [56]	0.875	-	-
$LLR^{>2}$ + HVAE (ours)	0.811	0.837	0.394
$LLR^{>2}$ + BIVA (ours)	**0.891**	**0.875**	**0.172**

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

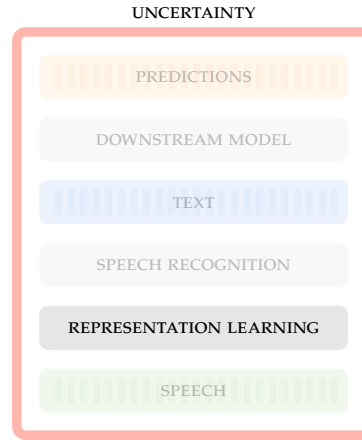
OOD dataset	Metric	AUROC↑	AUPRC↑	FPR80↓
Trained on FashionMNIST				
MNIST	$LLR^{>1}$	0.986	0.987	0.011
notMNIST	$LLR^{>1}$	0.998	0.998	0.000
KMNIST	$LLR^{>1}$	0.974	0.977	0.017
Omniglot28x28	$LLR^{>2}$	1.000	1.000	0.000
Omniglot28x28Inverted	$LLR^{>1}$	0.954	0.954	0.050
SmallNORB28x28	$LLR^{>2}$	0.999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

Conclusions

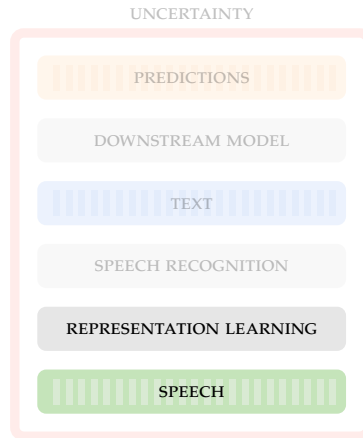
- Key observations:
  - The likelihood of a generative model is not a good score for OOD detection [40].
  - Strong correlations between some latent variables for different datasets.
  - Reconstructions of OOD data are good when using full approximate posterior.

→ Proposed a new score, $\text{LLR}^{>k}$, that uses the conditional prior for the top-most latent variables in the hierarchy.

Presentation

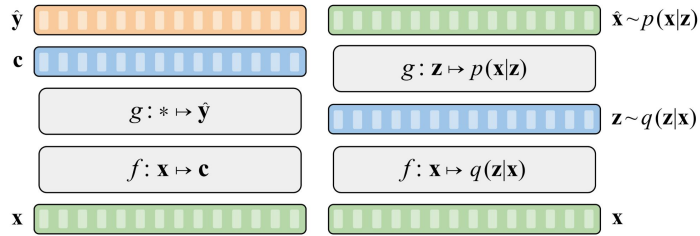


Presentation



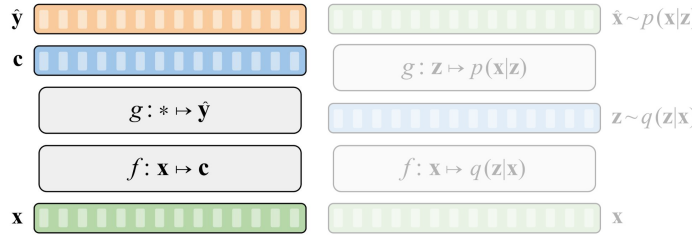
Overview: Representation Learning for Speech

- 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.

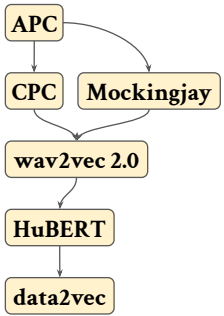


Overview: Representation Learning for Speech

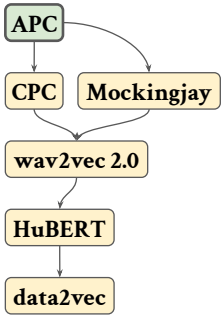
- We focus on two primary categories:
  - Self-supervised learning (SSL)
  - Probabilistic latent variable models (LVMs)
- Recent developments have been driven by self-supervised learning.
- A model-by-model overview: Focus on speech recognition.



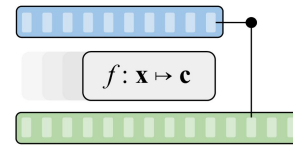
Development of SSL for speech



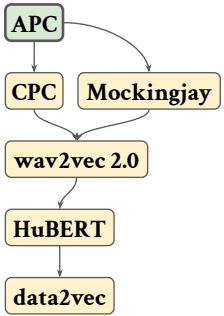
Autoregressive Predictive Coding (APC)



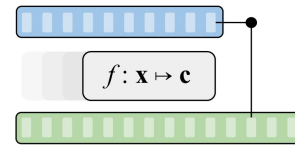
- **Task**: Predict future inputs.
- **Input/target**: Log-mel spectrogram.
- **Architecture**: RNN/Transformer decoder.
- **Slow features**: Predict k steps ahead.



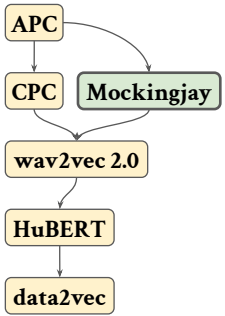
Autoregressive Predictive Coding (APC)



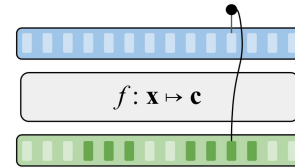
- **Challenges**:
  - Encodes only past inputs ✗
  - Uses the input as target ✗



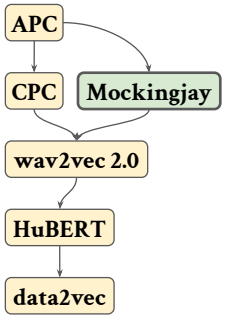
Mockingjay



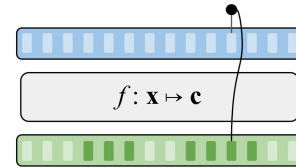
- **Task**: Reconstruct masked inputs.
- **Architecture**: Transformer encoder.
- **Masking**:
  - X% at random. (Mockingjay)
  - X% + N consecutive (wav2vec 2.0)
  - SpecAugment (Masked RNN)



