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

FILL ME OUT.
Background on Corti, motivation for project.

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 [30].
- Preventability: A considerable fraction of all hospital admissions have preventable adverse outcomes[a] [8].

[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 [15, 29, 31][a].
- Varying quality: Discharge summaries almost never meet *all* timeline, transmission, and content criteria. [14][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 number of 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.

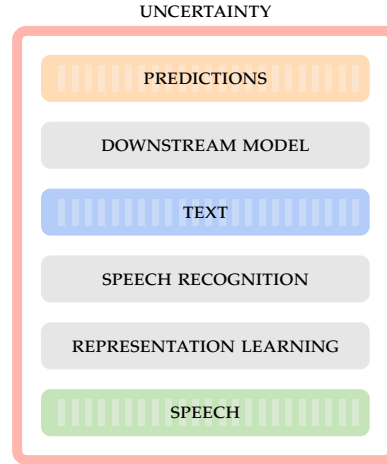
Reliability of machine learning systems

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

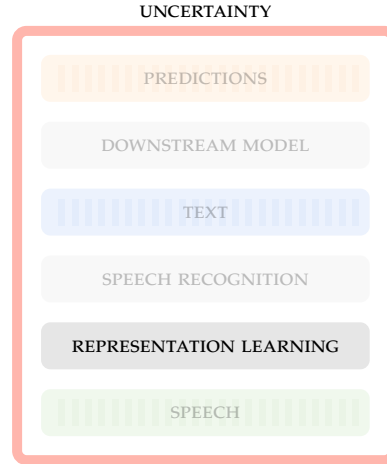
- Interpretability of how a model works (transparency, accountability, regulation).
- Explainability of model predictions (trust, understanding, feedback).
- Fairness in treatment of different groups of people.
- Robustness to noise, outliers, adversarial attacks, and distribution shift.

Thesis

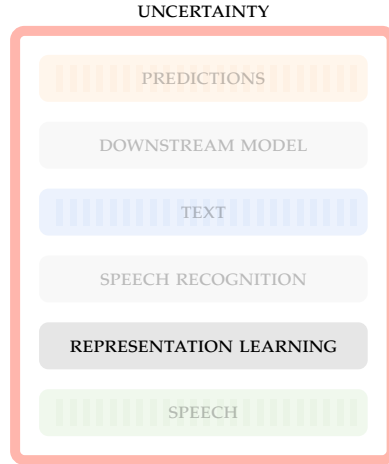
Chapter 1-3	**Introduction, Research Questions, and Background**
Chapter 4	**Hierarchical VAEs Know What They Don't Know**
Chapter 5	**Model-Agnostic Out-of-Distribution Detection Using Combined Statistical Tests**
Chapter 6	**A Brief Overview of Unsupervised Speech Representation Learning**
Chapter 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**



