

Boosting Deep Learning Risk Prediction with Generative Adversarial Networks for Electronic Health Records

Zhengping Che^{1*}, **Yu Cheng**^{2*}, **Shuangfei Zhai**³,
Zhaonan Sun⁴, **Yan Liu**¹

¹Department of Computer Science, University of Southern California

²AI Foundations, IBM T. J. Watson Research Center

³Department of Computer Science, Binghamton University, SUNY

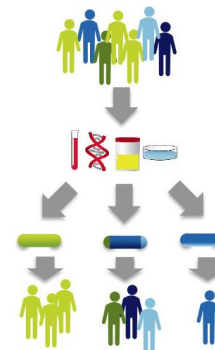
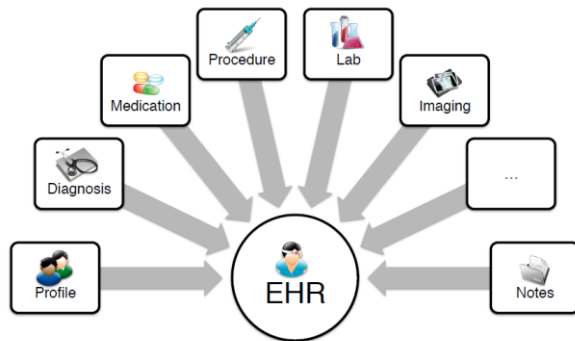
⁴Healthcare Analytics Group, IBM T. J. Watson Research Center



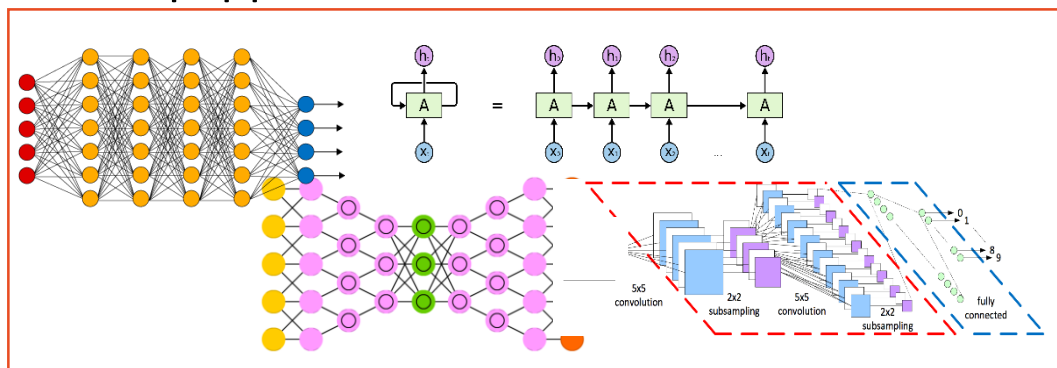
Nov 19, 2017

Background

- Rapid growth of Electronic Health Records (EHRs)
 - Urgent requirement of data-driven personalized healthcare



- Recent success and development in deep learning
 - If equipped with massive data...



[ImageNet]

