Deep Correlational Learning for Survival Prediction from Multi-modality Data

Jiawen Yao, Xinliang Zhu, Feiyun Zhu, and Junzhou Huang[✉]

Department of Computer Science and Engineering, University of Texas at Arlington,
Arlington, TX 76019, USA
`jzhuang@uta.edu`

Abstract. Technological advances have created a great opportunity to provide multi-view data for patients. However, due to the large discrepancy between different heterogeneous views, traditional survival models are unable to efficiently handle multiple modalities data as well as learn very complex interactions that can affect survival outcomes in various ways. In this paper, we develop a Deep Correlational Survival Model (DeepCorrSurv) for the integration of multi-view data. The proposed network consists of two sub-networks, view-specific and common subnetwork. To remove the view discrepancy, the proposed DeepCorrSurv first explicitly maximizes the correlation among the views. Then it transfers feature hierarchies from view commonality and specifically fine-tunes on the survival regression task. Extensive experiments on real lung and brain tumor data sets demonstrated the effectiveness of the proposed DeepCorrSurv model using multiple modalities data across different tumor types.

1   Introduction

Survival analysis aims at modeling the time that elapses from the beginning of follow-up until a certain event of interest (e.g. biological death) occurs. The most popular survival model is Cox proportional hazards model [6]. However, the Cox model and recent approaches [2–4,17] are still built based on the assumption that a patient's risk is a linear combination of covariates. Another limitation is that they mainly focus on one view and cannot efficiently handle multi-modalities data. Recently, Katzman *et al.* proposed a deep fully connected network (Deep-Surv) to learn highly complex survival functions [11]. They demonstrated that DeepSurv outperformed the standard linear Cox proportional hazard model. However, DeepSurv cannot process pathological images and also is unable to handle multi-view data.

To integrate multiple modalities and eliminate view variations, a good solution is to learn a joint embedding space which different modalities can be compared directly. Such embedding space will benefit the survival analysis since

recent study has suggested that common representation from different modalities provide important information for prognosis [18,21,22]. To learn the embedding space, one very popular method is canonical correlation analysis (CCA) [8] which aims to learn features in two views that are maximally correlated. Deep canonical correlation analysis [1] has been shown to be advantageous and such correlational representation learning (CRL) methods provide a very good chance for integrating different modalities of survival data. However, since these CRL methods are unsupervised learning models, they still have the risk of discarding important markers that are highly associated with patients' survival outcomes.

In this paper, we develop a Deep Correlational Survival Model (Deep-CorrSurv) to integrate views of pathological images and molecular data for survival analysis. The proposed method first eliminates the view variations and finds the maximum correlated representation. Then it transfers feature hierarchies from such common space and specifically fine-tunes on the survival regression task. It has the ability to discover important markers that are not found by previous deep correlational learning which will benefit the survival prediction. The contribution of this paper can be summarized as: (1) DeepCorrSurv can model very complex view distributions and learn good estimators for predicting patients' survival outcomes with insufficient training samples. (2) It used CNNs to represent much more abstract features from pathological images for survival prediction. Traditional survival models usually adopted hand-crafted imaging features. (3) Extensive experiments on TCGA-LUSC and GBM demonstrate that DeepCorrSurv model outperforms those state-of-the-art methods and achieves more accurate predictions across different tumor types.

2   Methodology

Given two sets $\mathbf{X}, \mathbf{Y}$ with m samples, the i-th sample is denoted as $\mathbf{x}_i$ and $\mathbf{y}_i$. Survival analysis is about predicting the time duration until an event occurs, and in our case the event is the death of a cancer patient. In survival data set, patient i has observation time and the censored status, denoted as (t_i, δ_i). δ_i is the indicator: 1 is for a uncensored instance (the death event occurs during the study), and 0 is for a censored instance (the event is not observed). The observation time t_i is either a survival time (S_i) or a censored time (C_i) which is determined by the status indicator δ_i. If and only if $t_i = \min(S_i, C_i)$ can be observed during the study, the dataset is said to be right-censored which is the most common case in real world.