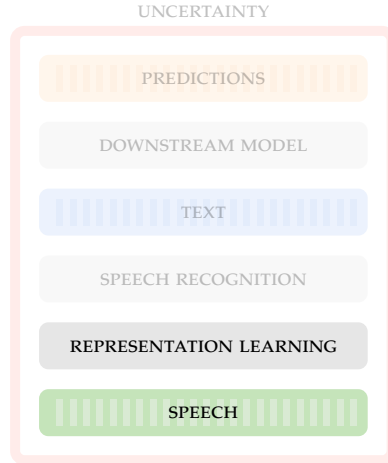
Thesis



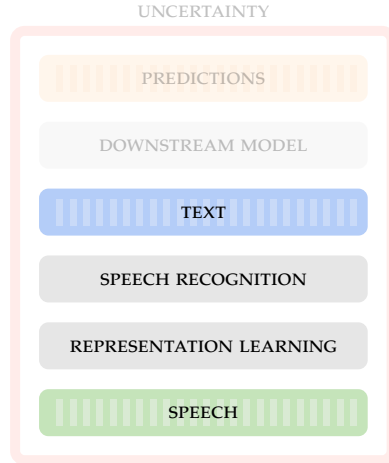
Thesis



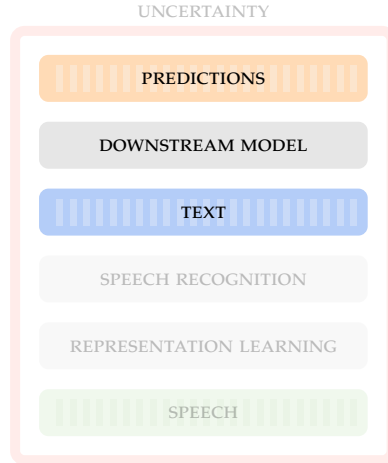
Thesis



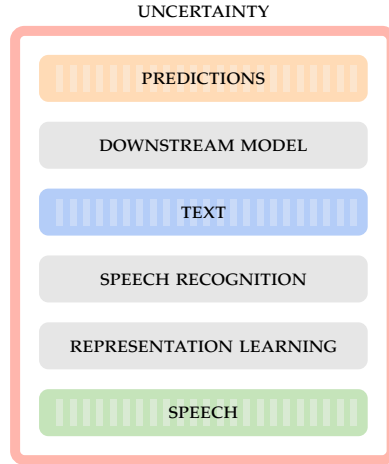
Thesis



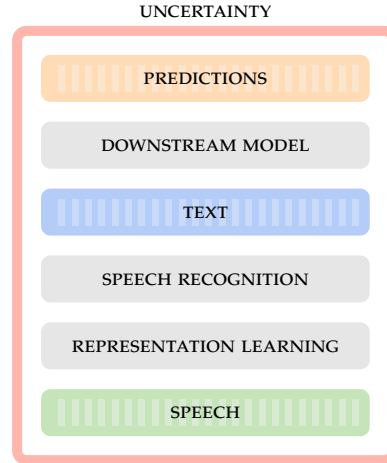
Thesis



Thesis

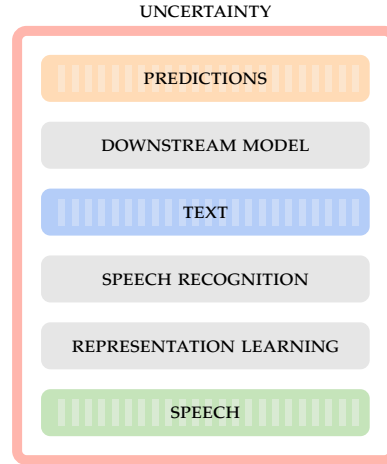


Presentation



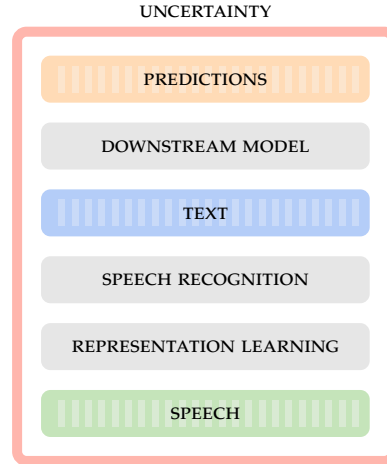
Presentation

Chapter 1-3	Introduction, Research Questions, and Background
Chapter 4	Hierarchical VAEs Know What They Don't Know
Chapter 5	Model-Agnostic Out-of-Distribution Detection Using Combined Statistical Tests
Chapter 6	A Brief Overview of Unsupervised Speech Representation Learning
Chapter 7	Benchmarking Latent Variable Models for Speech
Chapter 8	Automated Medical Coding on MIMIC-III and MIMIC-IV: A Critical Review and Replicability Study
Chapter 9	A Retrospective Study on Machine Learning-Assisted Stroke Recognition for Medical Helpline Calls
Chapter 10	Discussion and Conclusion

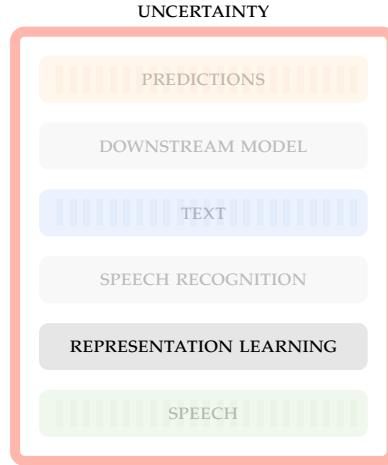


Presentation

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**



Presentation

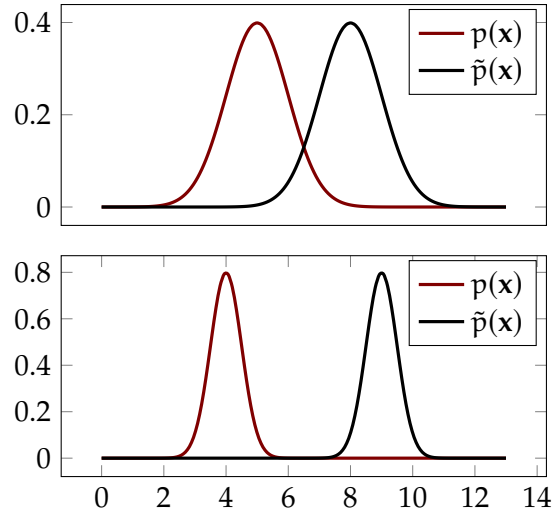


Defining OOD detection

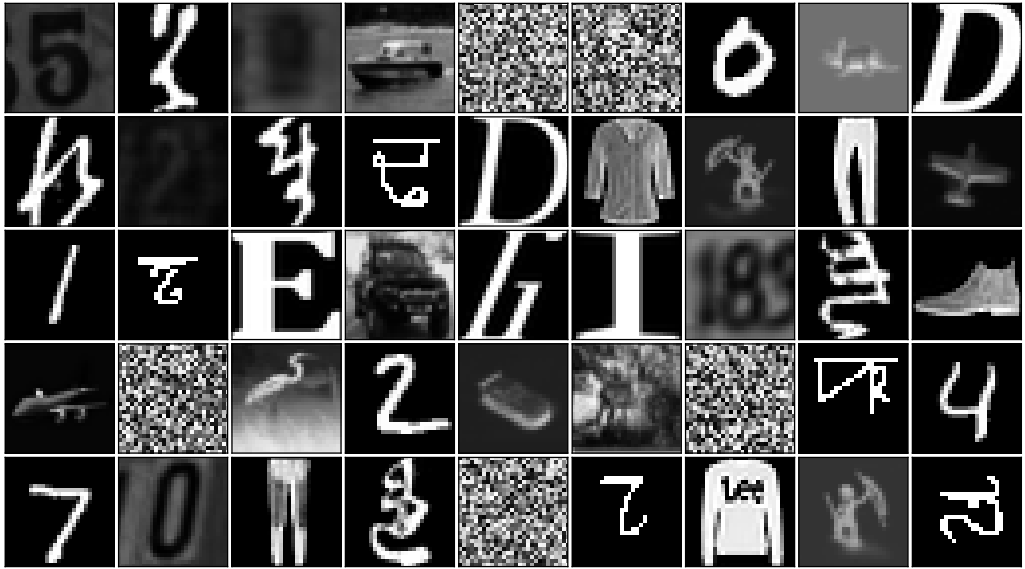
Enable models to distinguish the training data distribution $p(\mathbf{x})$ from any other distribution $\tilde{p}(\mathbf{x})$. Do this for any given

single observation, i.e. answer the question:

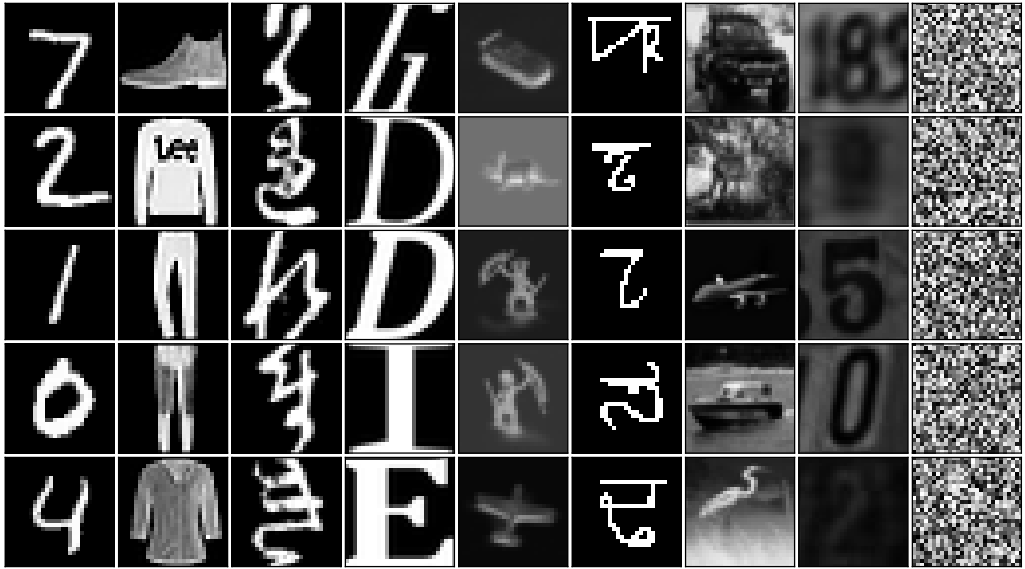
"Was $\mathbf{x}$ sampled from $p(\mathbf{x})$ or not?"



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 [23] 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 [17, 27].

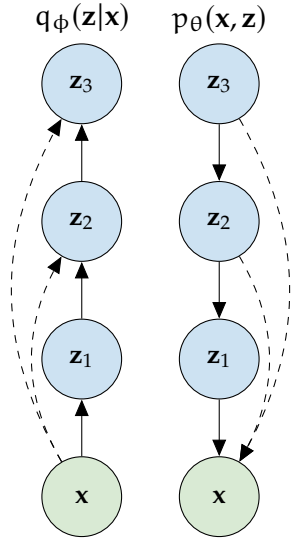
$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z}$$

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) [22].

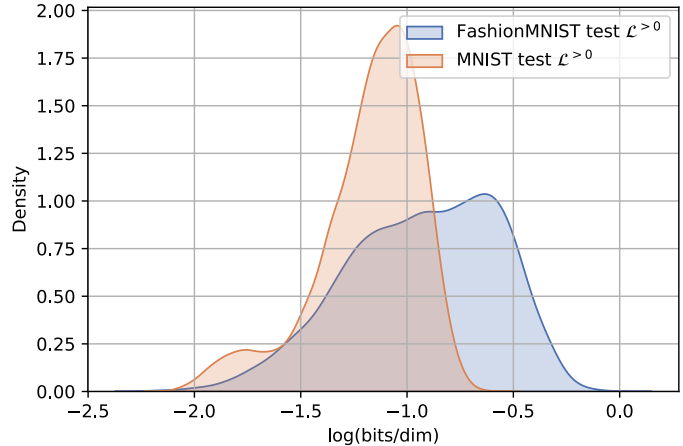


Out-of-distribution detection with hierarchical VAEs

- 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 [3].

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 [3].



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,

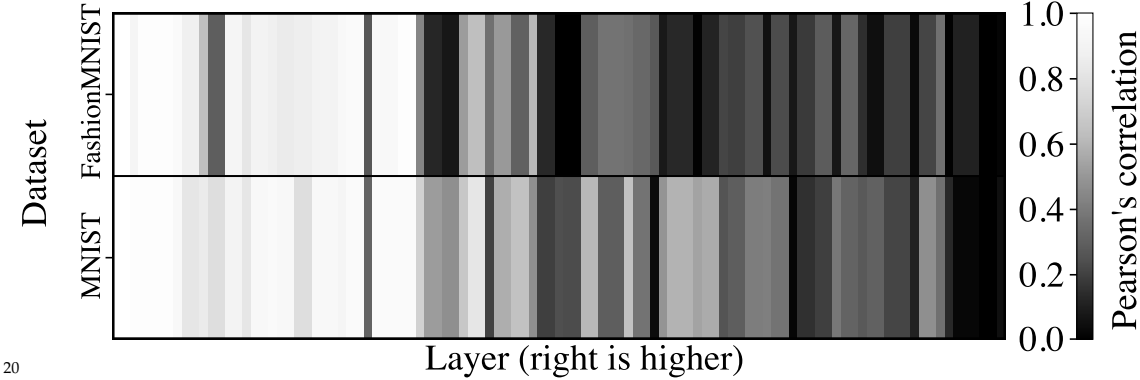
$$D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{i=1}^{L-1} \log \frac{p_\theta(\mathbf{z}_i|\mathbf{z}_{i+1})}{q_\phi(\mathbf{z}_i|\mathbf{z}_{i-1})} + \log \frac{p_\theta(\mathbf{z}_L)}{q_\phi(\mathbf{z}_L|\mathbf{z}_{L-1})}\right] . \quad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

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 [22]

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

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 [7].

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

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, $S \to \infty$ [6],

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

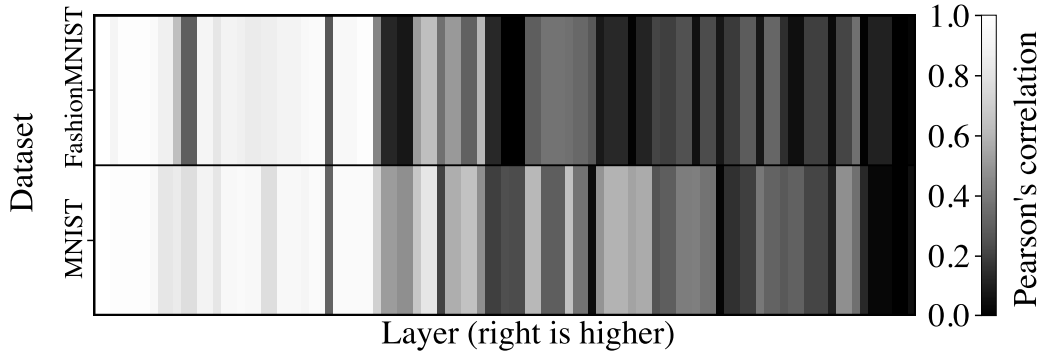
Consequently, by importance sampling the ELBO, the associated KL-divergence vanishes and our likelihood ratio reduces to the KL-divergence of $\mathcal{L}^{>k}$.