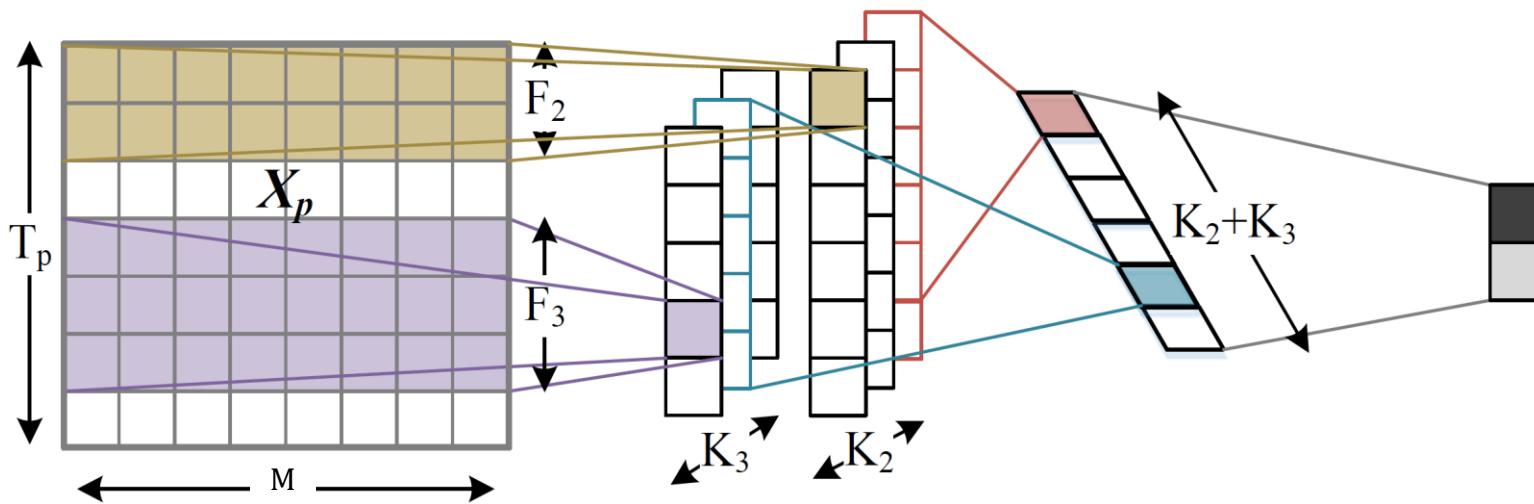
Motivation

- **High-quality labeled** medical data is limited
 - **Label Amount:** Rare diseases and conditions
 - ~7000 rare diseases in US with <200K cases
 - **Label Cleanness:** Difficult miscellaneous diseases
 - **Labelling Cost:** Time and money
 - **Data Availability:** Patient privacy
 - ...



- Good prediction model with limited labeled data?
 - Learn the structure of data manifold from limited data
 - **Generative adversarial network to get plausible labeled data**
 - Use augmented data to improve prediction performance
 - **Semi-supervised learning for deep learning models**

Basic Prediction Framework

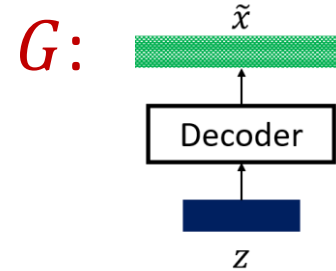
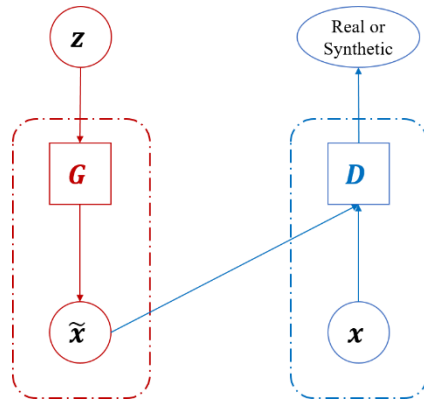


- Raw input: medical event sequence (T_p events)
- Feature learning: **Word2vec** embedding (dimension M) [Mikolov et al., 2013]
- Prediction model: 1-D convolutional neural network (**CNN**) [Che et al., 2017]
 - Competitive baseline

Heart Failure	CNN	GRU	LSTM	LR	SVM	RF
Accuracy	0.8630	0.8578	0.8511	0.8494	0.8443	0.8571
AUROC Score	0.9329	0.9129	0.9103	0.9052	0.9017	0.9225

Generative Adversarial Network (GAN)

