Towards modelling hazard factors in unstructured data spaces using gradient-based latent interpolation

Anonymous Author(s)
Affiliation
Address
email

Abstract

The application of deep learning in survival analysis (SA) gives the opportunity to utilize unstructured and high-dimensional data types uncommon in traditional survival methods. This allows to advance methods in fields such as digital health, predictive maintenance and churn analysis, but often yields less interpretable and intuitively understandable models due to the black-box character of deep learning-based approaches. We close this gap by proposing 1) a multi-task variational autoencoder (VAE) with survival objective, yielding survival-oriented embeddings, and 2) a novel method HazardWalk that allows to model hazard factors in the original data space. HazardWalk transforms the latent distribution of our autoencoder into areas of maximized/minimized hazard and then uses the decoder to project changes to the original domain. Our procedure is evaluated on a simulated dataset as well as on a dataset of CT imaging data of patients with liver metastases.

1 Introduction

Survival analysis (SA) is indispensable for estimating factors of increased risk, e.g., for an infection of a contagious infection or the onset of a disease. Survival is also an important clinical endpoint for assessing safety and efficacy of cancer therapies [1]. Despite its importance and practical relevance, modelling survival times remains challenging due to the often complex data generating process, involving censoring, truncation, time-varying features, recurrent events or competing risks. While there exist frameworks such as [2, 3] that reduce the complexity of modelling survival related tasks by relating SA to a Poisson regression using a special data transformation [4], many recent approaches adapt the semi-parametric Cox proportional hazard (PH) model [5]. Instead of relating directly to the survival probability, the Cox PH models the hazard rate, which is the instantaneous risk of experiencing an event (disease, death, etc.). Modern non-linear adaptions of the Cox PH model employ neural networks and advanced optimizers [6, 7]. Other approaches strive to model the hazard at every possible point in time using a discretized output time domain [8, 9, 10]. More recently, neural ordinary differential equations for use in SA were applied successfully [11]. Generative models and survival have been investigated in [12], who use an autoencoder in a semi-supervised setting to encode survival information and classify between long- and short-term survivors. Moreover, [13] regularize a denoising autoencoder with a survival loss to perform cardiac motion analysis.

SA is predominantly applied for tabular data. A few exceptions exist (see, e.g., [14]), but a majority of techniques is not designed for unstructured or high-dimensional data. Modern deep learning approaches address this problem, but in contrast to existing SA techniques, do not have a straightforward model interpretation – a crucial aspect in digital health and medical decision making. This interpretation and understanding is equally important for generative models, yet more intricate and especially challenging in the light of SA. We close this gap by proposing a two-step procedure combining ideas of deep generative models and counterfactuals [15]. In a first step, we estimate

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

an inherently interpretable latent representation of the unstructured data source using a multi-task variational autoencoder (VAE; [16]) with regularizing survival objective. In the second step we help the practitioner to understand the modelled relationship of this latent space and the survival outcome using our novel *HazardWalk* method. While our autoencoder architecture allows modelling and isolating survival-relevant factors, the HazardWalk can be used to traverse in the latent space guided by the survival-objective and utilizes the generative decoder to project back its reasoning to the original data space (e.g., by an respective visualization in the image).

2 Methods

Autoencoder with survival downstream objective Our proposed architecture consists of a probabilistic encoder enc_ϕ and a decoder dec_θ with parameters ϕ and θ, respectively. For a sample $\boldsymbol{x} \in \mathcal{X}$ from the sample space $\mathcal{X}$, the approximate inference distribution $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ of the encoder is given by $\mathcal{N}(\boldsymbol{\mu}_\phi, \text{diag}(\boldsymbol{\sigma}_\phi^2))$ where $\boldsymbol{\mu}_\phi, \boldsymbol{\sigma}_\phi = enc_\phi(\boldsymbol{x})$. By sampling $\boldsymbol{z} \sim q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$, a reconstructed version $\hat{\boldsymbol{x}}$ of the data sample $\boldsymbol{x}$ can be produced by dec_θ. In order to learn this reconstruction, we optimize the evidence lower bound (ELBO)

$$\mathcal{L}_{\beta-VAE}(\theta, \phi; \boldsymbol{x}) = -\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\right] + \beta \mathbb{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z}))\,, \tag{1}$$

where $\log p_\theta(\boldsymbol{x}|\boldsymbol{z})$ accounts for the reconstruction and β is an additional parameter to control the impact of the KL-divergence $\mathbb{KL}$ between $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and the Gaussian prior $p(\boldsymbol{z})$ [17].

