A General Framework for Visualizing Embedding Spaces of Neural Survival Analysis Models Based on Angular Information

George H. Chen (georgechen@cmu.edu)

Carnegie Mellon University Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Code: https://github.com/georgehc/anchor-vis

Motivation

- Many neural survival models are not straightforward to interpret: DeepSurv [Katzman et al 2018], DeepHit [Lee et al 2018], Neural MTLR [Fotso 2018], Extended Hazards [Zhong et al 2021], …

How can we make sense of representations learned by such models?

- While many visualization strategies have been developed for classification (e.g., saliency maps), these strategies do not easily extend to survival analysis

Main contribution: new framework for visualizing *any* intermediate embedding representation of *any* neural survival model

- Based on **angular information** (ignores embedding vector magnitudes and only considers their directions)
- Identifies **anchor directions** in an intermediate embedding space & relates these directions to raw features & survival time distributions
- Comes with statistical tests that can optionally be applied
- Serves as a debugging tool & provides info not conveyed in a t-SNE plot

Disclaimer: our framework focuses on a single embedding space at a time & does *not* "explain" how the overall neural survival model actually makes predictions

Remark: there are neural survival models designed to have interpretable components

- Clustering structure [Chapfuwa et al 2020, Nagpal et al 2021, …]
- Topic modeling structure [Li et al 2020]

Our framework can also be applied to such models to provide supplemental plots & statistical analyses

Survival Analysis

Training data (i.i.d.): $\{(x_i, y_i, \delta_i)\}_{i=1}^n$

raw input; observed time ≥ 0; event indicator: $1 \Rightarrow y_i$ is survival time, $0 \Rightarrow y_i$ is censoring time

Example model: DeepSurv [Faraggi & Simon 1995, Katzman et al 2018]

- Predict log partial hazard function $f(x;\theta) \in \mathbb{R}$ (model as neural net; θ: neural net parameters; x: raw input)
- Standard loss for learning parameters:

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^n \delta_i \Big[f(x_i;\theta) - \log \sum_{\substack{j=1,\dots,n \\ \text{s.t } y_j \geq y_i}} \exp(f(x_j;\theta))\Big]$$

Classical Cox model [Cox 1972]
$f(x;\theta) = \langle \theta, x \rangle$
$x, \theta \in \mathbb{R}^D$

Conditional survival function estimate

- Let $\widehat{\theta}$ denote the estimate of θ minimizing the loss above
- Sort unique times when critical event (e.g., death) happened: $t_{(1)} < t_{(2)} < \cdots < t_{(\tau)}$
- Let $d_{(\ell)}$ be # of critical events that happened at time $t_{(\ell)}$ for $\ell = 1, \dots, \tau$

$$\widehat{S}(t|x) = \exp\Big\{-\exp(f(x;\widehat{\theta})) \sum_{\substack{\ell=1,\dots,\tau \\ \text{s.t. } t_{(\ell)} \leq t}} \boxed{\widehat{h}_0(t_{(\ell)})}\Big\}$$

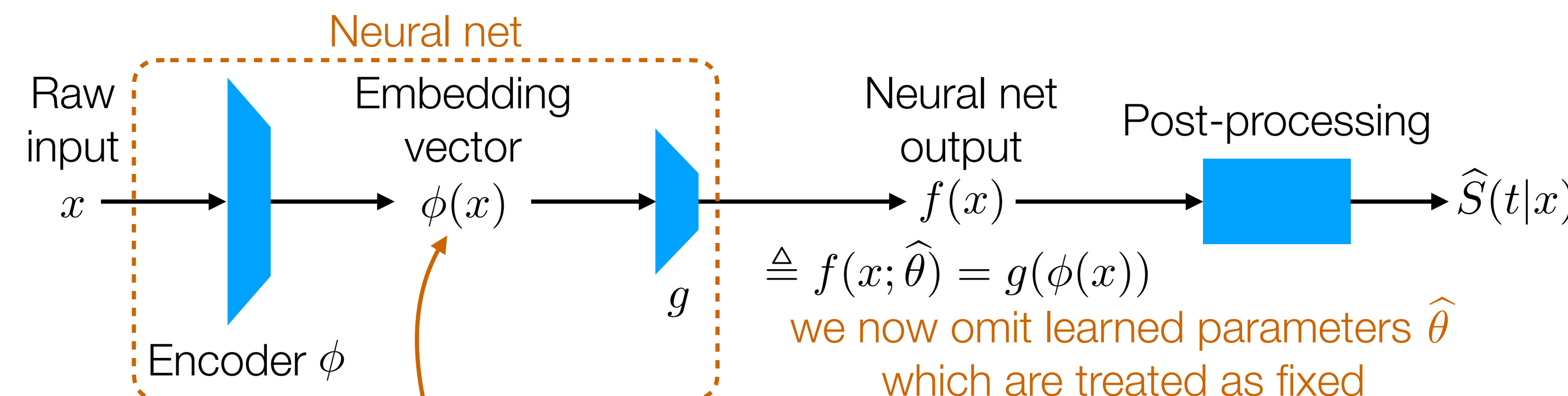
estimated prob. of surviving beyond time t given x