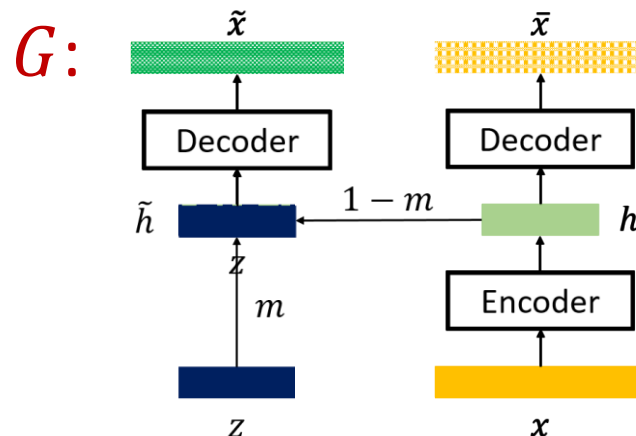
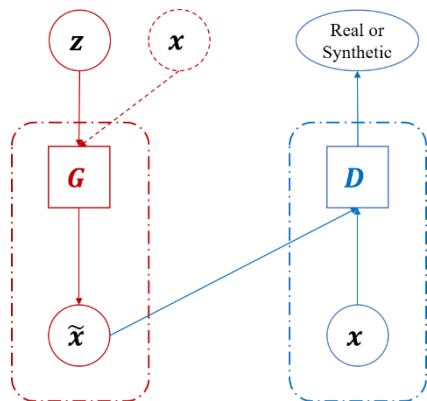
- Generative Adversarial Network [Goodfellow et al., 2014]



- Discriminator D** : distinguish whether the input x is from real data
- Generator G** : generate plausible data \tilde{x} given noise z
- Train by solving a mini-max game
$$\min_G \max_D \mathbb{E}_{x \sim p_d(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$
- Discriminator D** : 1-D convolutional neural network (CNN)
- Generator G** : 1-D de-convolutional neural network (DCNN)
- G generates **plausible synthetic data**.
 - Labeled data?**

Generating Labeled Data - ehrGAN

- Generative Adversarial Network with variational contrastive divergence [Zhai et al., 2016]



- Discriminator D** : distinguish whether the input x is from real data
- Generator G** : generate plausible data \tilde{x} given noise z and a input x
- Discriminator D** : 1-D convolutional neural network (CNN)
- Generator G** : 1-D DCNN with an encoder-decoder CNN
 - Generate plausible data \tilde{x} & reconstruct input x by \bar{x}

$$\mathbb{E}_{x \sim p_d(x)} \left[\rho \cdot \mathbb{E}_{\tilde{x} \sim p_g(\tilde{x}|x)} [-\log D(\tilde{x})] + (1 - \rho) \cdot \|\bar{x} - x\|_2^2 \right]$$

- ρ : control the similarity between the generated \tilde{x} and the input x
 - Small ρ works better

Semi-supervised Learning (SSL)

- G learns **rich structures of the data manifold** from real data
- G generates **plausible sample \tilde{x}_i around x_i** , possibly with label y_i
- \tilde{x}_i provides additional information of the data
- **SSL-GAN**: Data augmentation using data from *ehrGAN*
 - Objective function given N **training samples** and **generated samples**

$$\ell = \frac{1}{N} \sum_{i=1}^N \mathcal{L}(x_i, y_i) + \mu \cdot \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\tilde{x}_i \sim p_g(\tilde{x}_i|x_i)} \mathcal{L}(\tilde{x}_i, y_i)$$

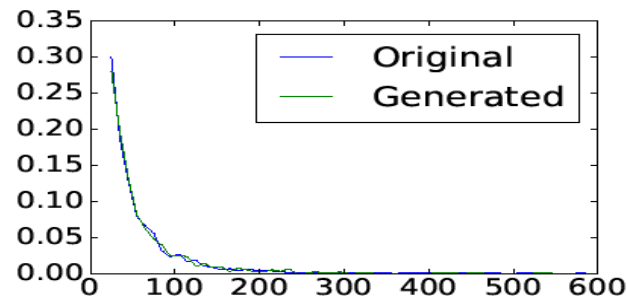
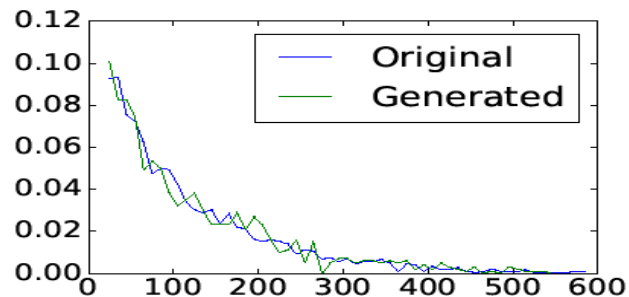
- \mathcal{L} : binary cross-entropy
- μ : leverage the usage of **real data** and **augmented data**
 - $\mu = 0.6$ performs best