While the plain (β-)VAE works well for reconstruction, it does not learn concepts to emphasize survival-relevant information in $\boldsymbol{x}$. We therefore propose to extend the (β)-VAE to a multi-task model with survival downstream objective by adding a network head cox_ψ with parameters ψ to the latent bottleneck. This leads to a latent space guided by survival information. The task of cox_ψ is to predict the log-hazard rate or risk score $r \in \mathbb{R}$ with $r = cox_\psi(\boldsymbol{z})$ for $\boldsymbol{z} \sim q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$. In SA, the hazard $h(t \mid \boldsymbol{x})$ describes the instantaneous risk of experiencing the event of interest exactly at a specific time point t. Following the original Cox-PH model, the hazard can be decomposed as $h(t \mid \boldsymbol{x}) = h_0(t)\exp(r)$, where $h_0(t)$ is a feature-independent baseline hazard and $\exp(r)$ serves as a feature-dependent scaling factor based on the risk score r. Let $\boldsymbol{X}$ be a dataset of n observations $\boldsymbol{x}_i$ with corresponding event times t_i for $i = 1, \ldots, n$. For a predicted risk score $r_i = cox_\psi(\boldsymbol{z}_i)$ based on $\boldsymbol{z}_i \sim enc_\phi(\boldsymbol{x}_i)$, the corresponding objective function is given by the negative partial Cox log-likelihood (see, e.g., [6]):

$$\mathcal{L}_{Cox}(\phi, \psi; \boldsymbol{X}) = -\frac{1}{\sum_{i=1}^n \delta_i} \sum_{i=1}^n \delta_i \left(r_i - \log \sum_{j:t_j \geq t_i} \exp(r_j) \right)\,, \tag{2}$$

where $\delta_i = 0$ if the ith observation is censored, else $\delta_i = 1$. By regularizing the autoencoder with the survival downstream task, latent features are forced to learn a representation that is inherently interpretable from a survival point of view. The joint objective of our multi-task approach is given by

$$\mathcal{L}(\phi, \theta, \psi; \boldsymbol{X}) = \frac{\tau}{n} \sum_{i=1}^n \left(\mathcal{L}_{\beta-VAE}(\phi, \theta; \boldsymbol{x}_i)\right) + (1-\tau)\mathcal{L}_{Cox}(\phi, \psi; \boldsymbol{X})\,, \tag{3}$$

with $\tau \in [0, 1]$ controlling the influence of each loss term in the convex combination of both. In our experiments, a value of $\tau = 0.5$ turns out to give both good reconstructions and survival-oriented latent spaces.

Hazard guided walk In order to visualize the impact of survival-relevant dimensions in the latent space, we propose the gradient-based hazard latent walk (*HazardWalk*). HazardWalk is a new method to traverse the latent space based on the estimated hazards by cox_ψ and utilizes dec_θ as generative model to project these changes to the original data space, making important factors on the overall survival directly observable. The latent variational distribution of the VAE thereby ensures smooth transitions in $\mathcal{X}$. First, the latent feature distribution is obtained by forwarding a sample $\boldsymbol{x}$ through enc_ϕ, yielding parameters $(\boldsymbol{\mu}_\phi, \boldsymbol{\sigma}_\phi) = \xi$. The goal is now to obtain the expected gradient of the hazard rate w.r.t. the whole latent distribution, which points in the direction of maximal hazard rate increase. We approximate this using Monte Carlo sampling with

$$\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\nabla_\xi \exp(cox_\psi(\boldsymbol{z}))\right] = \int \nabla_\xi \exp(r) q_\phi(\boldsymbol{z} \mid \boldsymbol{x})\, d\boldsymbol{z} \approx \frac{1}{B}\sum_{b=1}^B \nabla_\xi \exp(r_b) \tag{4}$$

based on samples $\boldsymbol{z}_b \sim q_\phi$. By updating the parameters of $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ in the direction of (4) and repeating this procedure for a pre-specified amount of iterations, we shift and scale the latent distribution into areas of maximized hazard. We can then use the decoder dec_θ to produce new samples $\widetilde{\boldsymbol{x}}$ from this transformed latent distribution. In this case, the disentanglement of latent features induced by the factorized Gaussian prior leads to samples, where survival-irrelevant characteristics are left relatively unchanged and survival-relevant characteristics are changed notably. In other words, the HazardWalk allows practitioners to understand the model's learned relationships directly in the original sample space by, e.g., plotting images based on the most survival-relevant changes in the latent space. This procedure can also be reversed, using the negative expected gradient for the parameter updates to generate samples with minimal risk.

