



Deep learning based Feature Selection Study

Sanghyuck Lee

2021. 10. 19.

Outline

- INVASE: Instance-wise variable selection using neural networks
- 중견 연구 과제
- 갑상선 연구 과제
- 산학 협력 과제

INVASE: Instance-wise variable selection using neural networks

Paper Title: INVASE: Instance-wise variable selection using neural networks

Authors: Yoon, Jinsung et al.

Conference: ICLR 2019

Main Algorithm

1: **Inputs:** learning rates $\alpha, \beta > 0$, mini-batch size $n_{mb} > 0$, dataset $\mathcal{D}$

2: **Initialize** parameters θ, ϕ, γ

3: **while** Converge **do**

4: Sample a mini-batch from the dataset $(\mathbf{x}_j, y_j)_{j=1}^{n_{mb}} \sim \mathcal{D}$

5: **for** $j = 1, \dots, n_{mb}$ **do**

6: Calculate selection probabilities

$$(p_1^j, \dots, p_d^j) \leftarrow \hat{S}^\theta(\mathbf{x}_j)$$

7: Sample selection vector

8: **for** $i = 1, \dots, d$ **do**

$$s_i^j \sim \text{Ber}(p_i^j)$$

9: Calculate loss

$$\hat{l}_j(\mathbf{x}_j, \mathbf{s}_j) \leftarrow - \left[\sum_{i=1}^c y_i^j \log(f_i^\phi(\mathbf{x}_j^{(\mathbf{s}_j)}, \mathbf{s}_j)) - \sum_{i=1}^c y_i^j \log(f_i^\gamma(\mathbf{x}_j)) \right]$$

10: Update the selector network parameters θ

$$\theta \leftarrow \theta - \alpha \frac{1}{n_{mb}} \sum_{j=1}^{n_{mb}} \left(\hat{l}_j(\mathbf{x}_j, \mathbf{s}_j) + \lambda \|\mathbf{s}_j\| \right) \nabla_\theta \log \pi_\theta(\mathbf{x}_j, \mathbf{s}_j)$$

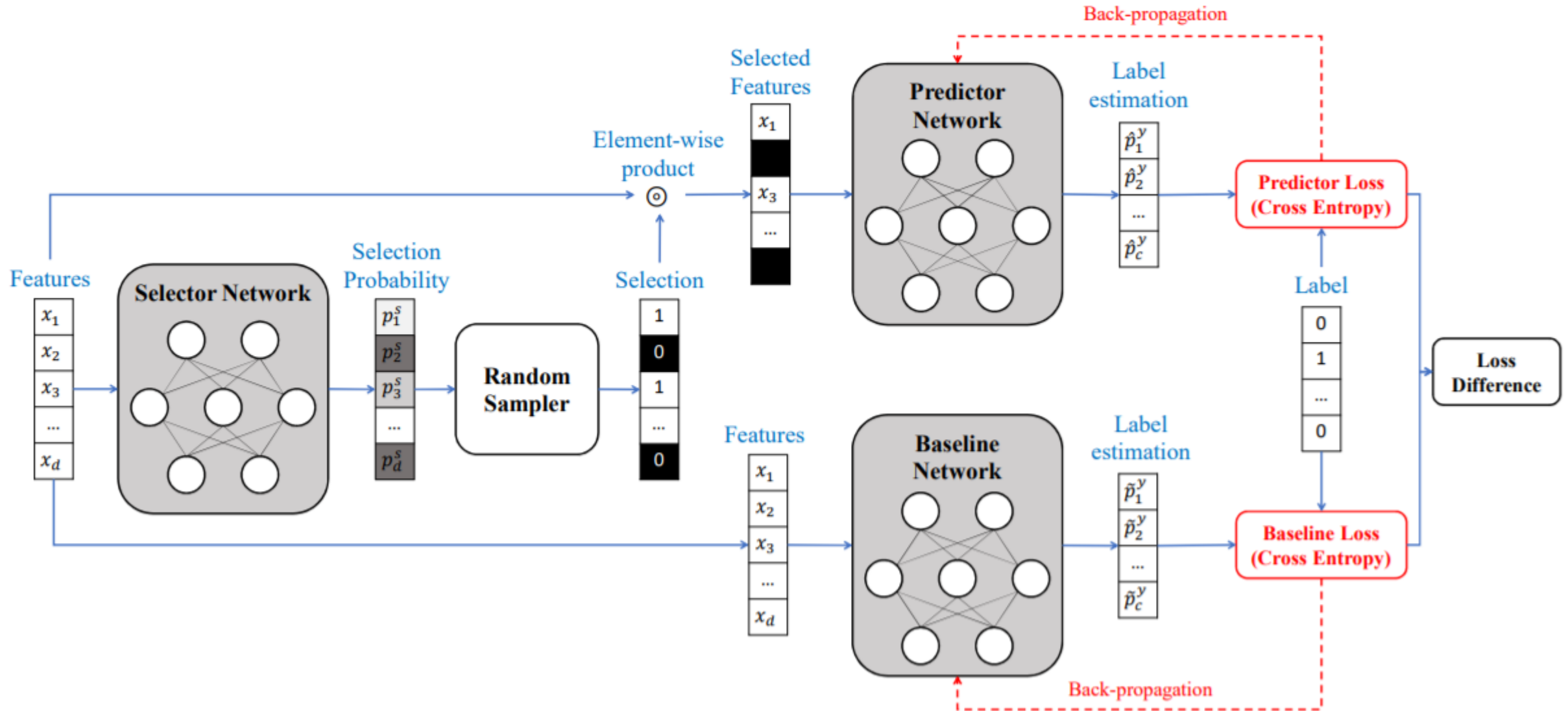
11: Update the predictor network parameters ϕ