$$\widehat{h}_0(t_{(\ell)}) = \frac{d_{(\ell)}}{\sum_{\substack{j=1,\dots,n \\ \text{s.t. } y_j \geq t_{(\ell)}}} \exp(f(x_j;\widehat{\theta}))}$$

Our framework works with any neural survival analysis model—not just DeepSurv

Framework

- Assume neural survival model has already been trained
- DeepSurv (and other neural survival models) can be depicted as:

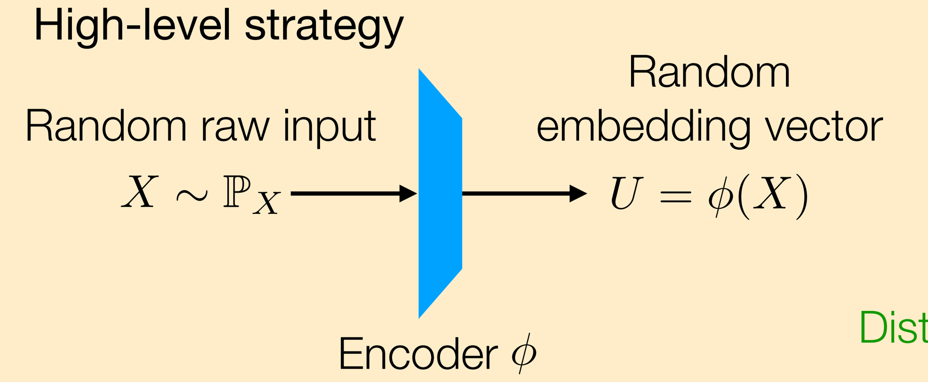

Neural net
Raw input x
Embedding vector $\phi(x)$
Encoder ϕ
g
Neural net output $f(x)$
Post-processing
$\widehat{S}(t|x)$
$\triangleq f(x;\widehat{\theta}) = g(\phi(x))$
we now omit learned parameters $\widehat{\theta}$ which are treated as fixed

Intermediate embedding space to be visualized (modeler can choose precisely which initial layers of f to treat as the encoder)

We treat learned neural net f as fixed & similarly encoder ϕ also as fixed

Example:
- f is a multilayer perceptron with ReLU activations
- We take ϕ to be everything in f prior to the last linear layer

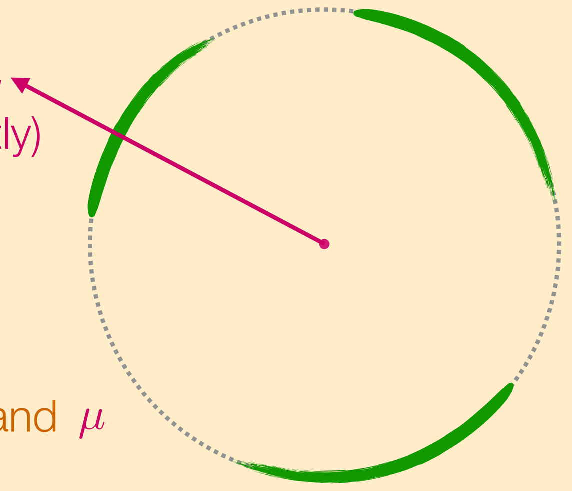
High-level strategy


Random raw input
$X \sim \mathbb{P}_X$
Encoder ϕ
Random embedding vector
$U = \phi(X)$

Consider the normalized version

$$\widetilde{U} = \frac{U - \mathbb{E}[U]}{\|U - \mathbb{E}[U]\|}$$

Distribution of $\widetilde{U}$ is on a hypersphere



Consider any user-specified "anchor direction" μ (we automatically estimate anchor directions shortly)

