Unsupervised patient representations with interpretable classification decisions

Madhumita Sushil, Simon Šuster, Kim Luyckx, Walter Daelemans

CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp

UZA

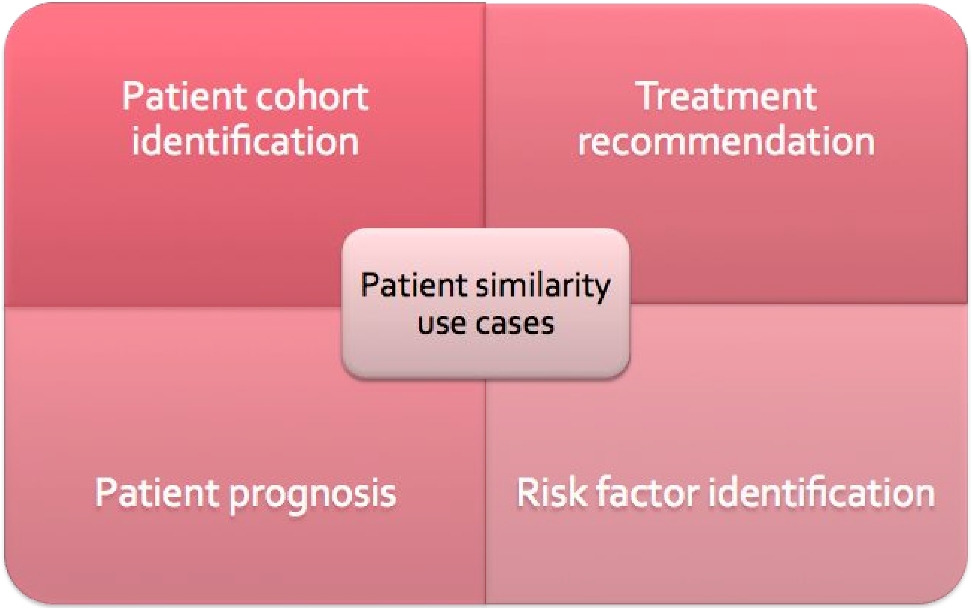
Patient Representations

Task-independent generalized semantic representations of patients, s.t.

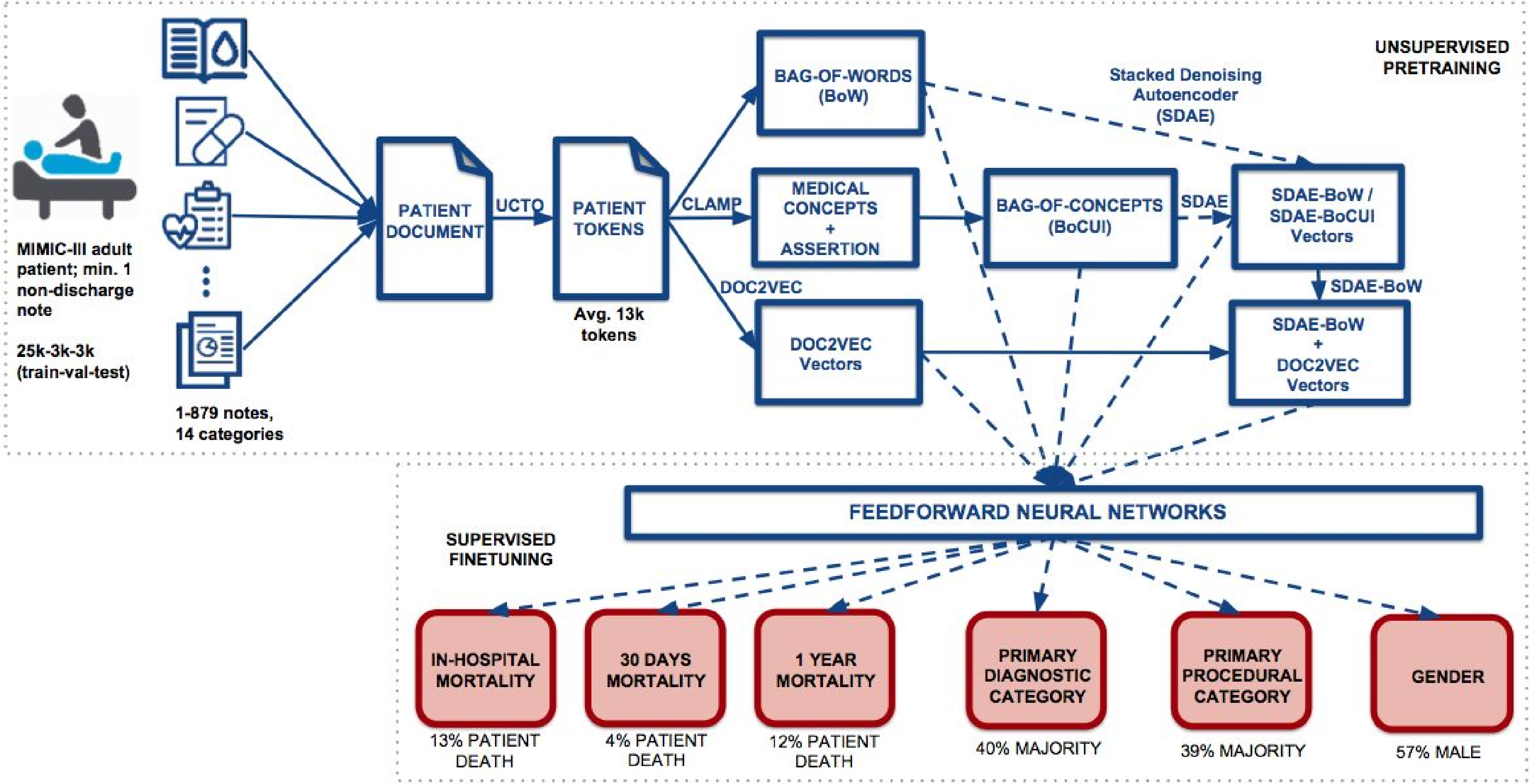
Similar patients - similar representations