Figure 1 illustrates the pipeline of the proposed DeepCorrSurv. It consists of two sub-networks, view-specific sub-network f_1, f_2 and the common sub-network g_c. We proposed Convolutional Neural Networks (CNNs) as one image-view sub-network f_1 and Fully Connected Neural Networks (FCNs) as another view-specific sub-network f_2 to learn deep representations from pathological images and molecular profiling data, respectively. The sub-network f_1 consists of 3 convolutional layers, 1 max-pooling layer and 1 fully-connected layer. In each convolutional layer, we employ ReLU as the nonlinear activation function. The
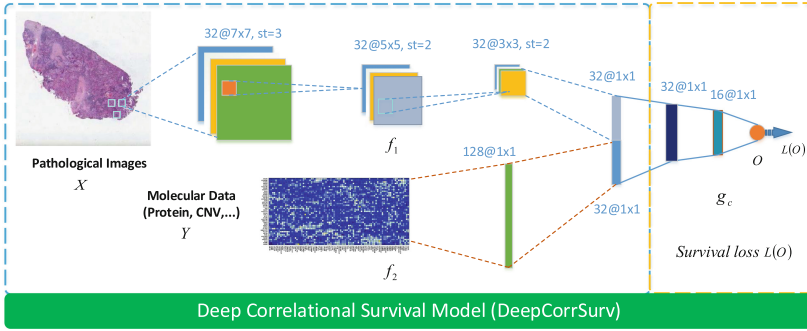
Fig. 1. The architecture of the DeepCorrSurv. 'st' is short for 'stride'.

sub-network f_2 includes two fully connected layers with 128 and 32 neurons equipped with ReLU activation function.

Deep Correlational Model: For any sample $\mathbf{x}_i, \mathbf{y}_i$ passing through the corresponding view sub-network, its representation is denoted as $f_1(\mathbf{x}_i; \mathbf{w_x})$ and $f_2(\mathbf{y}_i; \mathbf{w_y})$ respectively where $\mathbf{w_x}, \mathbf{w_y}$ represent all parameters of two sub-networks. The outputs of two branches will be connected to a correlation layer to form the common representation.

Deep correlational model seeks pairs of projections that maximize the correlation of two outputs from each networks $f_1(\mathbf{x}_i; \mathbf{w_x}), f_2(\mathbf{y}_i; \mathbf{w_y})$. If $\mathbf{w_x}, \mathbf{w_y}$ mean all parameters of two networks, then the commonality is enforced by maximizing the correlation between two views as follows

$$L = corr(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^{m}(f_1(\mathbf{x}_i) - \overline{f_1(\mathbf{X})})(f_2(\mathbf{y}_i) - \overline{f_2(\mathbf{Y})})}{\sqrt{\sum_{i=1}^{m}(f_1(\mathbf{x}_i) - \overline{f_1(\mathbf{X})})^2 \sum_{i=1}^{m}(f_2(\mathbf{y}_i) - \overline{f_2(\mathbf{Y})})^2}}, \tag{1}$$

where we omit networks' parameters $\mathbf{w_x}, \mathbf{w_y}$ in the loss function (1). We can maximize the correlation loss function to provide the shared representation indicating the most correlated features from two modalities.

Fine-Tune with Survival Loss: Denote the shared representation from the two views as $\mathbf{Z}$. Denote $\mathbf{O} = [o_1, ..., o_m]^\top$ as the outputs of common sub-network $\mathbf{g}_c$, i.e., $o_i = \mathbf{g}_c(\mathbf{z}_i)$. The survival loss function is set to be the negative log partial likelihood:

$$L(\mathbf{o}) = \sum_{i:R(t_i)=1} \left(-o_i + \log \sum_{j:t_j >= t_i} \exp(o_j)\right). \tag{2}$$

where o_i is the output of i-th patient. $R(t_i)$ is the risk set at time t_i, which is the set of all individuals who are still under study before time t_i. j is from the set whose survival time is not smaller than t_i $(t_j \geq t_i)$. Another understanding

is that all patients who live longer than i-th patient will be chosen into this set. Different from Cox-based models which only handle the linear condition in the risk function, the proposed model can better fit realistic data and learn complex interactions using deep representation.