3 Experiments

We evaluate the HazardWalk procedure on two datasets. First, on a simulated dataset based on MNIST, for which survival times are generated as described in [18]. The result is an MNIST dataset with the original images, where high hazard rates correspond to a high digit and vice versa. Second, we demonstrate the proposed approach on a dataset of 492 CT scans of patients with liver metastases, where the survival time after imaging is known.

Our enc_ϕ and dec_θ each consist of four residual blocks [19] with downsampling or upsampling operations respectively. cox_ψ was chosen to be a fully-connected network with no hidden layer and no bias, which equals a linear predictor. We use Adam [20] as optimizer with a learning rate of 1e-4 for θ, ϕ and 1e-5 for ψ.

HazardWalk on MNIST Based on a latent space of 4 dimensions, Figure 1 shows the differences in the embedding of a VAE with and without the regularizing Cox objective on the test dataset. As expected, the vanilla VAE (left plot) does not preserve the natural order of MNIST numbers, but structures the latent space with the intention of maximum visual reconstruction, hence the close neighborhood of digits like 0 and 9. In contrast, when adding the survival information (right plot), the behavior changes. The main source of variation is now based on the actual hazard, i.e., the natural digit order. Digits such as 0 and 1 are now placed next to each other instead of being maximally separated as for the vanilla VAE due to their visual dissimilarity.

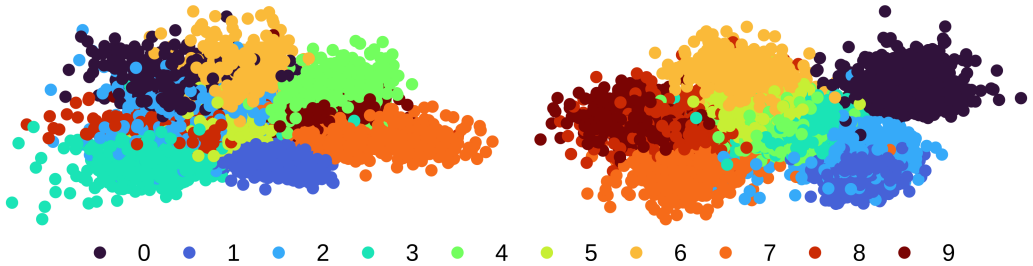


Figure 1: First two PCA components of MNIST embeddings from a VAE (**left**) versus a VAE with Cox objective (**right**). The color indicates the digit class. The VAE embedding is optimized to show structural similarities, whereas the Cox-regularized embedding groups classes according to their associated hazard.

In a second step we investigate the effect of the HazardWalk on these survival-oriented embeddings. We apply the HazardWalk with 1500 iterations, 128 samples and visualize the progress of the transformation. As depicted in Figure 2, the gradients of cox_ψ lead to a transition into different digit classes with higher (or lower) hazard. It is noteworthy that each digit shifts into a class that allows to keep some of the original core features (e.g., a 2 turns into a 3 first, then into an 8 instead of directly transforming into a 9 with the maximum associated hazard rate).

HazardWalk on computed tomography imaging data CT imaging is an important step in the diagnostic workup of cancer patients. We have trained the Cox-regularized VAE on imaging data of patients with tumor metastases in the liver. Here, we show a central slice of the volumetric CTs. The input to the network consists of two channels: the actual masked CT data and the segmentation of the

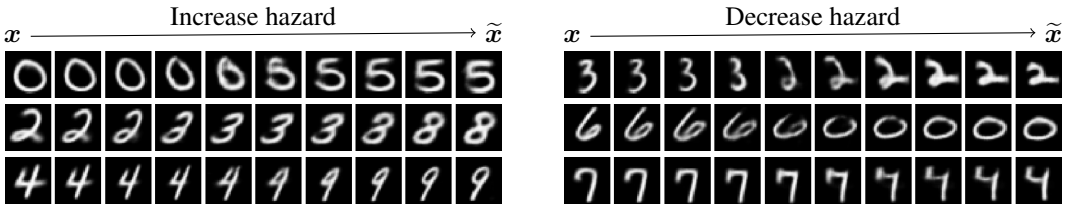
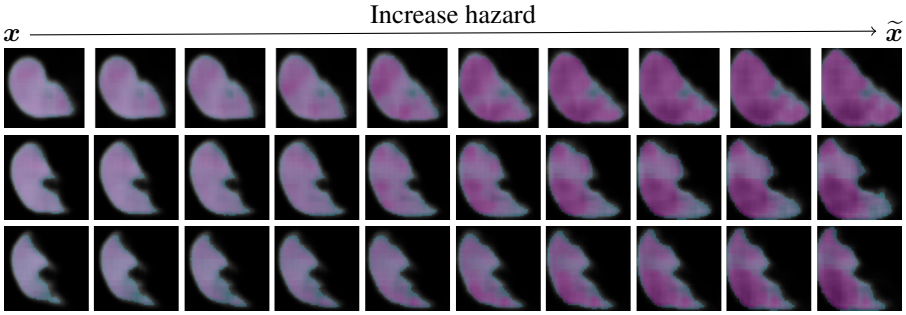