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

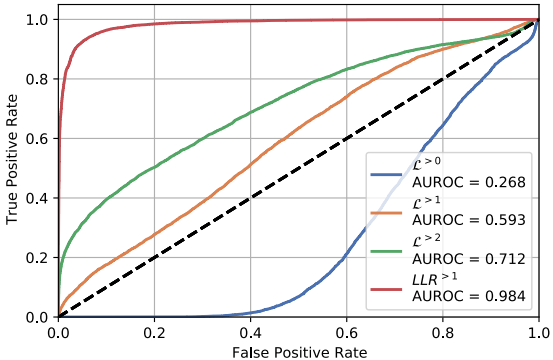
$\mathrm{LLR}_S^{>k}(\mathbf{x})$ performs KL-divergence-based OOD detection using top-most latent variables.

Selecting the value of k

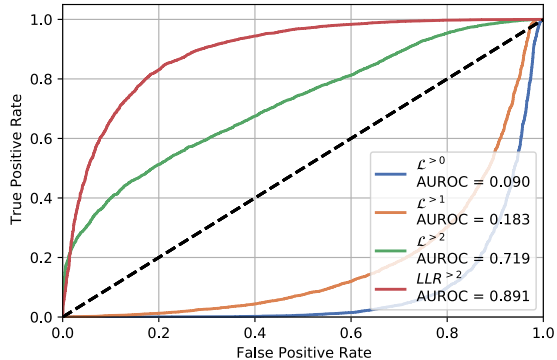
- Use validation OOD dataset(s).
- Compute $\text{LLR}^{>k}$ for different values of k and select the one that maximizes the AUROC.
- Compute feature correlations for different values of k and select k at the drop.



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



(a) FashionMNIST HVAE evaluated on MNIST



(b) CIFAR10 BIVA evaluated on SVHN

Results on FashionMNIST/MNIST

Method	AUROC↑	AUPRC↑	FPR80↓
FashionMNIST (in) / MNIST (out)			
Use prior knowledge of OOD			
Backgr. contrast. LR (PixelCNN) [26]	0.994	0.993	0.001
Backgr. contrast. LR (VAE) [9]	0.924	-	-
Binary classifier [26]	0.455	0.505	0.886
$p(\hat{y}\|\mathbf{x})$ with OOD as noise class [26]	0.877	0.871	0.195
$p(\hat{y}\|\mathbf{x})$ with calibration on OOD [26]	0.904	0.895	0.139
Input complexity (S, Glow) [13]	0.998	-	-
Input complexity (S, PixelCNN++) [13]	0.967	-	-
Use in-distribution data labels $\mathbf{y}$			
$p(\hat{y}\|\mathbf{x})$ [12, 26]	0.734	0.702	0.506
Entropy of $p(\mathbf{y}\|\mathbf{x})$ [26]	0.746	0.726	0.448
ODIN [21, 26]	0.752	0.763	0.432
VIB [1, 9]	0.941	-	-
Mahalanobis distance, CNN [26]	0.942	0.928	0.088
Mahalanobis distance, DenseNet [20]	0.986	-	-
Ensemble, 20 classifiers [19, 26]	0.857	0.849	0.240
No OOD-specific assumptions			
- Ensembles			
WAIC, 5 models, VAE [9]	0.766	-	-
WAIC, 5 models, PixelCNN [26]	0.221	0.401	0.911
- Not ensembles			
Likelihood regret [35]	**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) [26]	0.930	0.881	0.066
Backgr. contrast. LR (VAE) [35]	0.265	-	-
Outlier exposure [13]	0.984	-	-
Input complexity (S, Glow) [28]	0.950	-	-
Input complexity (S, PixelCNN++) [28]	0.929	-	-
Input complexity (S, HVAE) (Ours) [28]	0.833	0.855	0.344
Use in-distribution data labels y			
Mahalanobis distance [20]	0.991	-	-
No OOD-specific assumptions			
- Ensembles			
WAIC, 5 models, Glow [9]	1.000	-	-
WAIC, 5 models, PixelCNN [26]	0.628	0.616	0.657
- Not ensembles			
Likelihood regret [35]	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

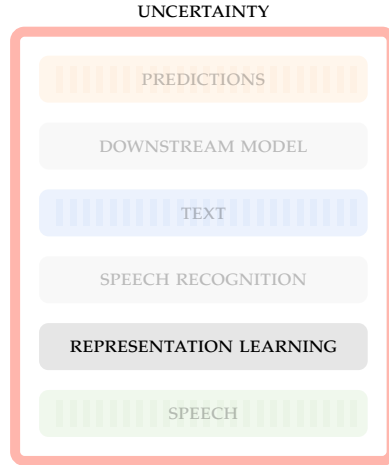
OOD dataset	Metric	AUROC↑	AUPRC↑	FPR80↓
Trained on FashionMNIST				
MNIST	$LLR^{>1}$	0.986	0.987	0.011
notMNIST	$LLR^{>1}$	0.998	0.998	0.000
KMNIST	$LLR^{>1}$	0.974	0.977	0.017
Omniglot28x28	$LLR^{>2}$	1.000	1.000	0.000
Omniglot28x28Inverted	$LLR^{>1}$	0.954	0.954	0.050
SmallNORB28x28	$LLR^{>2}$	0.999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

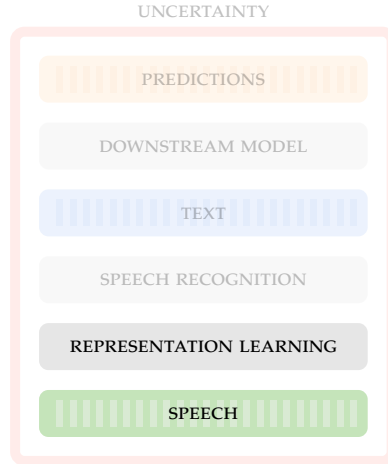
- Key observations:
  - The likelihood of a generative model is not a good score for OOD detection [23].
  - Strong correlations between some latent variables for different datasets.
  - Reconstructions of OOD data are good when using full approximate posterior.