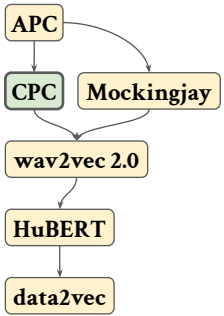
Mockingjay



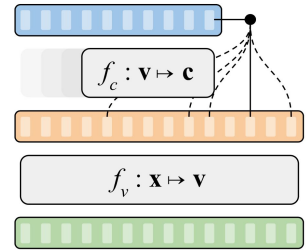
- **Challenges**:
  - Encodes the entire input ✓
  - Uses the input as target ✗



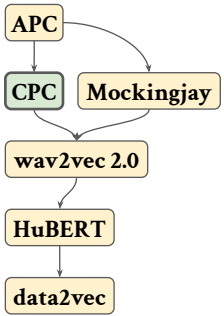
Contrastive Predictive Coding (CPC)



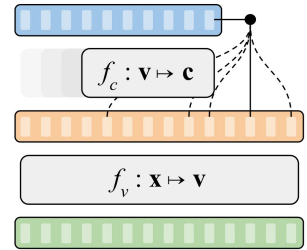
- **Contrastive models**: Distinguish target samples from negative samples.
- **Learned target**: Discard details.
- **Sampling negatives**:
  - Sample sequence?
  - Same speaker?



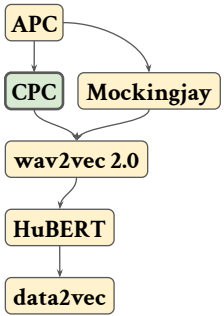
Contrastive Predictive Coding (CPC)



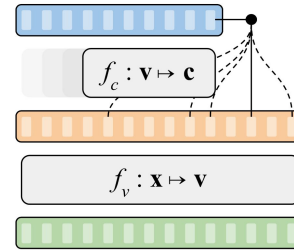
- **Challenges**:
  - Only encodes past inputs ✗
  - Uses a learned target ✓



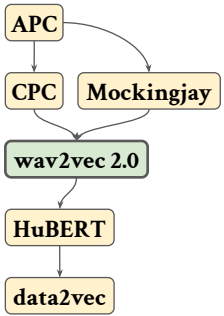
Contrastive Predictive Coding (CPC)



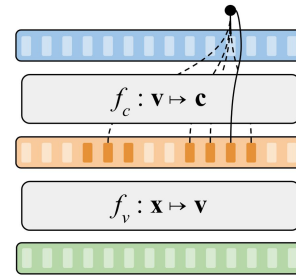
- **Challenges**:
  - Only encodes past inputs ✗
  - Uses a learned target ✓
  - Sampling negatives ✗



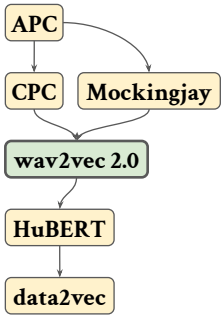
wav2vec 2.0



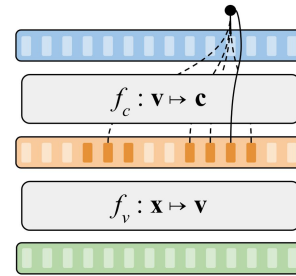
- Masking + contrastive learning.
- **Quantisation**: Better negative samples.
- **Results**:
  - 960 hours: **2.0%** WER.
  - 10 mimutes: **4.8%** WER.



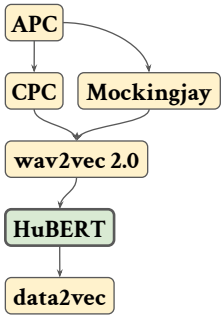
wav2vec 2.0



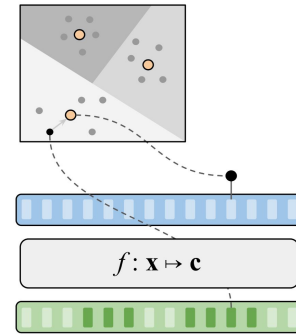
- **Challenges**:
  - Encodes the entire input ✓
  - Uses a learned target ✓
  - Sampling negatives ✗



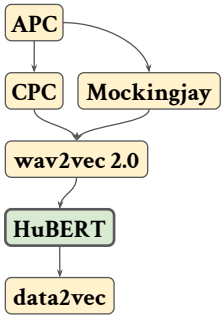
Hidden-unit BERT (HuBERT)



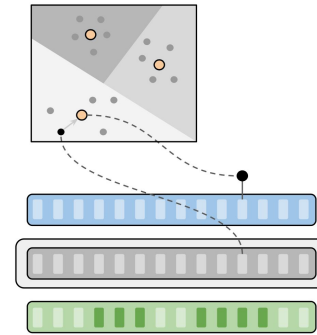
- **Target**: K-means teacher.
- **Training**: Simple cross-entropy loss.
- **1st iteration**: K-means on inputs.



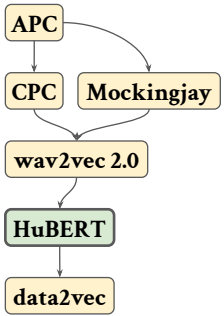
Hidden-unit BERT (HuBERT)



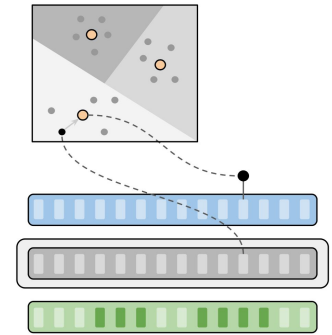
- **Target**: K-means teacher.
- **Training**: Simple cross-entropy loss.
- **1st iteration**: K-means on inputs.
- **2nd iteration**: K-means on hidden layers.