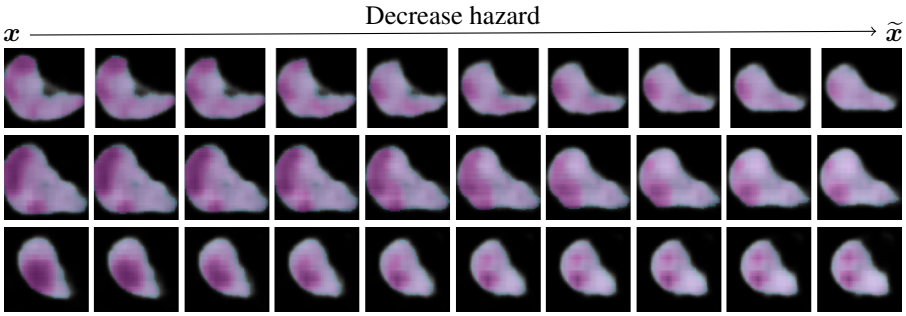
Figure 2: HazardWalk applied on three exemplary MNIST samples when increasing (**left**) or decreasing (**right**) the hazard rate. Applying the transformation results in a gradual change of the digit class while maintaining characteristics of the original sample $\boldsymbol{x}$.

tumor. The image was downscaled to a resolution of $64 \times 64 \times 64$ voxels and a latent space of 8 was chosen.

A major challenge for model training is the small sample size in combination with a relatively high dimensional input space. This results in gaps in the otherwise smooth latent space. By increasing the β-parameter and therefore the influence of the $\mathbb{KL}$-term, this can be mitigated at the cost of lower-quality reconstructions. Figure 3 depicts multiple examples, showing different liver cancer cases where the hazard ratio is tweaked by HazardWalk. Our method confirms the meaningfulness of the learned latent space of the autoencoder, increasing number and sizes of tumor patches when increasing hazard. The shape of the liver changes only slightly, which indicates that overall survival may be unrelated to this characteristic. In the given example, the sample in the first row of Figure 3a has a time-to-event of 1479 days and an estimated hazard rate of 0.601. This constitutes a hazard that is significantly lower than the baseline. After the transformation, the newly obtained $\widetilde{\boldsymbol{x}}$ is estimated with a hazard rate of 3.424. On the contrary, reducing the hazard leads to exactly the opposite – the tumor shrinks or vanishes. This can be seen, e.g., for the first sample in Figure 3b, with an original time-to-event of 61 days and a hazard rate of 3.165 that transformed to 0.785.



(a) Increasing the hazard rate leads to a spread of (larger) tumor patches.



(b) Decreasing the hazard rate results in shrinking and vanishing tumor patches.

Figure 3: HazardWalk applied on computed tomography of the liver when increasing (**top**) or decreasing (**bottom**) the hazard rate with HazardWalk. The images show reconstructed liver CTs in grayscale with red-tinted tumor patches.

4 Conclusion and Outlook

In this work we proposed a novel two-step procedure to learn and interpret the relationship between unstructured data sources such as images and survival times. While our generative model allows to emphasize hazard factors in unstructured data spaces, the proposed method HazardWalk yields a straightforward interpretation of the model using a gradient-based latent interpolation and reconstruction over the generative decoder.

One limiting factor in training generative models for SA are the typically small-scale datasets. This results in gaps in the latent space and/or insufficient representations. Adjusting the $\mathbb{KL}$-term turned out to be beneficial in our experiments to focus on the quality of the representation, resulting in more reliable and stable HazardWalk interpolations. While there are larger CT datasets available (e.g., [21]), these are mostly for segmentation tasks and not associated with a survival outcome. An extension of our approach to a semi-supervised setting could thus allow the integration of more information to improve the encoding and generation. Alternatively, adversarial autoencoders as in [22, 23] could help to improve the quality of generated images.