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

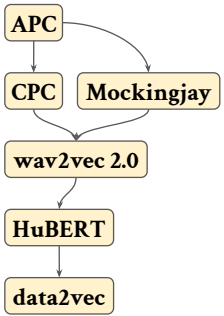
Presentation



Presentation

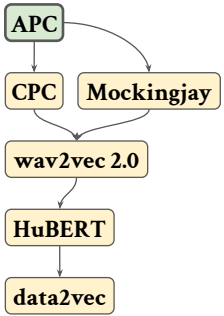


Development of SSL for speech

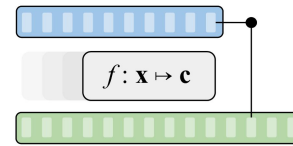


MODEL DESCRIPTION GOES HERE

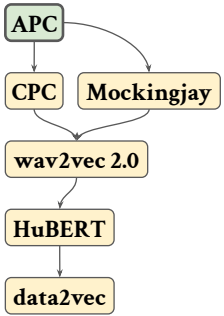
Autoregressive Predictive Coding (APC)



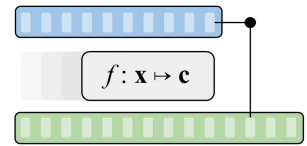
- **Task**: Predict future inputs.
- **Input/target**: Log-mel spectrogram.
- **Architecture**: RNN/Transformer decoder.
- **Slow features**: Predict k steps ahead.



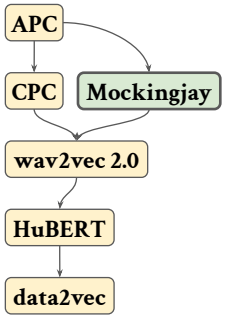
Autoregressive Predictive Coding (APC)



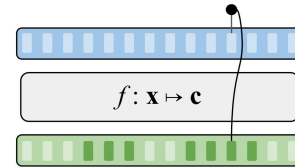
- **Challenges**:
  - Encodes only past inputs ✗
  - Uses the input as target ✗



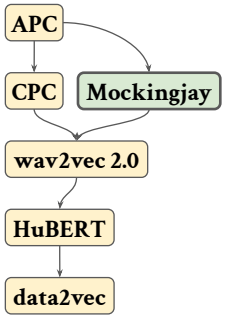
Mockingjay



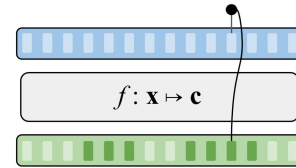
- **Task**: Reconstruct masked inputs.
- **Architecture**: Transformer encoder.
- **Masking**:
  - X% at random. (Mockingjay)
  - X% + N consecutive (wav2vec 2.0)
  - SpecAugment (Masked RNN)



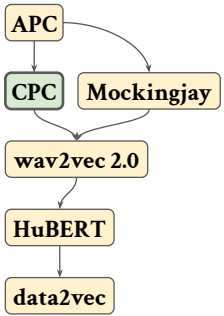
Mockingjay



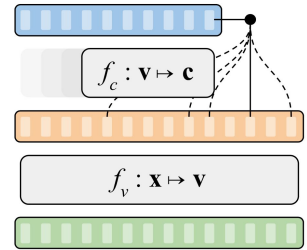
- **Challenges**:
  - Encodes the entire input ✓
  - Uses the input as target ✗



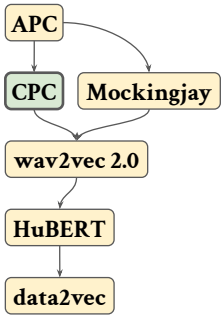
Contrastive Predictive Coding (CPC)



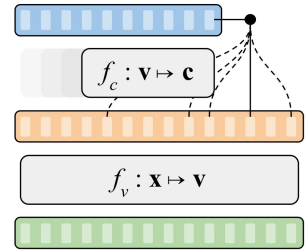
- **Contrastive models**: Distinguish target samples from negative samples.
- **Learned target**: Discard details.
- **Sampling negatives**:
  - Sample sequence?
  - Same speaker?



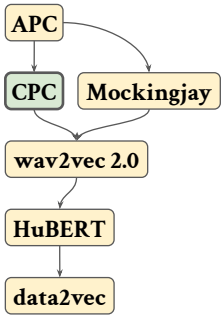
Contrastive Predictive Coding (CPC)



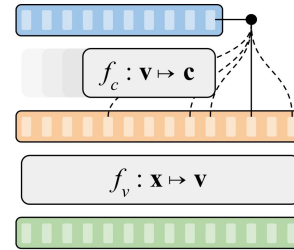
- **Challenges**:
  - Only encodes past inputs ✗
  - Uses a learned target ✓



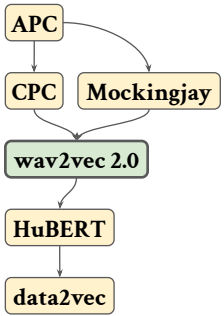
Contrastive Predictive Coding (CPC)



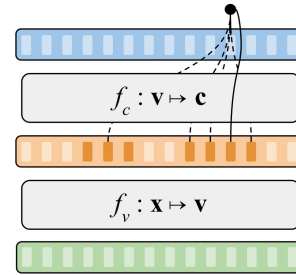
- **Challenges**:
  - Only encodes past inputs ✗
  - Uses a learned target ✓
  - Sampling negatives ✗



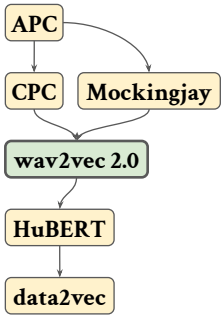
wav2vec 2.0



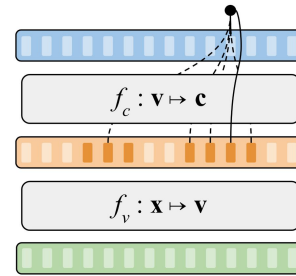
- Masking + contrastive learning.
- **Quantisation**: Better negative samples.
- **Results**:
  - 960 hours: **2.0%** WER.
  - 10 mimutes: **4.8%** WER.