For any raw input x, define projection onto μ:

$$\text{proj}_\mu(x) \triangleq \Big\langle \frac{\phi(x) - \mathbb{E}[U]}{\|\phi(x) - \mathbb{E}[U]\|}, \frac{\mu}{\|\mu\|} \Big\rangle$$

cosine similarity between $\phi(x) - \mathbb{E}[U]$ and μ (always between -1 and 1)

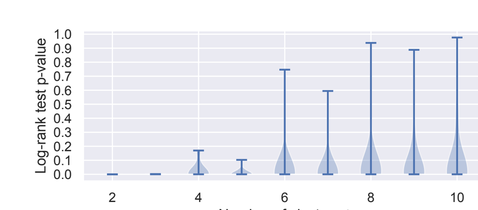
Visualization idea: look at how raw features/survival times of x vary vs $\text{proj}_\mu(x)$

How to automatically determine anchor directions to use?

- **Method #1:** cluster on embedding vectors (bring your own clustering algorithm)

$\mu_{\text{cluster } j} =$ average of embedding vectors in cluster j $-$ average of embedding vectors

We propose a novel violin plot based on the standard survival analysis log-rank test to select # of clusters


Log-rank test p-value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
2
4
6
8
10
Number of clusters k

- **Method #2:** if anchor direction should capture concept such as "female"

$\mu =$ average of embedding vectors of female patients $-$ average of embedding vectors

Technical details (see paper for details):

- Assume i.i.d. data split into train/anchor direction estimation/visualization sets (train: for training neural survival model)

assumption needed so various statistical tests make sense

- We show a systematic "clumping" pattern our framework exhibits when embedding vectors encode information mostly in magnitudes instead of directions

Tabular Data

- SUPPORT dataset [Knaus et al 1995]
- Train DeepSurv
 - Encoder ϕ is MLP (last activation normalizes vector to have norm 1)
 - g is a linear layer with 1 output
- Plot **raw feature vs projection onto anchor direction** as heatmap

Can run statistical test of association between raw feature & projection (details in paper; enables ranking raw features)

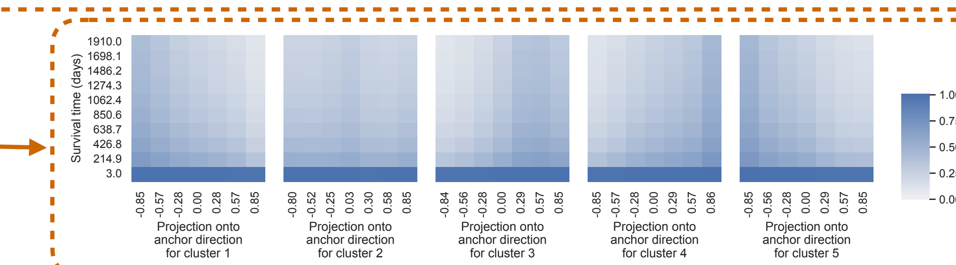
- Plot **conditional survival function vs projection onto anchor direction** as heatmap

Heatmaps help understand what semantics might be captured by different anchor directions