$$\phi \leftarrow \phi - \beta \frac{1}{n_{mb}} \sum_{j=1}^{n_{mb}} \sum_{i=1}^c y_i^j \times \nabla_\phi \log(f_i^\phi(\mathbf{x}_j^{(\mathbf{s}_j)}, \mathbf{s}_j))$$

12: Update the baseline network parameters γ

$$\gamma \leftarrow \gamma - \beta \frac{1}{n_{mb}} \sum_{j=1}^{n_{mb}} \sum_{i=1}^c y_i^j \times \nabla_\gamma \log(f_i^\gamma(\mathbf{x}_j))$$

Main Algorithm



Main Algorithm

1: **Inputs:** learning rates $\alpha, \beta > 0$, mini-batch size $n_{mb} > 0$, dataset $\mathcal{D}$

2: **Initialize** parameters θ, ϕ, γ

3: **while** Converge **do**

4: Sample a mini-batch from the dataset $(\mathbf{x}_j, y_j)_{j=1}^{n_{mb}} \sim \mathcal{D}$

5: **for** $j = 1, \dots, n_{mb}$ **do**

6: Calculate selection probabilities

$$(p_1^j, \dots, p_d^j) \leftarrow \hat{S}^\theta(\mathbf{x}_j)$$

7: Sample selection vector

8: **for** $i = 1, \dots, d$ **do**

$$s_i^j \sim \text{Ber}(p_i^j)$$

9: Calculate loss

$$\hat{l}_j(\mathbf{x}_j, \mathbf{s}_j) \leftarrow - \left[\sum_{i=1}^c y_i^j \log(f_i^\phi(\mathbf{x}_j^{(\mathbf{s}_j)}, \mathbf{s}_j)) - \sum_{i=1}^c y_i^j \log(f_i^\gamma(\mathbf{x}_j)) \right]$$

10: Update the selector network parameters θ

$$\theta \leftarrow \theta - \alpha \frac{1}{n_{mb}} \sum_{j=1}^{n_{mb}} \left(\hat{l}_j(\mathbf{x}_j, \mathbf{s}_j) + \lambda \|\mathbf{s}_j\| \right) \nabla_\theta \log \pi_\theta(\mathbf{x}_j, \mathbf{s}_j)$$

11: Update the predictor network parameters ϕ

$$\phi \leftarrow \phi - \beta \frac{1}{n_{mb}} \sum_{j=1}^{n_{mb}} \sum_{i=1}^c y_i^j \times \nabla_\phi \log(f_i^\phi(\mathbf{x}_j^{(\mathbf{s}_j)}, \mathbf{s}_j))$$

12: Update the baseline network parameters γ

$$\gamma \leftarrow \gamma - \beta \frac{1}{n_{mb}} \sum_{j=1}^{n_{mb}} \sum_{i=1}^c y_i^j \times \nabla_\gamma \log(f_i^\gamma(\mathbf{x}_j))$$