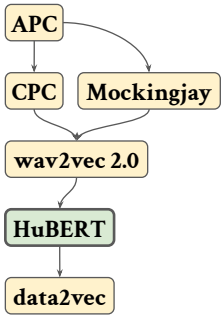
Hidden-unit BERT (HuBERT)



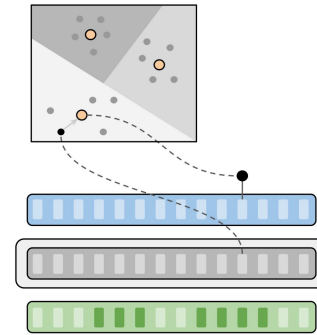
- **Challenges**:
  - Encodes the entire input ✓
  - Uses a learned target ✓
  - No need for negative samples ✓



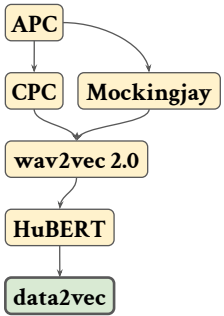
Hidden-unit BERT (HuBERT)



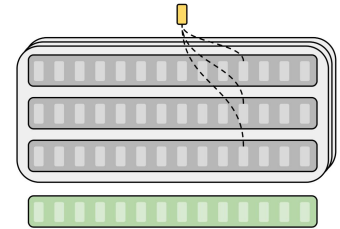
- **Challenges**:
  - Encodes the entire input ✓
  - Uses a learned target ✓
  - No need for negative samples ✓
  - Targets updated infrequently ✗
  - Quantized targets ✗



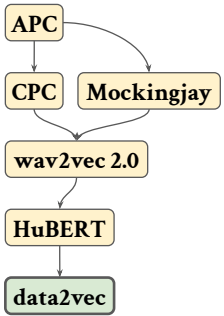
data2vec



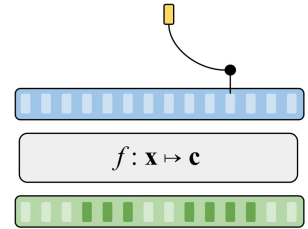
- Uses a teacher-student framework.
- **Teacher**:
  - EMA of student (online) ✓
  - Target is average of top K layers ✓



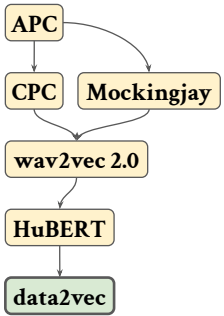
data2vec



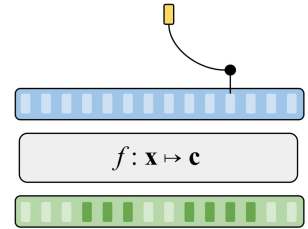
- Uses a teacher-student framework.
- **Teacher**:
  - EMA of student (online) ✓
  - Target is average of top K layers ✓
- **Student training**: Smooth $_1$ loss.



data2vec

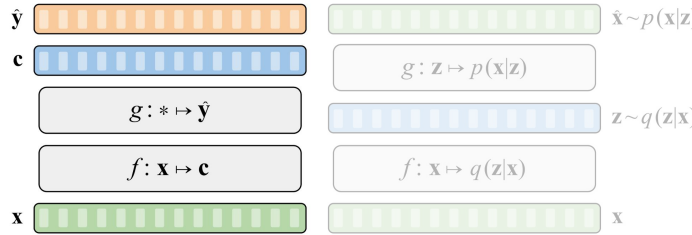


- **Challenges**:
  - Encodes the entire input ✓
  - Uses a learned target ✓
  - No need for negative samples ✓
  - Targets updated continuously ✓
  - Continuous-valued targets ✓



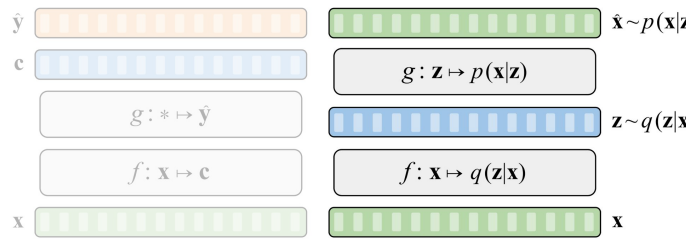
Overview: Representation Learning for Speech

- 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.



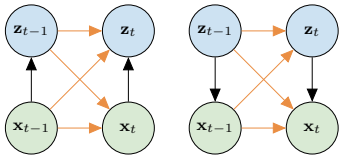
Overview: Representation Learning for Speech

- 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.

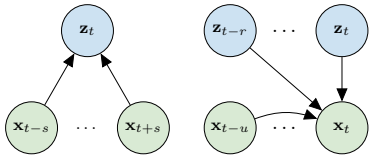


Graphical models for LVMs

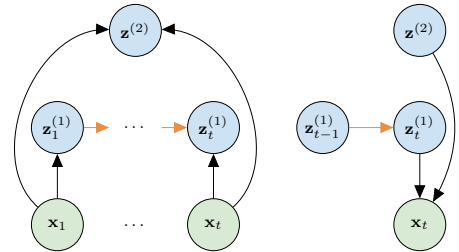
VRNN



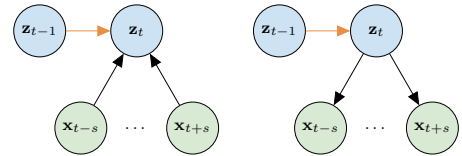
VQ-VAE



FDMM



ConvDMM



Orange edges indicate parameters shared between inference and generative models.