Patient similarity encompasses holistic patient condition

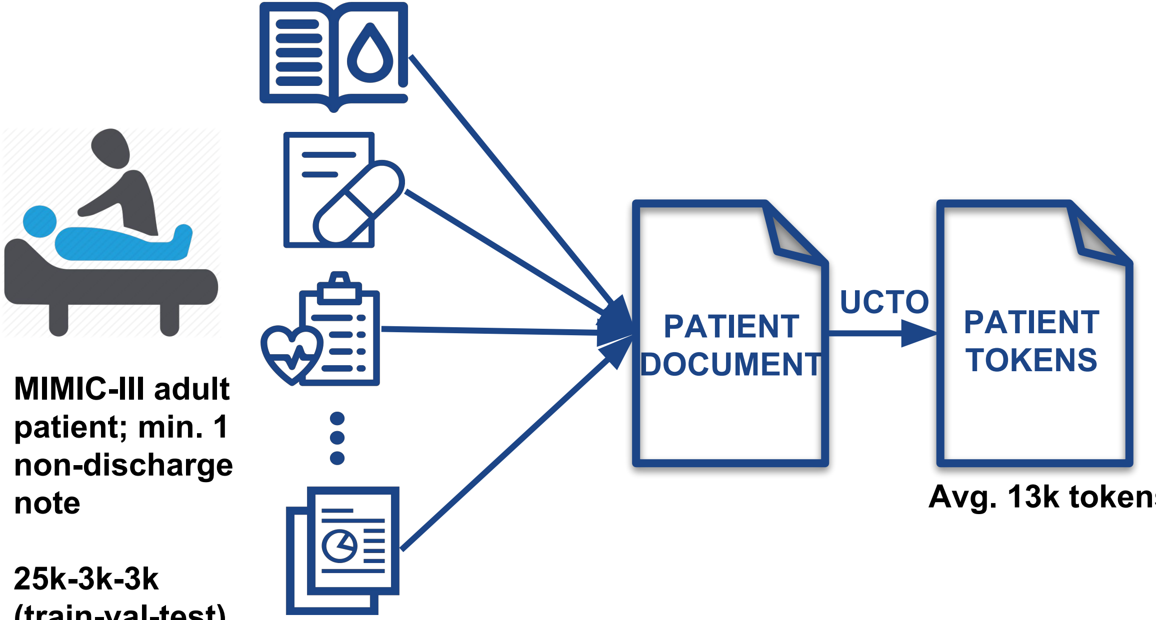
Patient Similarity



Pipeline - Representation Learning and Evaluation



Data



MIMIC-III adult patient; min. 1 non-discharge note