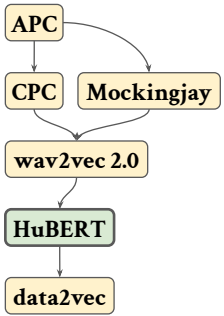
wav2vec 2.0



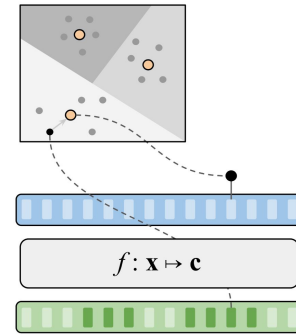
- **Challenges**:
  - Encodes the entire input ✓
  - Uses a learned target ✓
  - Sampling negatives ✗



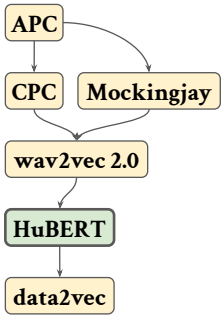
Hidden-unit BERT (HuBERT)



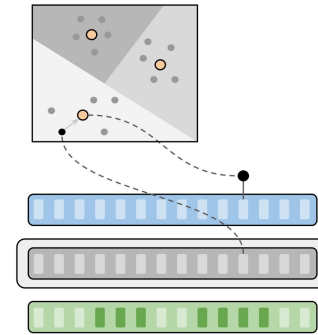
- **Target**: K-means teacher.
- **Training**: Simple cross-entropy loss.
- **1st iteration**: K-means on inputs.



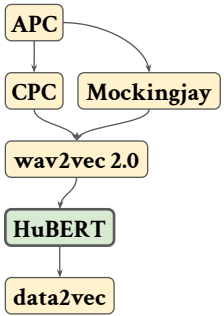
Hidden-unit BERT (HuBERT)



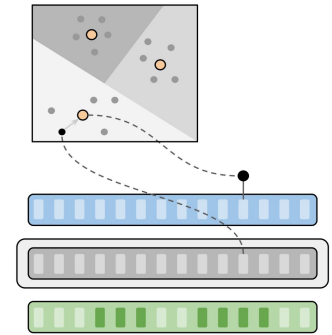
- **Target**: K-means teacher.
- **Training**: Simple cross-entropy loss.
- **1st iteration**: K-means on inputs.
- **2nd iteration**: K-means on hidden layers.



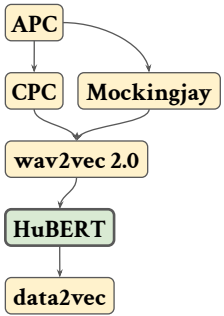
Hidden-unit BERT (HuBERT)



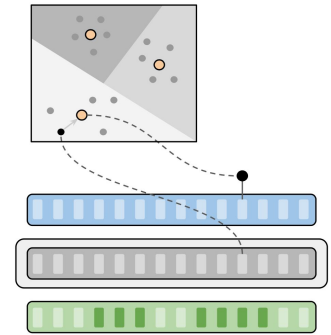
- **Challenges**:
  - Encodes the entire input ✓
  - Uses a learned target ✓
  - No need for negative samples ✓



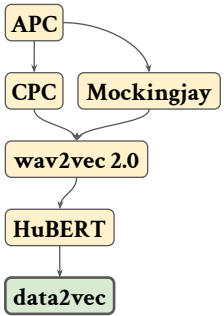
Hidden-unit BERT (HuBERT)



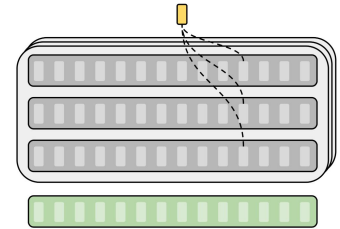
- **Challenges**:
  - Encodes the entire input ✓
  - Uses a learned target ✓
  - No need for negative samples ✓
  - Targets updated infrequently ✗
  - Quantized targets ✗



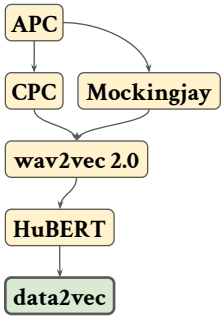
data2vec



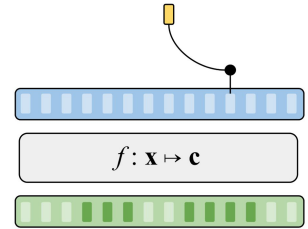
- Uses a teacher-student framework.
- **Teacher**:
  - EMA of student (online) ✓
  - Target is average of top K layers ✓



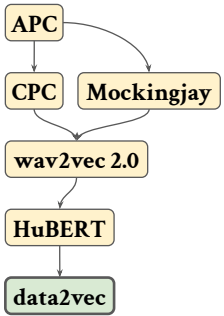
data2vec



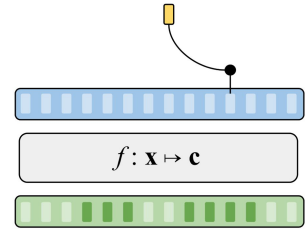
- Uses a teacher-student framework.
- **Teacher**:
  - EMA of student (online) ✓
  - Target is average of top K layers ✓
- **Student training**: Smooth $_1$ loss.