Experiments

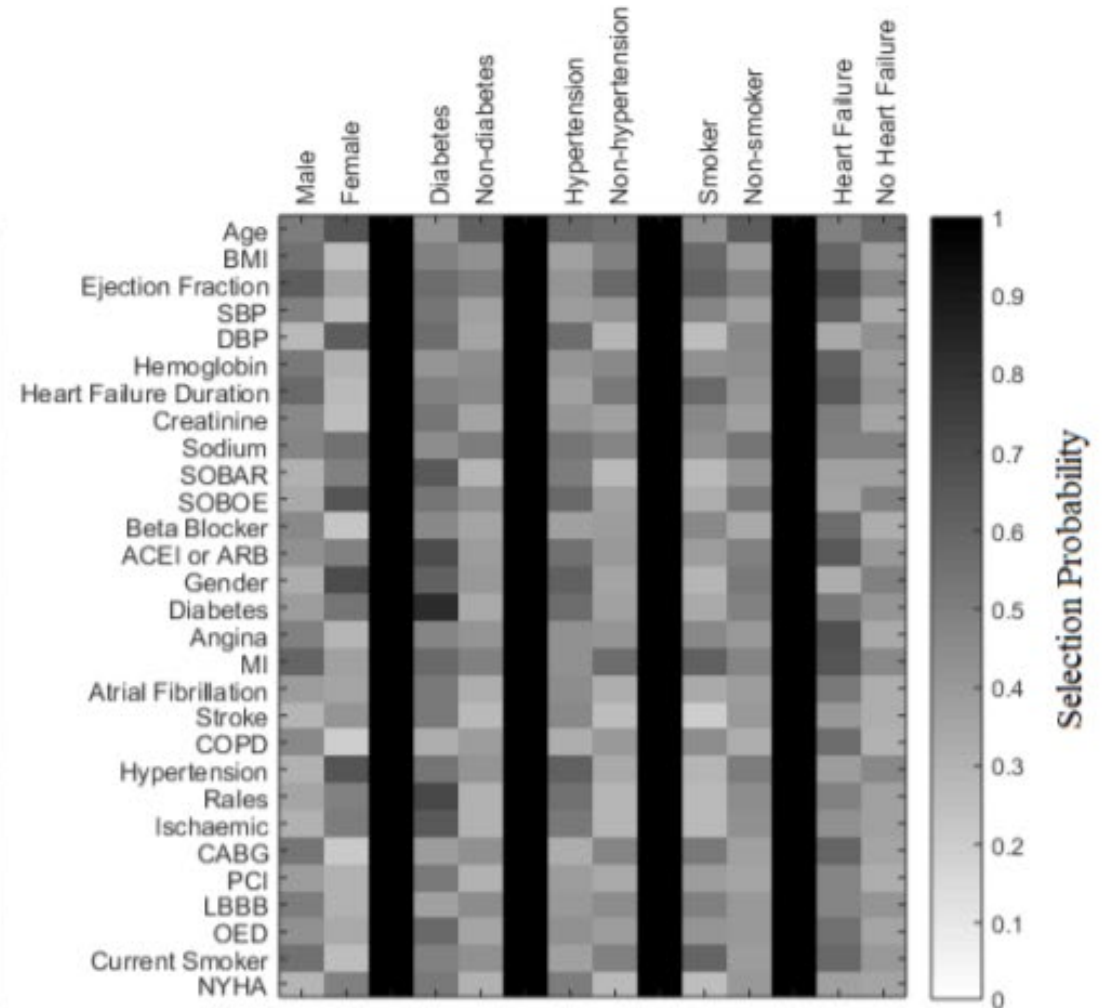
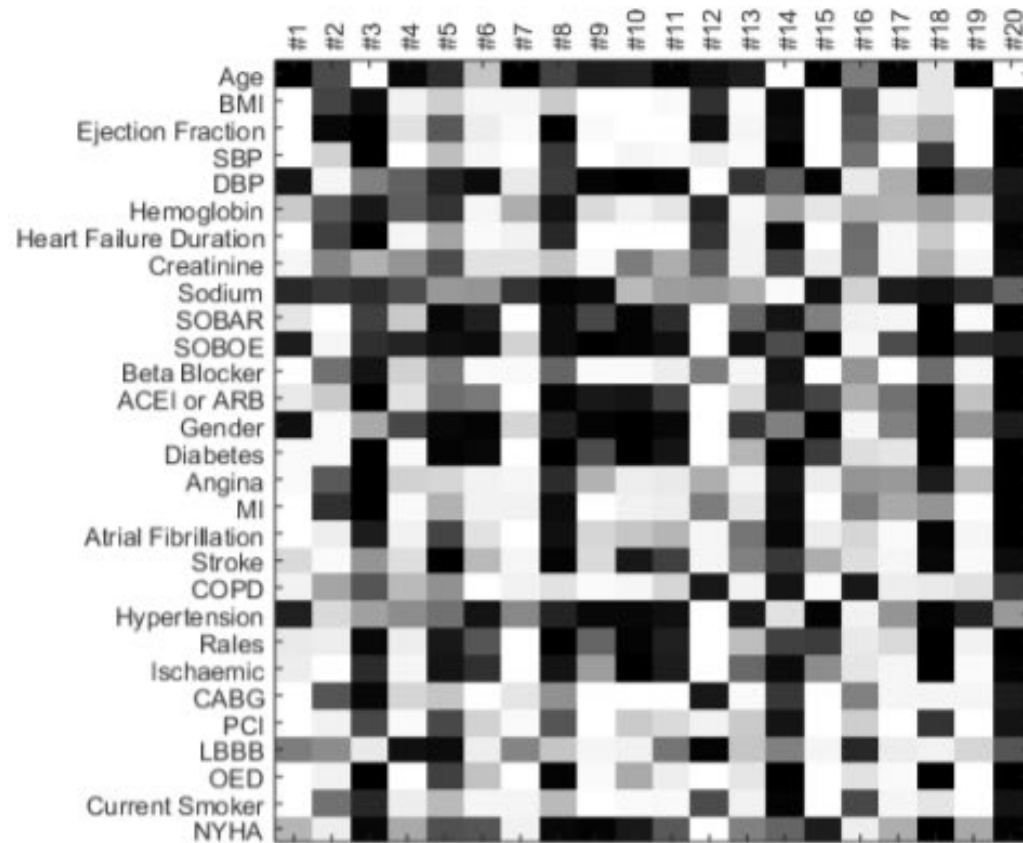
Dataset	Syn1		Syn2		Syn3		Syn4		Syn5		Syn6	
Metrics (%)	TPR	FDR	TPR	FDR	TPR	FDR	TPR	FDR	TPR	FDR	TPR	FDR
INVASE	100.0	0.0	100.0	0.0	92.0	0.0	99.8	10.3	84.8	1.1	90.1	7.4
L2X	100.0	0.0	100.0	0.0	69.4	30.6	79.5	21.8	74.8	26.3	83.3	16.7
LIME	13.8	86.2	100.0	0.0	98.1	1.9	40.7	49.4	41.1	50.6	50.5	49.5
Shapley	60.4	39.6	93.3	6.7	90.9	9.1	65.2	31.9	62.9	33.7	71.2	28.8
Knockoff	10.0	70.0	8.7	36.2	81.2	17.5	38.8	35.1	41.0	51.1	56.6	42.1
Tree	100.0	0.0	100.0	0.0	100.0	0.0	54.7	39.0	56.8	37.5	60.0	40.0
SCFS	23.5	76.5	39.5	60.5	78.3	22.0	48.9	52.4	42.4	51.2	56.1	43.9
LASSO	19.0	81.0	39.8	60.2	78.3	21.7	49.9	50.9	45.5	48.2	56.4	43.6