Discussions: Although different views of health data are very heterogeneous, they still do share common information for prognosis. Deep correlational learning is first trained to find such common representation using the correlation function (1). However, this procedure has a risk of discarding the discriminant markers for predicting patients' survival outcomes due to it belongs to unsupervised learning. To overcome this problem, the DeepCorrSurv transfers knowledge from the deep correlational learning and fine-tunes the network using the survival loss function (2). This will make DeepCorrSurv able to discover important markers that are ignored by correlational model and learn the best representation for survival prediction. Compared with the recent deep survival models [11,20] which can only handle one specific view of data, the DeepCorrSurv can achieve more complex architecture for the integration of multi-modalities data which can be used in the practical application on more challenging dataset.

3 Experiments

3.1 Dataset Description

We used a public cancer survival dataset TCGA (The Cancer Genome Atlas) project [10] which provides high resolution whole slide pathological images and molecular profiling data. We conducted experiments on two cancer types: glioblastoma multiforme (GBM) and lung squamous cell carcinoma (LUSC). For each cancer type, we adopted a core sample set from UT MD Anderson Cancer Center [19] in which each sample has information for the overall survival time, pathological images and molecular data related to gene expression.

- **TCGA-LUSC**: Non-Small-Cell Lung Carcinoma (NSCLC) is the majority of lung cancer. Lung squamous cell carcinoma (LUSC) is one major type in NSCLC. We collected 106 patients with pathological images and protein expression (reverse-phase protein array, 174 proteins).
- **TCGA-GBM**: Glioma is a type of brain cancer and it is the most common malignant brain tumor. 126 patients are selected from the core set with images and CNV data (Copy number variation, 106 dimension).

With the help of pathologists, we have annotations that locate the tumor regions in whole slide images (WSIs). We randomly extract patches of size $1024\times$ 1024 from the tumor regions. To analyze pathological images in comparison survival models, we calculated hand-crafted features using CellProfiler [5] which serves as a state-of-the-art medical image feature extracting and quantitative analysis tool. Similar to the pipeline in [16], a total of 1,795 quantitative features were calculated from each image tile. These types of image features include cell shape, size, texture of the cells and nuclei, as well as the distribution of pixel intensity in the cells and nuclei.

3.2   Comparison Methods

We compare our DeepCorrSurv with five state-of-the-art survival models and
three baselines deep survival models. Five survival methods include LASSO-
Cox [15], Parametric censored regression models with components with Weibull,
Logistic distribution [9], Boosting concordance index (BoostCI) [13] and Multi-
Task Learning model for Survival Analysis (MTLSA) [12]. To demonstrate the
effectiveness of the integration in our model, We adopted structured sparse CCA-
based feature selection (SCCA) [7] to identify stronger correlation patterns from
imaging genetic associations. Then we applied MTLSA using such associations
for survival analysis.

Three baseline deep survival models are listed as follows: (1) CNN-Surv: CNN
sub-network f_1 followed by survival loss [20]. (2)FCN-Surv: FCN sub-network
f_2 followed by survival loss. It will use molecular profiling data for prediction.
It can be also regarded as the DeepSurv [11] version on the dataset in this
paper. (3)DeepCorr+DeepSurv: Since finding the common space by maximizing
the correlation between two views belongs to unsupervised method, it cannot
ensure that the embedding space is highly correlated with survival outcomes.
We extract the shared representation by Deep correlational learning and feed
them to another DeepSurv model.

Overall speaking, the DeepCorrSurv is optimized by the gradient descent
following the chain rule, i.e., firstly compute the loss of objective, and then
propagate the loss to each layer and finally employ gradient descent to update the
whole network. These procedures can be automatically processed by Theano [14].
To make fair comparisons, the architectures of different deep survival models are
kept the same with that corresponding parts in the proposed DeepCorrSurv. The
source codes of MTLSA and SCCA are downloaded from the authors' websites.
All other methods in our comparisons were implemented in R. LASSO-Cox and
EN-Cox are built using the *cocktail* function from the *fastcox* package. The
implementation of BoostCI can be found in the supplementary materials of [13].
The parametric censored regression are from the *survival* package.

3.3   Results and Discussion

In order to evaluate the proposed approach with other state-of-the-arts methods,
we used a 5-fold cross-validation. For each of the 5 folds, models were established
using the other 4 folds as the training subset, and performance was evaluated
with the unused fold. To evaluate the performances in survival prediction, we
take the concordance index (CI) as our evaluation metric. The C-index quantifies
the ranking quality of rankings and is calculated as follows

$$c = \frac{1}{n} \sum_{i \in \{1...N | \delta_i = 1\}} \sum_{t_j > t_i} I[o_i > o_j] \qquad (3)$$

where n is the number of comparable pairs and $I[.]$ is the indicator function. $t.$ is
the actual observation and $o.$ represents the risk obtained from survival models.