References

[1] Amanda Delgado and Achuta Kumar Guddati. "Clinical endpoints in oncology-a primer". In: *American journal of cancer research* 11.4 (2021), p. 1121.

[2] Andreas Bender, David Rügamer, Fabian Scheipl, and Bernd Bischl. "A General Machine Learning Framework for Survival Analysis". In: *ArXiv e-prints* (2020). arXiv: 2006.15442.

[3] Philipp Kopper, Sebastian Pölsterl, Christian Wachinger, Bernd Bischl, Andreas Bender, and David Rügamer. "Semi-structured deep piecewise exponential models". In: *ArXiv e-prints* (2020). arXiv: 2011.05824.

[4] Michael Friedman. "Piecewise Exponential Models for Survival Data with Covariates". In: *The Annals of Statistics* (1982).

[5] David R. Cox. "Regression Models with Life Tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* (1972).

[6] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. "DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network". In: *BMC Medical Research Methodology* 18.1 (2018).

[7] Margaux Luck, Tristan Sylvain, Héloïse Cardinal, Andrea Lodi, and Yoshua Bengio. "Deep Learning for Patient-Specific Kidney Graft Survival Analysis". In: *ArXiv e-prints* (2017). arXiv: 1705.10245.

[8] Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela Van Der Schaar. "DeepHit: A deep learning approach to survival analysis with competing risks". In: *32nd AAAI Conference on Artificial Intelligence, AAAI 2018* 32 (2018), pp. 2314–2321.

[9] Changhee Lee, Jinsung Yoon, and Mihaela Van Der Schaar. "Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis with Competing Risks Based on Longitudinal Data". In: *IEEE Transactions on Biomedical Engineering* 67.1 (2020), pp. 122–133.

[10] Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. "Deep Recurrent Survival Analysis". In: *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019* (2018), pp. 4798–4805.

[11] Stefan Groha, Sebastian M. Schmon, and Alexander Gusev. "A General Framework for Survival Analysis and Multi-State Modelling". In: *ArXiv e-prints* (2020). arXiv: 2006.04893.

[12] Sveinn Pálsson, Stefano Cerri, Andrea Dittadi, and Koen Van Leemput. "Semi-Supervised Variational Autoencoder for Survival Prediction". In: *International MICCAI Brainlesion Workshop* (2019), pp. 124–134.

[13] Ghalib A. Bello, Timothy J.W. Dawes, Jinming Duan, Carlo Biffi, Antonio de Marvao, Luke S.G.E. Howard, J. Simon R. Gibbs, Martin R. Wilkins, Stuart A. Cook, Daniel Rueckert, and Declan P. O'Regan. "Deep-learning cardiac motion analysis for human survival prediction". In: *Nature Machine Intelligence* 1.2 (2019), pp. 95–104.

[14] Christoph Haarburger, Philippe Weitz, Oliver Rippel, and Dorit Merhof. "Image-based Survival Analysis for Lung Cancer Patients using CNNs". In: *Proceedings - International Symposium on Biomedical Imaging* 2019-April (2018), pp. 1197–1201.

[15] Sahil Verma, John Dickerson, and Keegan Hines. "Counterfactual Explanations for Machine Learning: A Review". In: *ArXiv e-prints* (2020). arXiv: 2010.10596.

[16] Diederik P. Kingma and Max Welling. "Auto-encoding variational bayes". In: *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2014.

[17] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. "Understanding disentangling in beta-VAE". In: *ArXiv e-prints* (2018). arXiv: 1804.03599.

[18] Michael F. Gensheimer and Balasubramanian Narasimhan. "A scalable discrete-time survival model for neural networks". In: *PeerJ* 7 (2019), e6257.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2016-Decem. IEEE Computer Society, 2016, pp. 770–778.

[20] Diederik P. Kingma and Jimmy Lei Ba. "Adam: A method for stochastic optimization". In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2015.

[21] Michela Antonelli et al. "The Medical Segmentation Decathlon". In: *ArXiv e-prints* (2021). arXiv: 2106.05735.

[22] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. "Autoencoding beyond pixels using a learned similarity metric". In: *33rd International Conference on Machine Learning, ICML 2016* 4 (2015), pp. 2341–2349.

[23] Stanislav Pidhorskyi, Donald Adjeroh, and Gianfranco Doretto. "Adversarial Latent Autoencoders". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2020), pp. 14092–14101.