Overview of LVM probabilistic components

Type		Form
	Observation model	
ARX	Autoregressive on $\mathbf{x}_t$	$p(\mathbf{x}_t \vert \mathbf{x}_{1:t-1})$
LOC	Local latent variable	$p(\mathbf{x}_t \vert \mathbf{z}_{1:t})$
GLB	Global latent variable	$p(\mathbf{x}_t \vert \mathbf{z})$
	Prior	
ARX	Autoregressive on $\mathbf{x}_t$	$p(\mathbf{z}_t \vert \mathbf{x}_{1:t-1})$
ARZ	Autoregressive on $\mathbf{z}_t$	$p(\mathbf{z}_t \vert \mathbf{z}_{1:t-1})$
IND	Locally independent $\mathbf{z}_t$	$p(\mathbf{z}_t)$
GLB	Global latent variable	$p(\mathbf{z})$
	Inference model	
ARZ	Autoregressive on $\mathbf{z}_t$	$q(\mathbf{z}_t \vert \mathbf{z}_{1:t-1})$
FLT	Filtering	$q(\mathbf{z}_t \vert \mathbf{x}_{1:t})$
LSM	Local smoothing	$q(\mathbf{z}_t \vert \mathbf{x}_{t-r:t+r})$
GSM	Global smoothing	$q(\mathbf{z}_t \vert \mathbf{x}_{1:T})$
GLB	Global latent variable	$q(\mathbf{z} \vert \mathbf{x}_{1:T})$

Classification of selected LVMs for speech

	Observation			Prior				Inference					
Model	**ARX**	**LOC**	**GLB**	**ARX**	**ARZ**	**IND**	**GLB**	**ARZ**	**FLT**	**LSM**	**GSM**	**GLB**	**HIE**
VRNN [14]	✓	✓	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗
SRNN [17]	✓	✓	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗
HMM-VAE [16]	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✓
ConvVAE [25]	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗
FHVAE [26]	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓	✓	✓
VQ-VAE [43]	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
BHMM-VAE [19]	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗
STCN [1]	✗	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
FDMM [29]	✗	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✓	✓
ConvDMM [30]	✗	✓	✗	✗	✓	✗	✗	✓	✗	✓	✗	✗	✗

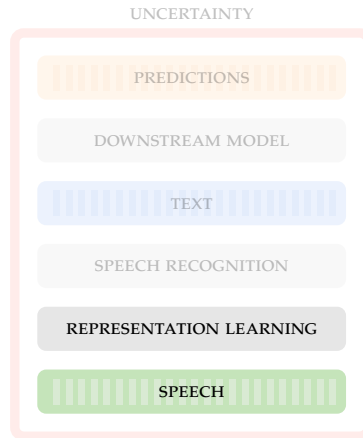
Comparison of LVMs and SSL methods

		MODEL AND TASK DESIGN						RESOLUTION			USAGE	
	MODEL	MSK	PRD	CON	REC	QTZ	GEN	LOC	GLB	VAR	FRZ	FTN
SELF-SUPERVISED MODELS	**CPC** [42]	✗	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗
	APC [13]	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗
	wav2vec [48]	✗	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗
	Mockingjay [38]	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✓
	wav2vec 2.0 [4]	✓	✗	✓	✗	✓	✗	✓	✗	✗	✗	✓
	NPC [37]	✓	✗	✗	✓	✓	✗	✓	✗	✗	✓	✗
	DeCoAR 2.0 [36]	✓	✗	✗	✓	✓	✗	✓	✗	✗	✓	✗
	HuBERT [24]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✓
	data2vec [3]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓
LATENT VARIABLE MODELS	**VRNN** [14]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
	SRNN [17]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
	ConvVAE [25]	✗	✗	✗	✓	✗	✓	✗	✓	✗	✓	✗
	FHVAE [26]	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✗
	VQ-VAE [43]	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗
	STCN [1]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
	FDMM [29]	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✗
	ConvDMM [30]	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗

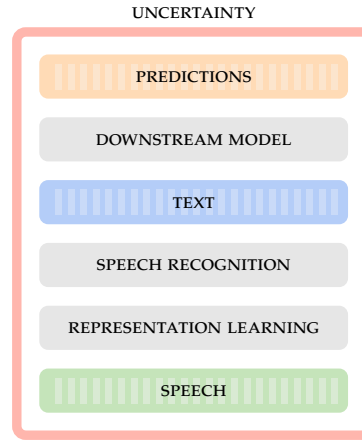
Conclusions

- **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.

Presentation



Presentation



Stroke

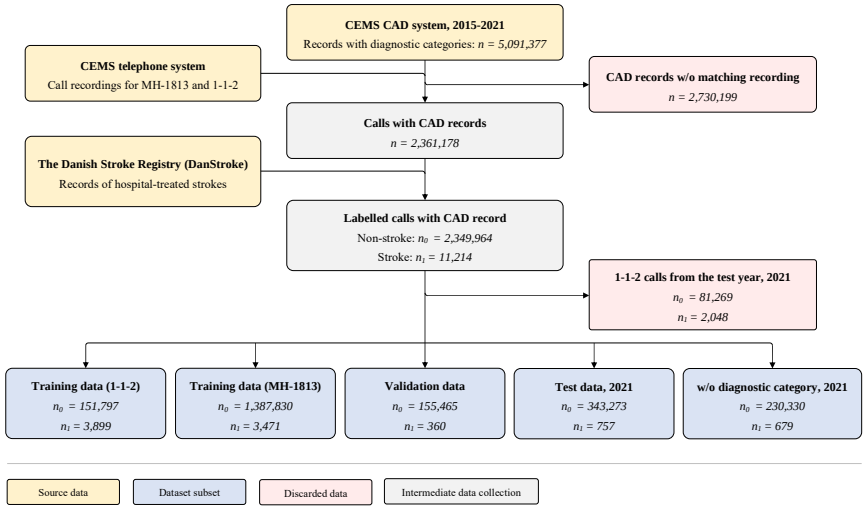
- Stroke is a leading cause of disability and death worldwide [18, 28, 32].
- Effective treatment is very time-sensitive. [5, 53].
- The gateway to ambulance transport and hospital admittance is through prehospital telehealth services.
- Mobile stroke units has made it possible to deliver advanced treatment faster [20, 41].
- The effectiveness of mobile stroke units hinges on call-taker recognition of stroke [20, 41].
- But stroke

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).
- Approximately half of all patients with stroke do not receive the correct triage for their condition from call-takers [8, 44, 55].

- We wanted to investigate if a machine learning model could assist call-takers of 1813 in recognizing stroke.