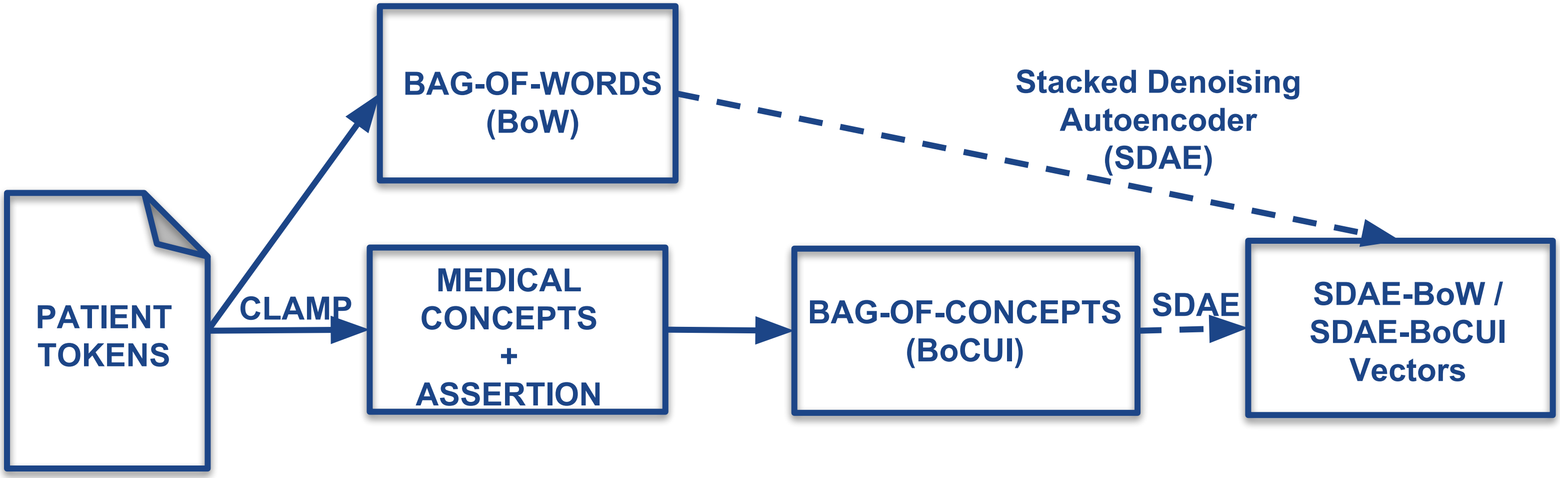
25k-3k-3k (train-val-test)

1-879 notes, 14 categories

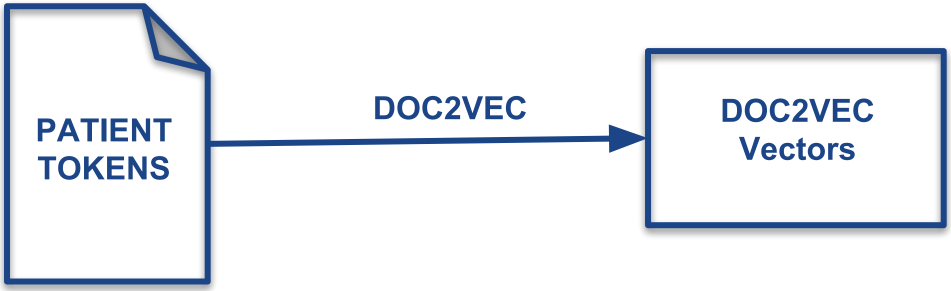
Avg. 13k tokens

icons by pictohaven, LAFS, ProSymbols, Lucas Almeida, Juraj Sedlák; thenounproject.com