Experimental Settings

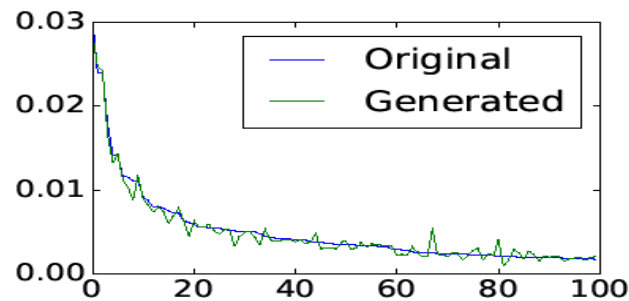
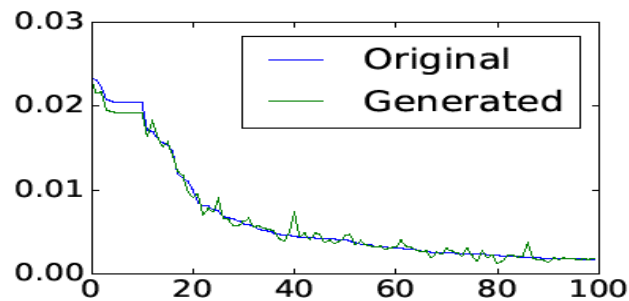
- A private real-world EHR dataset from a hospital
 - Medication (generic name) and diagnosis (ICD-9 code) records
 - 218,680 patients, 14,969,489 observations, 14,690 unique features
 - The entire dataset is used to train the word2vec embedding
- Two binary classification tasks on a real-world dataset
 - Congestive heart failure (**HF**): 3357 confirmed patients (case group)
 - Diabetes (**Dia**): 2248 confirmed patients (case group)
 - Two settings:
 - HF-1/2, Dia-1/2: case group : control group = 1:1
 - HF-1/3, Dia-1/3: case group : control group = 1:2
 - The ehrGAN is trained on only the training dataset for each setting

Generated Data Analyses (Left: HF; Right: Dia)

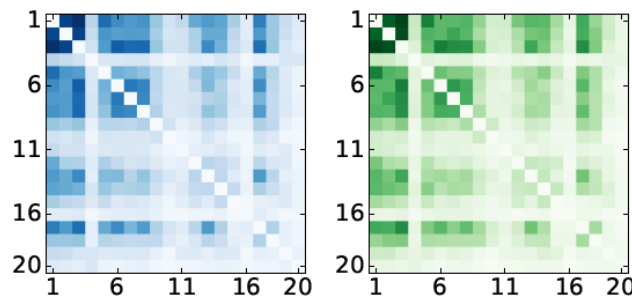
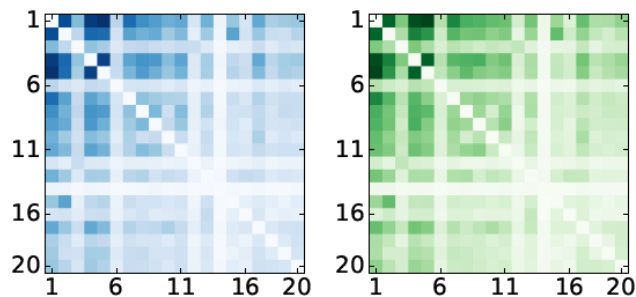
- Length distribution - **similar**