Population selection and datasets

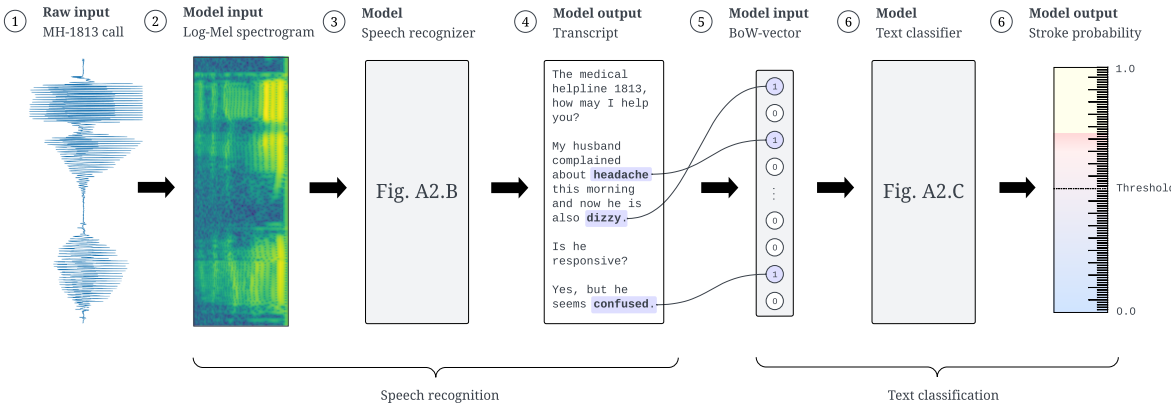


Population characteristics

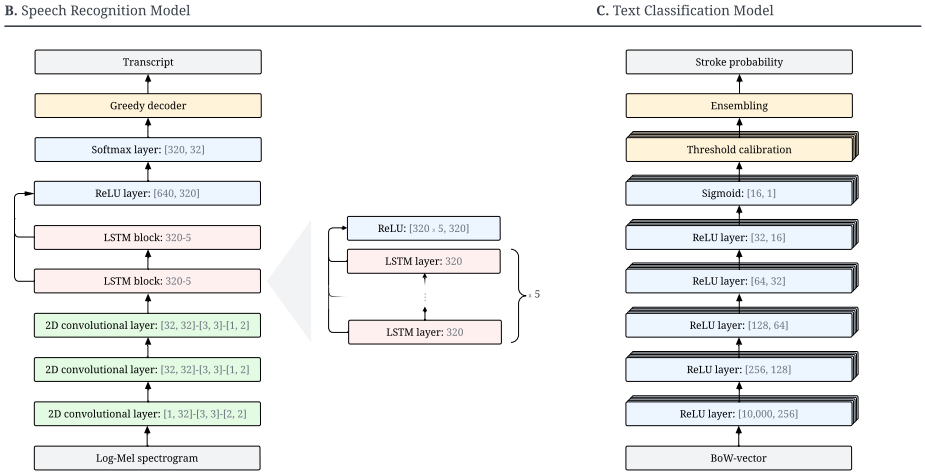
		Training (112)	Training (MH-1813)	Validation	Test	2021 w/o category
All calls	**Num. calls**	155,696	1,391,301	155,825	344,030	231,009
	Female	74,640 (47.94%)	792,783 (56.98%)	86,959 (55.81%)	190,974 (55.51%)	134,324 (58.14%)
	Male	79,564 (51.10%)	596,760 (42.89%)	68,866 (44.19%)	153,050 (44.49%)	96,258 (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

Model design

A. Schematic Overview of Stroke Classification Pipeline



Model design



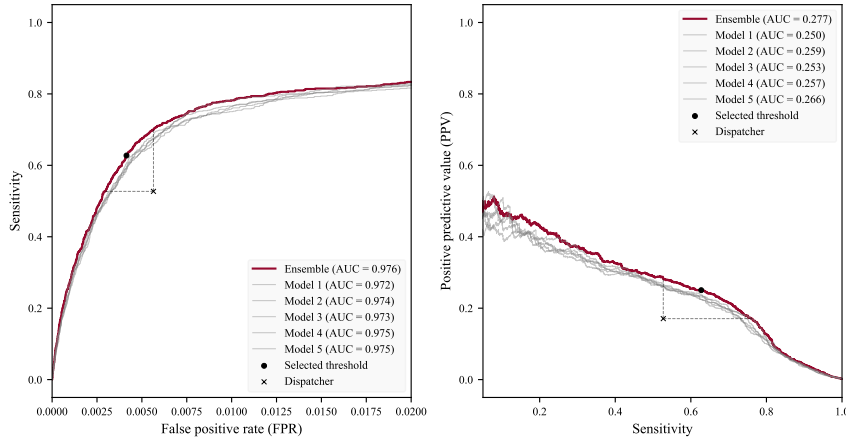
Main results

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)

Main results

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



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 positves 1,938
	Negatives	False negatives 358	True negatives 341,335

		Ground truth labels	
		Positives	Negatives
Model predictions	Positives	True positives 477	False positves 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.

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

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

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

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]

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]

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]

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]

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]

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]

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]

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]

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].

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**
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, 45].
- Might BERT be useful for recognizing stroke?

Fine-tuning a large language model

- Large language models are effective in a wide range of NLP tasks [15, 45].
- 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)

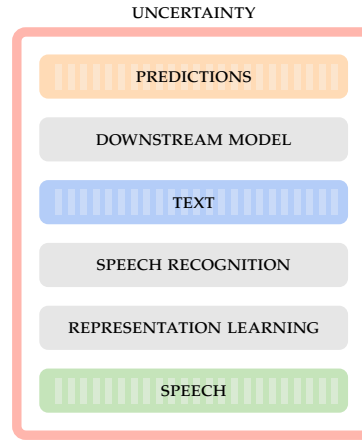
Future work

- Machine learning
  - Learning to predict directly from audio data (SSL).
  - Investigate learning to defer to predict methods [54].