data2vec



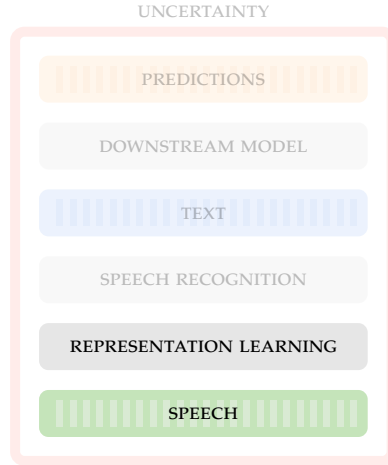
- **Challenges**:
  - Encodes the entire input ✓
  - Uses a learned target ✓
  - No need for negative samples ✓
  - Targets updated continuously ✓
  - Continuous-valued targets ✓



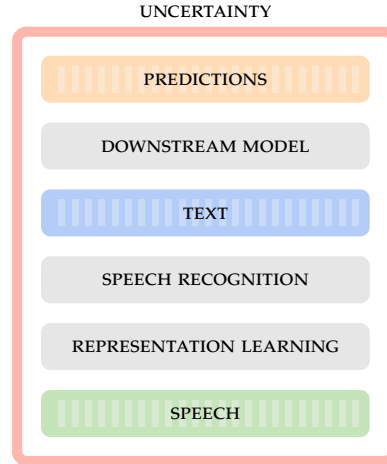
Conclusions

- **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

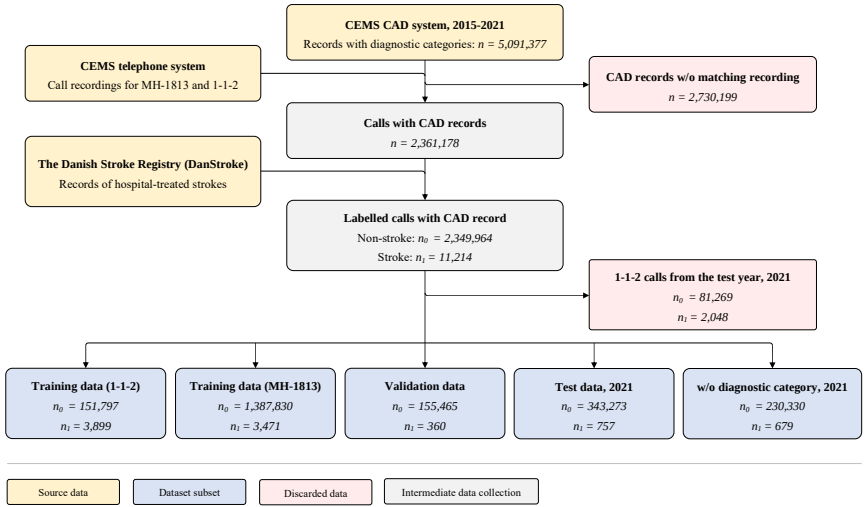
- Stroke is a leading cause of disability and death worldwide [10, 16, 18].
- Effective treatment is very time-sensitive. [2, 32].
- The gateway to ambulance transport and hospital admittance is through prehospital telehealth services.
- Mobile stroke units has made it possible to deliver advanced treatment faster [11, 24].
- The effectiveness of mobile stroke units hinges on call-taker recognition of stroke [11, 24].
- 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 [5, 25, 34].

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

Population selection and datasets

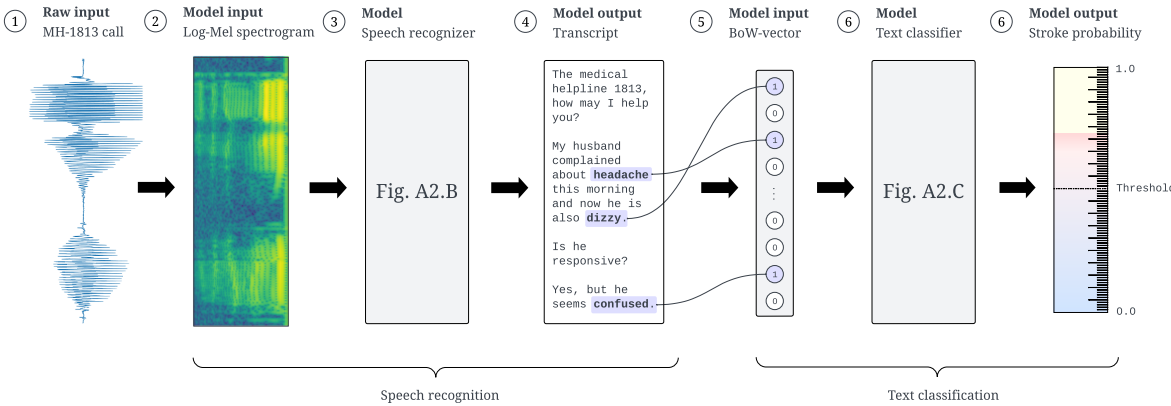


Population characteristics

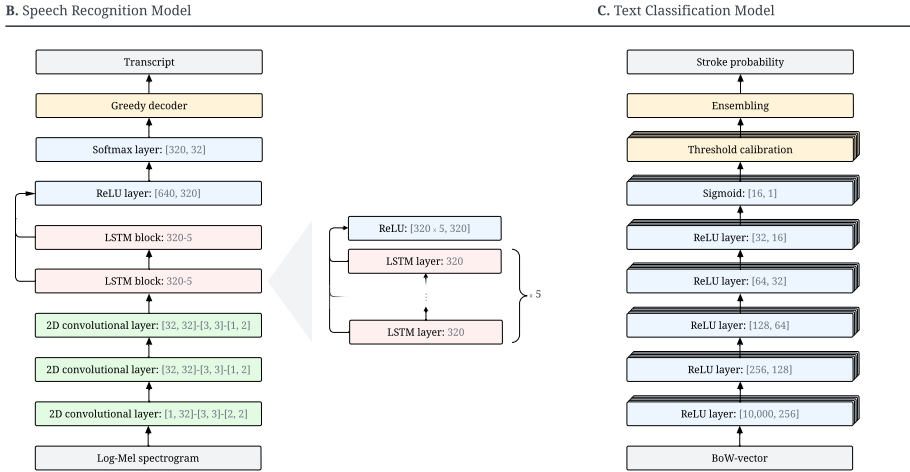
		Training (112)	Training (MH-1813)	Validation	Test	2021 w/o category
All calls	Num. calls	155,696	1,391,301	155,825	344,030	231,009
	Female	74,640 (47.94%)	792,783 (56.98%)	86,959 (55.81%)	190,974 (55.51%)	134,324 (58.14%)
	Male	79,564 (51.10%)	596,760 (42.89%)	68,866 (44.19%)	153,050 (44.49%)	96,258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