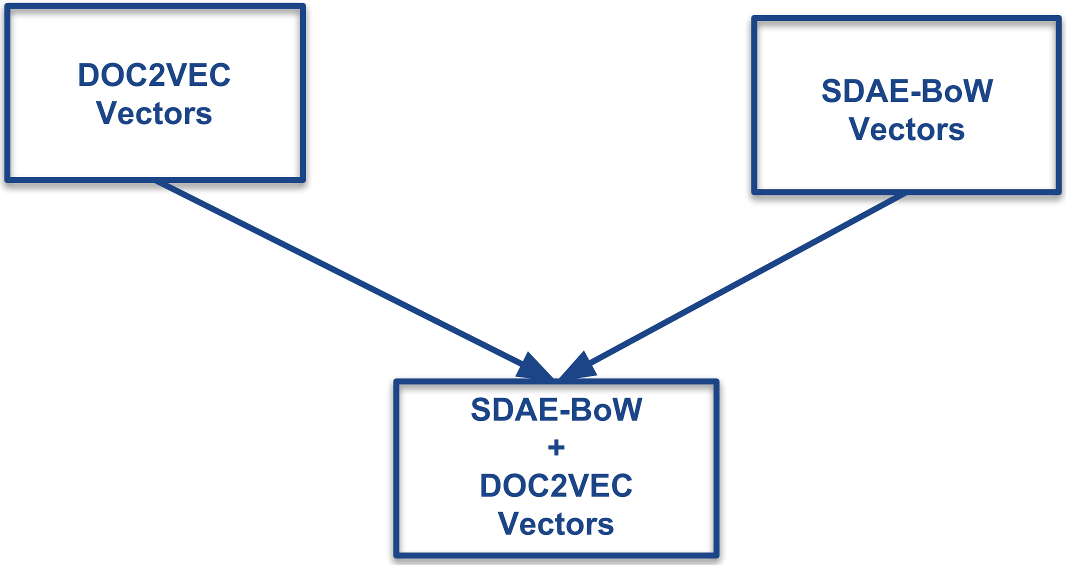
Unsupervised Representation Learning



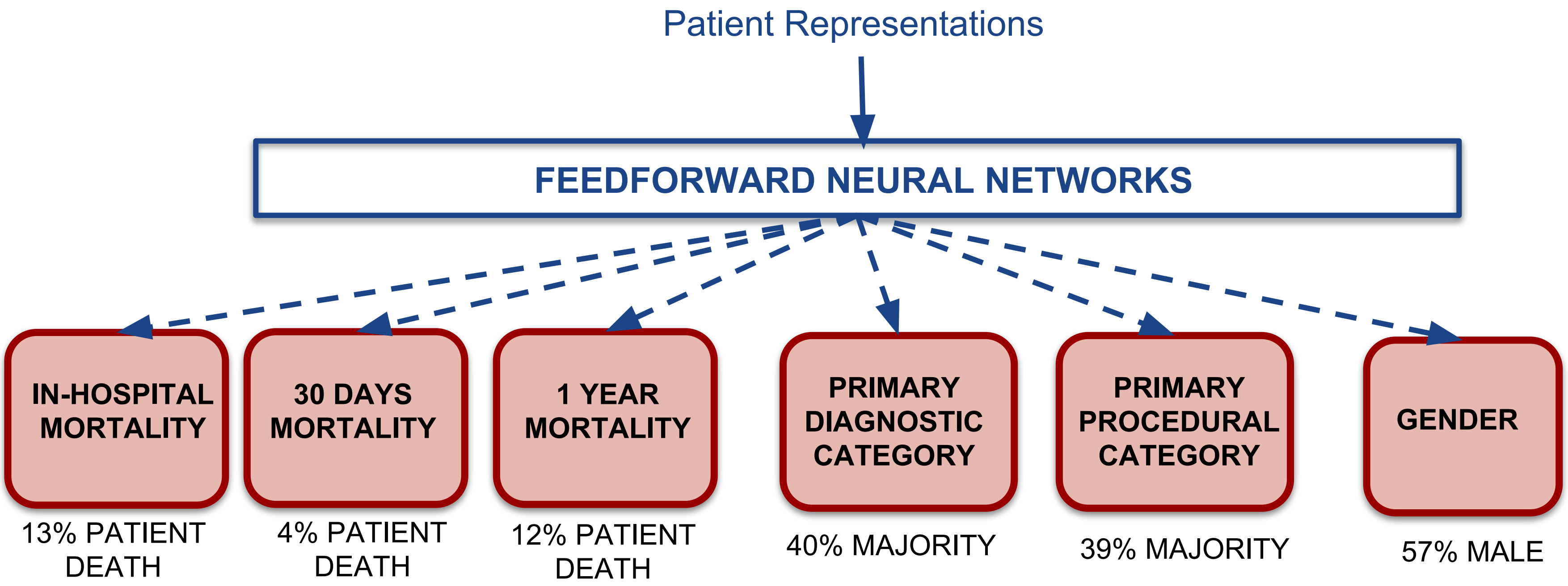
Unsupervised Representation Learning



Unsupervised Representation Learning



Supervised Representation Evaluation



Results - Representation Evaluation

Approach	In_hosp (AUC)	30_days (AUC)	1_year (AUC)	Pri_diag_cat (F-score-wt)	Pri_proc_cat (F-score-wt)	Gender (F-score-wt)
BoW	94.57	59.49	79.42	70.16	73.66	98.47
SDAE-BoW	91.94	79.65	79.80	65.00	67.46	87.75
Doc2vec	91.95	76.80	81.34	68.07	65.83	97.70
SDAE-BoW + Doc2vec	93.83	81.13	83.02	67.88	70.30	97.47

All generalized representation models **significantly outperform** sparse models **when no. of positive instances is low** (30 days mortality).

Results - Representation Evaluation

Approach	In_hosp (AUC)	30_days (AUC)	1_year (AUC)	Pri_diag_cat (F-score-wt)	Pri_proc_cat (F-score-wt)	Gender (F-score-wt)
BoW	94.57	59.49	79.42	70.16	73.66	98.47
SDAE-BoW	91.94	79.65	79.80	65.00	67.46	87.75
Doc2vec	91.95	76.80	81.34	68.07	65.83	97.70
SDAE-BoW + Doc2vec	93.83	81.13	83.02	67.88	70.30	97.47

For all tasks except distant patient mortality, BoW model is a strong baseline (strong lexical features present).

Results - Representation Evaluation