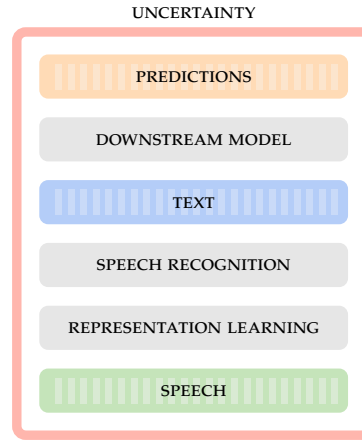
Future work

- Machine learning
  - Learning to predict directly from audio data (SSL).
  - Investigate learning to defer to predict methods [54].
- Clinical applications
  - Mental health (Screening for suicide risk in emergency and medical helpline calls).
  - Maternity ward (Screening for serious pregnancy complications.)

Presentation



Presentation



DISCUSSION

First slide for discussion

Building an operational decision support system

- Do we need true uncertainty estimates? Bayesian methods versus pragmatic methods.

From global to local In table 1, we see that work on global representations within self-supervised learning pre- cedes work on local representations. However, we find that the core ideas underlying the recent successes in learning lo- cal representation models have also been used for global rep- resentation learning; masking (Chung et al. 2016), context prediction (Chung and Glass 2018), and contrastive train- ing (Milde and Biemann 2018) have been applied in both settings. Furthermore, where work on global representation learning has taken inspiration from Word2vec (Mikolov et al. 2013), the techniques used for learning local repre- sentations are inspired by contextualized word embeddings (Devlin et al. 2019). Thus, the gap between these two model classes is largely a product of the developments in related fields and the general increase in computational resources.

Representations beyond inference Predictive tasks are commonly used for self-supervised models, but they are not directly compatible with LVM training. However, an LVM prior with an autoregressive parameterization, $p(z_t \mid z_{1:t-1})$ or $p(z_t \mid x_{1:t-1})$, can be seen as predictive in the sense that it tries to match the approximate posterior. Hence, the prior might be considered for feature extraction. Jones and Moore (2020) examine the importance of the prior in the VQ-VAE and show that the ability of this model to estimate

den- sities p(x1:T) lies solely in its prior. Other work has also explored representations beyond the latent variable such as hidden units of the observation model (Khurana et al. 2020; Chorowski et al. 2019b).

Masking and missing data Masking may also improve representations learned with VAEs. Masking in VAEs has already been explored in the literature in the context of missing data imputation. Here, x is only partially observed, and often represented as a segmentation into observed and miss- ing parts and a mask m indicating where the data is missing. The model is then trained to infer the latent variable from the observed data. Reconstruction also deals only with the ob- served data. Previous work has largely focused on the abil- ity of these models to yield high-quality imputations within the tabular and image data domains, without probing for the effects on the learned latent representation (Mattei and Frellsen 2019; Ipsen, Mattei, and Frellsen 2021). The idea of using VAEs to impute missing data was already examined in the seminal paper by Rezende, Mohamed, and Wierstra (2014). Here the model was trained with fully observed data and used to impute data in an iterative sampling approach post hoc leaving the learned representations unchanged.

Evaluating representations Although this review has a primarily methodological focus, we should briefly touch upon evaluation procedures. Training metrics for self- supervised tasks and the likelihood of LVMs offer little guid- ance as to the quality of the learned representations (Huszár 2017). Thus, a common approach is to evaluate the representations in terms of their usefulness for downstream tasks. Such tasks may be chosen to target specific attributes of the representation (e.g. semantic or speaker information).
The SUPERB benchmark (Yang et al. 2021) gathers mul- tiple tasks grouped into categories such as recognition, de- tection, semantics, speaker, paralinguistics and generation. The recently proposed SLUE benchmark focuses on spo- ken language understanding (Shon et al. 2021). The long- standing zero resource speech challenge (ZeroSpeech) offers a new set of tasks for each edition (Versteegh et al. 2015; Dunbar et al. 2017, 2019, 2020, 2021) usually featuring a minimal-pair ABX task (Schatz et al. 2013, 2014).
Tasks that evaluate representations in terms of speaker- related information include speaker verification (Hsu, Zhang, and Glass 2017b; Khurana et al. 2019; Milde and Biemann 2018), speaker identification (van den Oord, Li, and Vinyals 2018; Jati and

Georgiou 2019; Chung et al. 2019; Liu, Chung, and Glass 2021), dialect classification (Khurana et al. 2019), emotion recognition (Pascual et al. 2019; Yang et al. 2021) and gender classification (Lee et al. 2009). The semantic content of representations are evaluated using tasks such as intent classification (Morais et al. 2021; Yang et al. 2021), slot filling (Lai et al. 2021; Yang et al. 2021), sentiment analysis (Liu et al. 2020), question answer- ing (Chung, Zhu, and Zeng 2020), named entity recogni- tion (Shon et al. 2021; Borgholt et al. 2021a; Pasad et al. 2021) and speech translation (Bansal et al. 2017; Chung and Glass 2020a). Cardiac arrest detection for emergency calls has also been used to evaluate speech representations (Borgholt et al. 2021a). For local representations, phoneme classification is very common (Lee et al. 2009; Hsu, Zhang, and Glass 2017a; Chorowski et al. 2019b; Chung et al. 2019; Liu, Li, and Lee 2021). However, automatic speech recogni- tion has become the de facto standard benchmark task (Ling and Liu 2020; Chung and Glass 2020a; Hsu et al. 2021).

Moving forward Most of the seminal work has focused on improving speech recognition (Schneider et al. 2019; Baevski et al. 2020). This focus has gained traction over the last couple of years, as computational resources have be- come more accessible and

end-to-end models (Graves et al. 2006; Chan et al. 2016) have been established as the dominant approach to speech recognition (Gulati et al. 2020). It is important to stress that self-supervised models, such as wav2vec 2.0 (Baevski et al. 2020), represent a breakthrough, and recent successful approaches build upon this method. That is, deep self-attention models combined with masking (Hsu et al. 2021; Wang et al. 2021; Chen et al. 2021). This development mirrors years of rapid progress in masked lan- guage modeling within natural language processing (Devlin et al. 2019; Clark et al. 2020) and we expect this to continue for unsupervised neural speech representation learning.

Thank you for your attention

Bibliography I

[1] Emre Aksan and Otmar Hilliges. “STCN: Stochastic Temporal Convolutional Networks”. In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. New Orleans, LA, USA, 2019 (cited on pages 79, 80).