Table 1. Performance comparison of the proposed methods and other existing related methods using C-index values on TCGA-LUSC and GBM

Data	Model	LUSC	GBM
Images	LASSO-Cox [15]	0.5945 (0.1849)	0.5476 (0.0949)
	BoostCI [13]	0.5769 (0.2599)	0.5235 (0.1263)
	Weibull [9]	0.4988 (0.1924)	0.4885 (0.0127)
	Logistic [9]	0.4498 (0.1432)	0.4865 (0.0061)
	MTLSA [12]	0.6074 (0.1128)	0.6223 (0.1897)
	CNN-Surv [20]	0.5540 (0.2170)	0.5053 (0.0264)
Protein/CNV	LASSO-Cox [15]	0.5005 (0.1565)	0.5779 (0.0609)
	BoostCI [13]	0.4309 (0.1160)	0.4610 (0.1470)
	Weibull [9]	0.4334 (0.1587)	0.5131 (0.0895)
	Logistic [9]	0.5821 (0.1653)	0.5013 (0.1406)
	MTLSA [12]	0.5911 (0.2532)	0.6150 (0.1773)
	FCN-Surv [11]	0.5989 (0.1131)	0.5596 (0.0934)
Integration	SCCA [7] + MTLSA	0.5518 (0.0882)	0.5915 (0.1195)
	DeepCorr+DeepSurv	0.5760 (0.1645)	0.5842 (0.0450)
	DeepCorrSurv	**0.6287 (0.0596)**	**0.6452 (0.0389)**

Table 1 presents the C-index values by various survival regression methods on two datasets. Results of using each individual view in the table present that pathological images and molecule data can provide predictive powers while the integration of both modalities in the proposed DeepCorrSurv achieves the best performance for both lung and brain cancer. Because the proposed Deep-CorrSurv can remove view discrepancy as well as learn the survival-related common representations from both views, it obtains the highest C-index with low standard variation. When looking at deep survival models, CNN-Surv cannot achieve good prediction using imaging data alone. But when integrating with information from another view, DeepCorr+DeepSurv and the proposed Deep-CorrSurv can achieve better performances than CNN-Surv using the same imaging data. This demonstrates that the common representation by maximizing the correlation between both views can benefit the survival analysis when the samples are not sufficient.

Another observation is DeepCorr+DeepSurv and SCCA+MTLSA cannot obtain a very good estimation compared with some predictions from one view. This demonstrates that the common representation by maximizing the correlation in an unsupervised manner still has the risk of discarding markers that are highly associated with survival outcomes. On the contrary, the DeepCorrSurv can consider discriminant as well as view discrepancy which can ensure a representation that is robust to view discrepancy and also discriminant for survival prediction.

Results on TCGA-GBM dataset suggest that most models using CNV data can have better predictions than same models using imaging data. This observation is different from that in LUSC cohort. This reminds us, due to the heterogeneous of different tumor types, it is not easy to find a general model that can successfully estimate patients' survival outcomes across different tumor types using only one specific view. In addition, the original data in each view might contain variations or noises which are not survival-related and might affect the estimation of survival models. The proposed DeepCorrSurv can effectively integrate with two views and thus achieve good prediction performances across different tumor types.

4   Conclusion

In this paper, we proposed Deep Correlational Survival model (DeepCorrSurv) that is able to efficiently integrate multi-modalities censored data with small samples. One challenge is the view-discrepancy between different views in recent real cancer database. To eliminate the view discrepancy between imaging data and molecular profiling data, deep correlational learning provides a good solution to maximize the correlation of two views and find the common embedding space. However, deep correlational learning is an unsupervised approach which cannot ensure the common space is suitable for survival prediction. In order to find the truly deep representations for prediction, the proposed DeepCorrSurv transfers knowledge from the embedding space and fine-tunes the whole network using survival loss. Experiments have shown that DeepCorrSurv can discover important markers that are ignored by correlational learning and extract the best representation for survival prediction. The results have shown that since Deep-CorrSurv can model non-linear relationships between factors and prognosis, it achieved quite promising performances with improvements. In the future, we will extend the framework with other kinds of data sources.