Approach	In_hosp (AUC)	30_days (AUC)	1_year (AUC)	Pri_diag_cat (F-score-wt)	Pri_proc_cat (F-score-wt)	Gender (F-score-wt)
BoW	94.57	59.49	79.42	70.16	73.66	98.47
SDAE-BoW	91.94	79.65	79.80	65.00	67.46	87.75
Doc2vec	91.95	76.80	81.34	68.07	65.83	97.70
SDAE-BoW + Doc2vec	93.83	81.13	83.02	67.88	70.30	97.47

Recommended to combine SDAE and doc2vec vectors
for unknown tasks.

Model Interpretability

Critical for

- Error analysis
- Exploratory analysis

Goal: Quantitative explanation of a trained model without retraining

Feature Extraction - Pretraining Autoencoders

Rank features according to **mean squared feature reconstruction error across all instances**

Feature Extraction - Pretraining Autoencoders

Rank features according to **mean squared feature reconstruction error across all instances**

Best Reconstruction	Worst Reconstruction
stumnz	picc
jajhnx	woman
a-fibril	osh
lsc.o	fall
potentiallly	man
yesh	stent
forcal	he
contbributing	wife
hyponatremia-on	repair
pre-exiusting	bleed

Feature Extraction - Pretraining Autoencoders

Rank features according to **mean squared feature reconstruction error across all instances**

0.87-0.88 Spearman correlation coefficient between feature reconstruction error and frequency

Feature entropy too high?