[2] Alexander A. Alemi, Ian Fischer, and Joshua V. Dillon. *Uncertainty in the Variational Information Bottleneck*. 2018. arXiv: 1807.00906 (cited on page 50).

[3] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. *Data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language*. Facebook AI Research blog, 2022 (cited on page 80).

[4] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. “Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”. In: *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*. Conference on Neural Information Processing Systems. Virtual, 2020. arXiv: 2006.11477 (cited on page 80).

[5] Eivind Berge, William Whiteley, Heinrich Audebert, Gian Marco De Marchis, Ana Catarina Fonseca, Chiara Padiglioni, Natalia Pérez de la Ossa, Daniel Strbian, Georgios Tsivgoulis, and Guillaume Turc. “European Stroke Organisation (ESO) Guidelines on Intravenous Thrombolysis for Acute Ischaemic Stroke”. In: *European Stroke Journal* 6.1 (2021), pages I–LXII (cited on page 84).

[6] Christopher M. Bishop. “Novelty Detection and Neural-Network Validation”. In: *IEE Proceedings - Vision, Image and Signal Processing* 141.4 (1994), pages 217–222. ISSN: 1350245x, 13597108. DOI: 10.1049/ip-vis:19941330 (cited on pages 41, 42).

Bibliography II

[7] Stig Nikolaj Blomberg, Helle Collatz Christensen, Freddy Lippert, Annette Kjær Ersbøll, Christian Torp-Petersen, Michael R Sayre, Peter J Kudenchuk, and Fredrik Folke. "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), e2032320–e2032320 (cited on page 102).

[8] K Bohm and Lisa Kurland. "The Accuracy of Medical Dispatch - A Systematic Review". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 26 (2018), pages 1–10 (cited on page 85).

[9] Yuri Burda, Roger Grosse, and Ruslan R. Salakhutdinov. "Importance Weighted Autoencoders". In: *Proceedings of the 4th International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. San Juan, Puerto Rico, 2016, page 8 (cited on page 47).

[10] Adolf Buse. "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note". In: *The American Statistician* 36 (3a 1982), pages 153–157 (cited on page 46).

[11] Niki Carver, Vikas Gupta, and John E. Hipskind. "Medical Errors". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024. pmid: 28613514 (cited on pages 8, 9).

[12] Hyunsun Choi, Eric Jang, and Alexander A. Alemi. *WAIC, but Why? Generative Ensembles for Robust Anomaly Detection*. 2019. arXiv: 1810.01392 (cited on pages 50, 51).

[13] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. *An Unsupervised Autoregressive Model for Speech Representation Learning*. 2019. arXiv: 1904.03240 (cited on page 80).

Bibliography III

[14] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. "A Recurrent Latent Variable Model for Sequential Data". In: *Proceedings of the 29th Conference on Neural Information Processing Systems (NeurIPS)*. Conference on Neural Information Processing Systems. Montréal, Quebec, Canada, 2015, page 9 (cited on pages 79, 80).

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805 (visited on February 11, 2019). preprint (cited on pages 104, 105).

[16] Janek Ebbers, Jahn Heymann, Lukas Drude, Thomas Glarner, Reinhold Haeb-Umbach, and Bhiksha Raj. "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, pages 488–492. DOI: 10.21437/Interspeech.2017-1160 (cited on page 79).

[17] Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. "Sequential Neural Models with Stochastic Layers". In: *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*. Conference on Neural Information Processing Systems. Barcelona, Spain, 2016 (cited on pages 79, 80).

[18] GBD 2019 Stroke Collaborators et al. "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), pages 795–820. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(21)00252-0 (cited on page 84).

Bibliography IV

[19] Thomas Glarner, Patrick Hanebrink, Janek Ebbers, and Reinhold Haeb-Umbach. "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)*. Interspeech 2018. Hyderabad, India: ISCA, 2018, pages 2688–2692. DOI: 10.21437/Interspeech.2018-2148 (cited on page 79).

[20] Praveen Hariharan, Muhammad Bilal Tariq, James C Grotta, and Alexandra L Czap. "Mobile Stroke Units: Current Evidence and Impact". In: *Current Neurology and Neuroscience Reports* 22.1 (2022), pages 71–81 (cited on page 84).

[21] Dan Hendrycks and Kevin Gimpel. "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks". In: *Proceedings of the 5th International Conference on Learning Representations (ICRL)*. International Conference on Learning Representations. Toulon, France, 2017 (cited on page 50).

[22] Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. "Deep Anomaly Detection with Outlier Exposure". In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. New Orleans, LA, USA, 2019 (cited on pages 50, 51).

[23] Leora I. Horwitz, Grace Y. Jenq, Ursula C. Brewster, Christine Chen, Sandhya Kanade, Peter H. Van Ness, Katy L. B. Araujo, Boback Ziaeian, John P. Moriarty, Robert Fogerty, and Harlan M. Krumholz. "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), pages 436–443. ISSN: 1553-5592. DOI: 10.1002/jhm.2021. pmid: 23526813 (cited on pages 10, 11).

Bibliography V

[24] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units". In: (2021) (cited on page 80).

[25] Wei-Ning Hsu, Yu Zhang, and James Glass. *Learning Latent Representations for Speech Generation and Transformation*. 2017. arXiv: 1704.04222 (cited on pages 79, 80).

[26] Wei-Ning Hsu, Yu Zhang, and James Glass. "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data". In: *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Neural Information Processing Systems. Long Beach, CA, USA, 2017 (cited on pages 79, 80).

[27] Erik Joukes, Ameen Abu-Hanna, Ronald Cornet, and Nicolette De Keizer. "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), pages 046–053. ISSN: 1869-0327. DOI: 10.1055/s-0037-1615747 (cited on pages 10, 11).

[28] Mira Katan and Andreas Luft. "Global Burden of Stroke". In: *Seminars in Neurology*. Volume 38. 02. Thieme Medical Publishers, 2018, pages 208–211 (cited on page 84).