More plots + 1 more tabular dataset are in paper

Raw feature heatmap


age bin#1(-inf,50.40)
age bin#2[50.40,61.78)
age bin#3[61.78,68.77)
age bin#4[68.77,76.24)
age bin#5[76.24,inf)
number of comorbidities bin#1(-inf,1.00)
number of comorbidities bin#2[1.00,2.00)
number of comorbidities bin#3[2.00,3.00)
number of comorbidities bin#4[3.00,inf)
mean arterial blood pressure bin#1(-inf,61.00)
mean arterial blood pressure bin#2[61.00,71.00)
mean arterial blood pressure bin#3[71.00,93.00)
mean arterial blood pressure bin#4[93.00,112.00)
mean arterial blood pressure bin#5[112.00,inf)
heart rate bin#1(-inf,69.00)
heart rate bin#2[69.00,82.00)
heart rate bin#3[82.00,110.00)
heart rate bin#4[110.00,125.00)
heart rate bin#5[125.00,inf)
respiration rate bin#1(-inf,14.00)
respiration rate bin#2[14.00,20.00)
respiration rate bin#3[20.00,24.00)
respiration rate bin#4[24.00,30.00)
respiration rate bin#5[30.00,inf)
temperature bin#1(-inf,36.09)
temperature bin#2[36.09,36.50)
temperature bin#3[36.50,37.59)
temperature bin#4[37.59,38.40)
temperature bin#5[38.40,inf)
white blood count bin#1(-inf,6.10)
white blood count bin#2[6.10,8.90)
white blood count bin#3[8.90,12.00)
white blood count bin#4[12.00,16.50)
white blood count bin#5[16.50,inf)
serum sodium bin#1(-inf,133.00)
serum sodium bin#2[133.00,136.00)
serum sodium bin#3[136.00,139.00)
serum sodium bin#4[139.00,142.00)
serum sodium bin#5[142.00,inf)
serum creatinine bin#1(-inf,0.80)
serum creatinine bin#2[0.80,1.00)
serum creatinine bin#3[1.00,1.40)
serum creatinine bin#4[1.40,2.20)
serum creatinine bin#5[2.20,inf)
female
diabetes
dementia
race cat#1(unspecified)
race cat#2(asian)
race cat#3(black)
race cat#4(hispanic)
race cat#5(other)
race cat#6(white)
cancer cat#1(no)
cancer cat#2(yes)
cancer cat#3(metastatic)
1.0
0.8
0.6
0.4
0.2
0.0
-0.85 -0.57 -0.28 0.00 0.28 0.57 0.85
Projection onto anchor direction for cluster 1
-0.80 -0.52 -0.25 0.03 0.30 0.58 0.85
Projection onto anchor direction for cluster 2
-0.84 -0.56 -0.28 0.00 0.29 0.57 0.85
Projection onto anchor direction for cluster 3
-0.85 -0.57 -0.28 0.00 0.29 0.57 0.86
Projection onto anchor direction for cluster 4
-0.85 -0.56 -0.28 0.00 0.29 0.57 0.85
Projection onto anchor direction for cluster 5

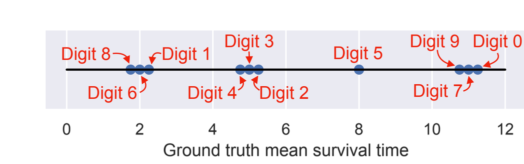

Survival time (days)
1910.0
1698.1
1486.2
1274.3
1062.4
850.6
638.7
426.8
214.9
3.0
1.00
0.75
0.50
0.25
0.00
Projection onto anchor direction for cluster 1
Projection onto anchor direction for cluster 2
Projection onto anchor direction for cluster 3
Projection onto anchor direction for cluster 4
Projection onto anchor direction for cluster 5

Survival probability heatmap (each column is a conditional survival function)

Image Data

- Survival MNIST dataset (MNIST [LeCun et al 2010] modified to have survival outcomes [Pölsterl 2019])

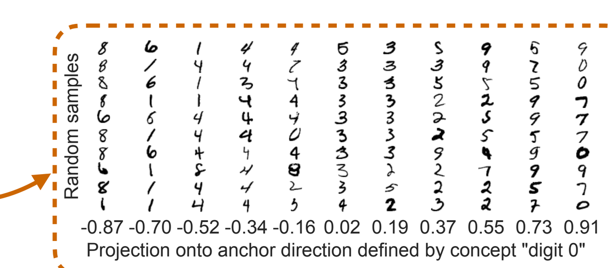
Digits with higher ground truth mean survival time are censored more often


Digit 8
Digit 1
Digit 3
Digit 5
Digit 9
Digit 0
Digit 6
Digit 4
Digit 2
Digit 7
0
2
4
6
8
10
12
Ground truth mean survival time

Example: digit 0 is always censored

- Train DeepSurv (encoder is CNN & outputs vectors with norm 1)
- Plot **raw input images (random samples) vs projection onto anchor direction**

Also possible to plot convolution filter output images (random samples) vs projections


Random samples
-0.87 -0.70 -0.52 -0.34 -0.16 0.02 0.19 0.37 0.55 0.73 0.91
Projection onto anchor direction defined by concept "digit 0"

Raw inputs vs projection plot

More plots are in the paper

With only survival labels (no classification labels), learned embedding space does not easily distinguish different digits (digits 0, 7, and 9 have similar mean survival times)

This work was supported by NSF CAREER award #2047981