Table 1: 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 (7.61-11.18)	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)

Model performance

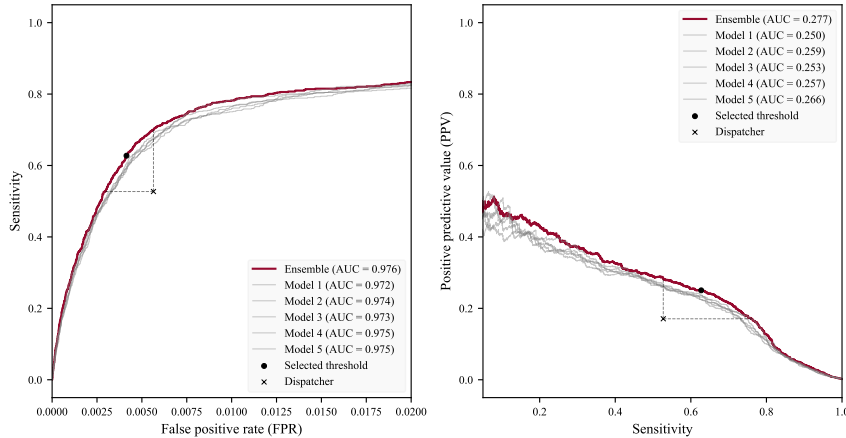


Figure 2: Left, the ROC curve and, right, PPV-sensitivity curve (precision-recall curve). Models 1-5 are the individual models that make up the ensemble model.

Model performance

Figure 3: 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?

	Positive ranking score, $r^{(w)}$		Negative ranking score, $r^{(w)}$	
	Stroke predictions, D = 1,897		Non-stroke predictions, D = 342,133	
	Word, *w (translated)*	Occurrences, $D^{(w)}$	Word, *w (translated)*	Occurrences, $D^{(w)}$
1.	Ambulance	1,680	Tetanus	4,378
2.	Blood clot	895	Pregnant	8,749
3.	Left	1,108	Cut	7,592
4.	Right	1,050	Bandage	4,561
5.	Double vision	84	Amager (a location)	23,776
6.	The words	344	O'clock	94,436
7.	Suddenly	783	The emergency room	42,809
8.	Arm	709	The police	2,903
9.	Side	1,139	Swollen	60,559
10.	Stroke	117	Over the counter (OTC)	4,641
11.	Double	113	The neck	30,151
12.	Control	134	Fever	112,586
13.	Call	39	Prescription	5,450

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 [4].

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)

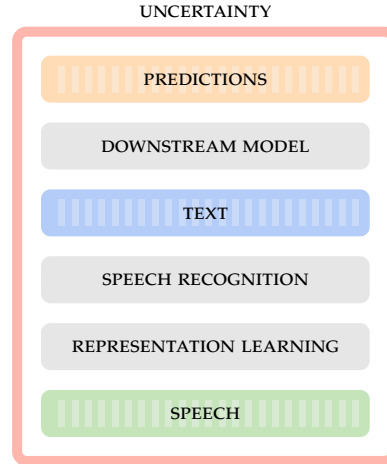
Fine-tuning a large language model

	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)

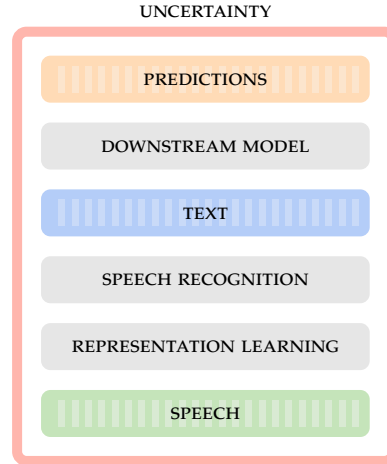
Future work

- Self-supervised learning directly from audio data.
- Investigate learning to defer to predict methods [33].

Presentation



Presentation



First slide for discussion

Building an operational decision support system

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

Thank you for your attention

Bibliography I

[1] Alexander A. Alemi, Ian Fischer, and Joshua V. Dillon. *Uncertainty in the Variational Information Bottleneck*. 2018. arXiv: 1807.00906 (cited on page 45).

[2] 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 71).

[3] 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 36, 37).

[4] 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 82).

[5] 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 72).

[6] 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 42).

Bibliography II

[7] 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 41).

[8] 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).

[9] Hyunsun Choi, Eric Jang, and Alexander A. Alemi. *WAIC, but Why? Generative Ensembles for Robust Anomaly Detection*. 2019. arXiv: 1810.01392 (cited on pages 45, 46).

[10] 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 71).

[11] 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 71).

[12] 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 45).

[13] 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 45, 46).

Bibliography III

[14] 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).

[15] 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).

[16] 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 71).

[17] 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 35).

[18] 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 71).

Bibliography IV

[19] 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 45).

[20] 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 45, 46).

[21] 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 45).

[22] 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 35, 40).

[23] 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 34, 48).

Bibliography V

[24] 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 71).

[25] 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 72).

[26] 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 45, 46).

[27] 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 35).

[28] 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 46).

Bibliography VI

[29] 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).

[30] 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).

[31] 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).

[32] 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 71).

[33] 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 page 85).

Bibliography VII

[34] 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 72).

[35] 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 45, 46).