[29] Sameer Khurana, Shafiq Rayhan Joty, Ahmed Ali, and James Glass. "A Factorial Deep Markov Model for Unsupervised Disentangled Representation Learning from Speech". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom: IEEE, 2019, pages 6540–6544. ISBN: 978-1-4799-8131-1. DOI: 10.1109/ICASSP.2019.8683131 (cited on pages 79, 80).

Bibliography VI

[30] Sameer Khurana, Antoine Laurent, Wei-Ning Hsu, Jan Chorowski, Adrian Lancucki, Ricard Marxer, and James Glass. *A Convolutional Deep Markov Model for Unsupervised Speech Representation Learning*. 2020. arXiv: 2006.02547 (cited on pages 79, 80).

[31] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. Banff, AB, Canada, 2014. arXiv: 1312.6114 (cited on page 40).

[32] Hmwe Hmwe Kyu, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. "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), pages 1859–1922 (cited on page 84).

[33] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles". In: *In Proceddings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Conference on Neural Information Processing Systems. Long Beach, CA, USA, 2017 (cited on page 50).

[34] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. "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)*. Conference on Neural Information Processing Systems. Montréal, Quebec, Canada, 2018, page 11 (cited on pages 50, 51).

Bibliography VII

[35] Shiyu Liang, Yixuan Li, and R. Srikant. "Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks". In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. Vancouver, Canada, 2018 (cited on page 50).

[36] Shaoshi Ling and Yuzong Liu. "DeCoAR 2.0: Deep Contextualized Acoustic Representations with Vector Quantization". 2020. arXiv: 2012.06659 [cs, eess] (cited on page 80).

[37] Alexander H. Liu, Yu-An Chung, and James Glass. "Non-Autoregressive Predictive Coding for Learning Speech Representations from Local Dependencies". 2020. arXiv: 2011.00406 [cs] (cited on page 80).

[38] Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. "Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020, pages 6419–6423. DOI: 10.1109/ICASSP40776.2020.9054458 (cited on page 80).

[39] Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. "BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling". In: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*. Conference on Neural Information Processing Systems. Vancouver, Canada, 2019, pages 6548–6558 (cited on pages 40, 45).

[40] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. "Do Deep Generative Models Know What They Don't Know?" In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. New Orleans, LA, USA, 2019. arXiv: 1810.09136 (cited on pages 39, 53).

Bibliography VIII

[41] Babak B Navi, Heinrich J Audebert, Anne W Alexandrov, Dominique A Cadilhac, James C Grotta, and PRESTO (Prehospital Stroke Treatment Organization) Writing Group. "Mobile Stroke Units: Evidence, Gaps, and next Steps". In: *Stroke* 53.6 (2022), pages 2103–2113 (cited on page 84).

[42] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. *Representation Learning with Contrastive Predictive Coding*. 2018. arXiv: 1807.03748 (cited on page 80).

[43] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. "Neural Discrete Representation Learning". In: *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*. Conference on Neural Information Processing Systems. Long Beach, CA, USA, 2018 (cited on pages 79, 80).

[44] John Adam Oostema, Trevor Carle, Nadine Talia, and Mathew Reeves. "Dispatcher Stroke Recognition Using a Stroke Screening Tool: A Systematic Review". In: *Cerebrovascular Diseases* 42.5-6 (2016), pages 370–377 (cited on page 85).

[45] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. "Improving Language Understanding by Generative Pre-Training". In: (2018) (cited on pages 104, 105).

[46] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. "Likelihood Ratios for Out-of-Distribution Detection". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*. International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019, page 12 (cited on pages 50, 51).

Bibliography IX

[47] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *Proceedings of the 31st International Conference on Machine Learning (ICML)*. International Conference on Machine Learning. Volume 32. Beijing, China: PMLR, 2014, pages 1278–1286 (cited on page 40).

[48] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. "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 80).

[49] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. "Input Complexity and Out-of-Distribution Detection with Likelihood-Based Generative Models". In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020 (cited on page 51).

[50] Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. "Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties". In: *Annals of Internal Medicine* 165.11 (2016), pages 753–760. ISSN: 1539-3704. DOI: 10.7326/M16-0961. pmid: 27595430 (cited on pages 10, 11).

[51] Amy J Starmer, Nancy D Spector, Rajendu Srivastava, Daniel C West, Glenn Rosenbluth, April D Allen, Elizabeth L Noble, Lisa L Tse, Anuj K Dalal, Carol A Keohane, et al. "Changes in Medical Errors after Implementation of a Handoff Program". In: *New England Journal of Medicine* 371.19 (2014), pages 1803–1812 (cited on pages 8, 9).

Bibliography X

[52] Matthew D. Tipping, Victoria E. Forth, Kevin J. O'Leary, David M. Malkenson, David B. Magill, Kate Englert, and Mark V. Williams. "Where Did the Day Go?–A Time-Motion Study of Hospitalists". In: *Journal of Hospital Medicine* 5.6 (2010), pages 323–328. ISSN: 1553-5606. DOI: 10.1002/jhm.790. pmid: 20803669 (cited on pages 10, 11).

[53] Guillaume Turc, Pervinder Bhogal, Urs Fischer, Pooja Khatri, Kyriakos Lobotesis, Mikaël Mazighi, Peter D Schellinger, Danilo Toni, Joost De Vries, Philip White, et al. "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), pages 535–538 (cited on page 84).

[54] Rajeev Verma and Eric Nalisnick. "Calibrated Learning to Defer with One-vs-All Classifiers". In: *International Conference on Machine Learning*. PMLR, 2022, pages 22184–22202 (cited on pages 106, 107).

[55] Søren Viereck, Thea Palsgaard Møller, Helle Klingenberg Iversen, Hanne Christensen, and Freddy Lippert. "Medical Dispatchers Recognise Substantial Amount of Acute Stroke during Emergency Calls". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 24 (2016), pages 1–7 (cited on page 85).

[56] Zhisheng Xiao, Qing Yan, and Yali Amit. "Likelihood Regret: An Out-of-Distribution Detection Score for Variational Auto-Encoder". In: *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*. Conference on Neural Information Processing Systems. Virtual, 2020 (cited on pages 50, 51).

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.