Best Reconstruction	Worst Reconstruction
stumnz	picc
jajhnx	woman
a-fibril	osh
lsc.o	fall
potentiallly	man
yesh	stent
forcal	he
contbributing	wife
hyponatremia-on	repair
pre-exiusting	bleed

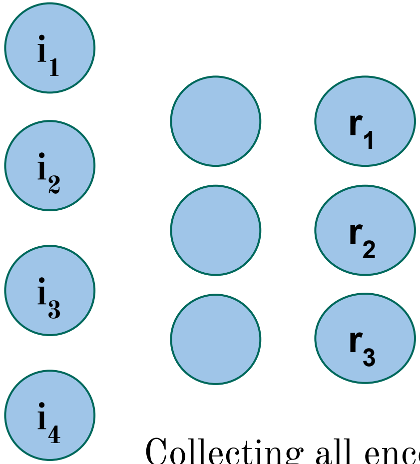
Finding Influential features - Classification Phase

Gradient-based feature sensitivity calculation across two neural networks

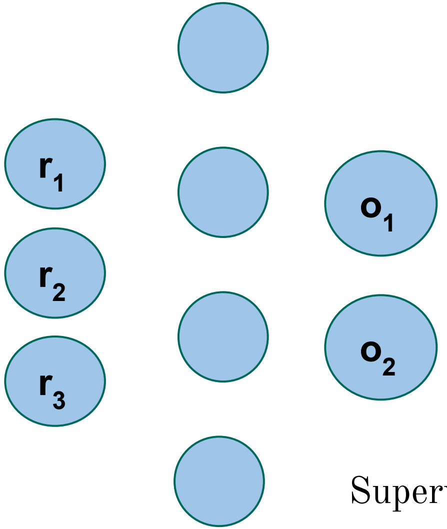
For arbitrary number of instances and output classes

Transferable to different representation learning architectures

Feature Extraction - Classification Phase



Collecting all encoder layers of SDAE

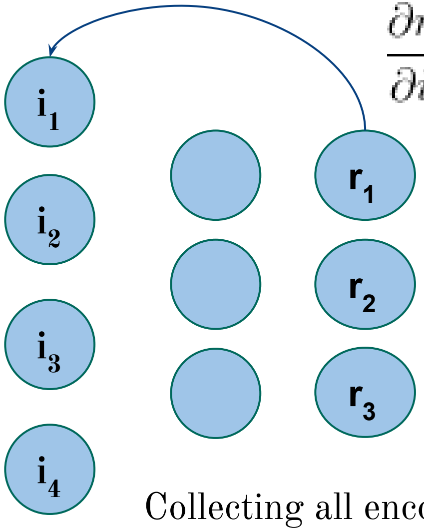


Supervised classifier

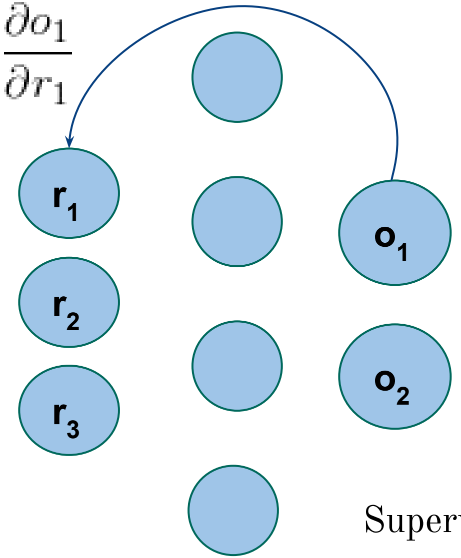
Feature Extraction - Classification Phase

$$S_{o1,\, i1}^{(1)} = \frac{\partial r_1}{\partial i_1} * \frac{\partial o_1}{\partial r_1}$$