References

1. Andrew, G., Arora, R., Bilmes, J.A., Livescu, K.: Deep canonical correlation analysis. In: ICML, pp. 1247–1255 (2013)
2. Bair, E., Hastie, T., Paul, D., Tibshirani, R.: Prediction by supervised principal components. J. Am. Stat. Assoc. **101**(473), 119–137 (2006)
3. Bair, E., Tibshirani, R.: Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol. **2**(4), E108 (2004)
4. Bøvelstad, H.M., Nygård, S., Størvold, H.L., Aldrin, M., Borgan, Ø., Frigessi, A., Lingjærde, O.C.: Predicting survival from microarray dataa comparative study. Bioinformatics **23**(16), 2080–2087 (2007)
5. Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al.: Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. **7**(10), R100 (2006)

6. Cox, D.R.: Regression models and life-tables. J. Roy. Stat. Soc.: Ser. B (Methodol.) **34**, 187–220 (1972)
7. Du, L., Huang, H., Yan, J., Kim, S., Risacher, S.L., Inlow, M., Moore, J.H., Saykin, A.J., Shen, L., Initiative, A.D.N., et al.: Structured sparse canonical correlation analysis for brain imaging genetics: an improved graphnet method. Bioinformatics **32**(10), 1544 (2016)
8. Hotelling, H.: Relations between two sets of variates. Biometrika **28**(3/4), 321–377 (1936)
9. Kalbfleisch, J.D., Prentice, R.L.: The Statistical Analysis of Failure Time Data, vol. 360. Wiley, Hoboken (2011)
10. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al.: Mutational landscape and significance across 12 major cancer types. Nature **502**(7471), 333–339 (2013)
11. Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deep survival: a deep cox proportional hazards network. arXiv preprint arXiv:1606.00931 (2016)
12. Li, Y., Wang, J., Ye, J., Reddy, C.K.: A multi-task learning formulation for survival analysis. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1715–1724 (2016)
13. Mayr, A., Schmid, M.: Boosting the concordance index for survival data-a unified framework to derive and evaluate biomarker combinations. PLoS one **9**(1), e84483 (2014)
14. Theano Development Team: Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints abs/1605.02688, May 2016. http://arxiv.org/abs/1605.02688
15. Tibshirani, R., et al.: The lasso method for variable selection in the Cox model. Stat. Med. **16**(4), 385–395 (1997)
16. Yao, J., Ganti, D., Luo, X., Xiao, G., Xie, Y., Yan, S., Huang, J.: Computer-assisted diagnosis of lung cancer using quantitative topology features. In: Zhou, L., Wang, L., Wang, Q., Shi, Y. (eds.) MLMI 2015. LNCS, vol. 9352, pp. 288–295. Springer, Cham (2015). doi:10.1007/978-3-319-24888-2_35
17. Yao, J., Wang, S., Zhu, X., Huang, J.: Imaging biomarker discovery for lung cancer survival prediction. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 649–657. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_75
18. Yuan, Y., Failmezger, H., Rueda, O.M., Ali, H.R., Gräf, S., Chin, S.F., Schwarz, R.F., Curtis, C., Dunning, M.J., Bardwell, H., et al.: Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. Sci. Transl. Med. **4**(157), 157ra143 (2012)
19. Yuan, Y., Van Allen, E.M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., Byers, L.A., Xu, Y., Hess, K.R., Diao, L., et al.: Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat. Biotechnol. **32**(7), 644–652 (2014)
20. Zhu, X., Yao, J., Huang, J.: Deep convolutional neural network for survival analysis with pathological images. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 544–547. IEEE (2016)

21. Zhu, X., Yao, J., Luo, X., Xiao, G., Xie, Y., Gazdar, A., Huang, J.: Lung cancer survival prediction from pathological images and genetic data - an integration study. In: IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp. 1173–1176 (2016)
22. Zhu, X., Yao, J., Xiao, G., Xie, Y., Rodriguez-Canales, J., Parra, E.R., Behrens, C., Wistuba, I.I., Huang, J.: Imaging-genetic data mapping for clinical outcome prediction via supervised conditional gaussian graphical model. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 455–459. IEEE (2016)