- Feature distribution - **similar**



- Feature co-occurrence - **similar**



Classification Results

- SSL-GAN outperforms CNN and other SSL baselines [Niu et al., 2013, Zhou et al., 2004]

	HF-1/2		HF-1/3	
	Acc.	AUC	Acc.	AUC
CNN	0.8096	0.8784	0.8347	0.8953
SSL-SMIR	0.8207	0.8842	0.8466	0.9102
SSL-LGC	0.8119	0.8767	0.8325	0.9011
SSL-GAN	0.8574	0.9075	0.8662	0.9246

	Dia-1/2		Dia-1/3	
	Acc.	AUC	Acc.	AUC
CNN	0.8990	0.9156	0.9129	0.9386
SSL-SMIR	0.9089	0.9197	0.9038	0.9277
SSL-LGC	0.8844	0.9102	0.8815	0.9128
SSL-GAN	0.9135	0.9354	0.9330	0.9563

- SSL-GAN achieves close performance as fully real data is given

	HF-1/2		HF-1/3	
	Acc.	AUC	Acc.	AUC
<i>CNN-Basic</i>	0.8096	0.8784	0.8347	0.8953
<i>CNN-Rand</i>	0.7418	0.7856	0.7788	0.8117
<i>CNN-Full</i>	<i>0.8631</i>	<i>0.9212</i>	<i>0.8749</i>	<i>0.9329</i>
SSL-GAN	0.8574	0.9075	0.8662	0.9246

	Dia-1/2		Dia-1/3	
	Acc.	AUC	Acc.	AUC
<i>CNN-Basic</i>	0.8990	0.9156	0.9129	0.9386
<i>CNN-Rand</i>	0.7734	0.8011	0.7969	0.8486
<i>CNN-Full</i>	<i>0.9335</i>	<i>0.9528</i>	<i>0.9486</i>	<i>0.9714</i>
SSL-GAN	0.9135	0.9354	0.9330	0.9563

Conclusion

- Summary
 - **ehrGAN**: learn generates plausible labeled EHR data
 - **SSL-GAN**: boost prediction performance by generated data
 - Superior performance on a real-world EHR dataset
- Future work
 - Clinical analysis on generated data
 - More challenging healthcare prediction tasks
 - Jointly-trained predictive and generative models
- Contact
 - IBM Research
 - USC Melady Lab

Thanks!

Zhengping Che
zche@usc.edu



References

- [Che & Cheng et al., 2017]
 - Che, Zhengping, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. "Boosting Deep Learning Risk Prediction with Generative Adversarial Networks for Electronic Health Records." arXiv preprint arXiv:1709.01648 (2017).
- [Che et al., 2017]
 - Che, Zhengping, Yu Cheng, Zhaonan Sun, and Yan Liu. "Exploiting Convolutional Neural Network for Risk Prediction with Medical Feature Embedding." arXiv preprint arXiv:1701.07474 (2017).
- [Goodfellow et al., 2014]
 - Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In Advances in neural information processing systems, pp. 2672-2680. 2014.
- [ImageNet]
 - <http://image-net.org/>
- [Mikolov et al., 2013]
 - Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
- [Niu et al., 2013]
 - Niu, Gang, Wittawat Jitkrittum, Bo Dai, Hirotaka Hachiya, and Masashi Sugiyama. "Squared-loss mutual information regularization: A novel information-theoretic approach to semi-supervised learning." In International Conference on Machine Learning, pp. 10-18. 2013.
- [Zhai et al., 2016]
 - Zhai, Shuangfei, Yu Cheng, Rogerio Feris, and Zhongfei Zhang. "Generative Adversarial Networks as Variational Training of Energy Based Models." arXiv preprint arXiv:1611.01799 (2016).
- [Zhou et al., 2004]
 - Zhou, Denny, Olivier Bousquet, Thomas N. Lal, Jason Weston, and Bernhard Schölkopf. "Learning with local and global consistency." In Advances in neural information processing systems, pp. 321-328. 2004.