Sensitivity



Collecting all encoder layers of SDAE

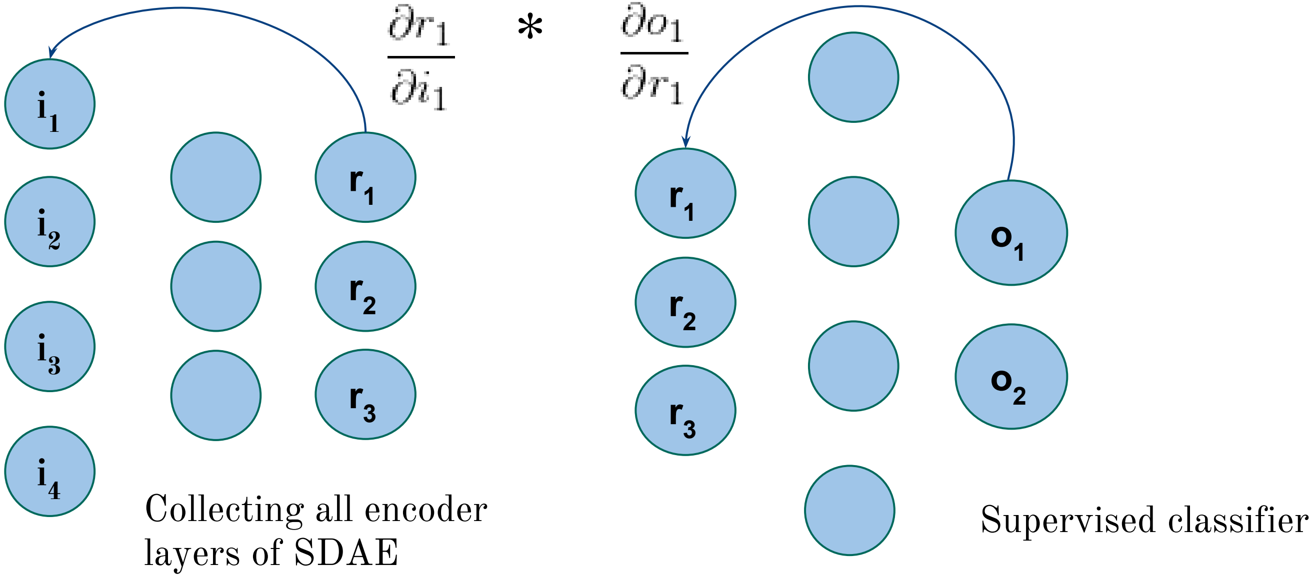


Supervised classifier

Feature Extraction - Classification Phase

$$S_{o1,\ i1}^{(1)} = \frac{\partial r_1}{\partial i_1} * \frac{\partial o_1}{\partial r_1}$$

Sensitivity



Feature influence = Maximum root mean squared sensitivity across all instances

Most influential features for one instance each

In_hosp	30_days	1_year	Pri_diag_cat	Pri_proc_cat	Gender
vasopressin	leaflet	magnevist	numeric_val	numeric_val	woman
pressors	structurally	signal	previous	no	female
focused	pacemaker	decisions	rhythm	of	she
dnr	sda	periventricular	no	enzymes	man
dopamine	periventricular	embolus	flexure	extubated	he
acidosis	excursion	underestimated	dementia	rhythm	male
levophed	non-coronary	calcified	brbpr	and	her
pressor	dosages	screws	of	the	his
cvvhd	microvascular	rib	sinus	vent	wife
cvvh	left-sided	shadowing	for	uncal	uterus
emergency	chronic	gadolinium	to	mso	him

Conclusions

Generalized patient representations help for low number of positive instances

Recommended to combine autoencoder and doc2vec representations

During representation learning, autoencoder has high reconstruction error for frequent terms, perhaps due to their high entropy

The most influential features for classification however are frequent terms, and context of features plays a role.

Unsupervised patient representations from clinical notes with interpretable classification decisions

Madhumita Sushil, Simon Šuster, Kim Luyckx, Walter Daelemans

Workshop on Machine Learning for Health, NIPS 2017.

THANK YOU!

Results

Approach	In_hosp (AUC)	30_days (AUC)	1_year (AUC)	Pri_diag_cat (F-score-wt)	Pri_proc_cat (F-score-wt)	Gender (F-score-wt)
BoW	94.57	59.49	79.42	70.16	73.66	98.47
SDAE-BoW	91.94	79.65	79.80	65.00	67.46	87.75
Doc2vec	91.95	76.80	81.34	68.07	65.83	97.70
SDAE-BoW + Doc2vec	93.83	81.13	83.02	67.88	70.30	97.47
BoCUI	90.88	50.65	69.93	71.04	72.65	75.04
SDAE-BoCUI	90.07	78.32	80.16	66.47	67.77	62.45