Experiments

Dataset	AUROC					
	w/o FS	with Global	with INVASE	with Tree	with L2X	with LASSO
Syn1	.578±.004	.686±.005	.690±.006	.574±.101	.498±.005	.498±.006
Syn2	.789±.003	.873±.003	.877±.003	.872±.003	.823±.029	.555±.061
Syn3	.854±.004	.900±.003	.902±.003	.899±.001	.862±.009	.886±.003
Syn4	.558±.021	.774±.006	.787±.004	.684±.017	.678±.024	.514±.031
Syn5	.662±.013	.784±.005	.784±.005	.741±.004	.709±.008	.691±.024
Syn6	.692±.015	.858±.004	.877±.003	.771±.031	.827±.017	.727±.025

Dataset	AUPRC					
	w/o FS	with Global	with INVASE	with Tree	with L2X	with LASSO
Syn1	.567±.007	.690±.006	.694±.006	.577±.102	.498±.007	.499±.008
Syn2	.799±.005	.878±.005	.886±.004	.878±.004	.817±.031	.591±.037
Syn3	.861±.003	.905±.002	.907±.003	.904±.002	.860±.012	.890±.002
Syn4	.572±.019	.794±.006	.804±.004	.681±.031	.672±.025	.536±.025
Syn5	.665±.019	.796±.005	.797±.006	.765±.003	.719±.011	.680±.040
Syn6	.709±.018	.870±.005	.886±.004	.779±.027	.835±.017	.757±.036

Experiments



Experiments

Datasets	Metrics	AUROC	AUPRC	AUROC	AUPRC
	Labels	3 year		5 year	
MAGGIC	INVASE	.722±.005	.655±.010	.740±.005	.867±.006
	Without INVASE	.720±.006	.639±.009	.730±.006	.855±.004
	Labels	5 year		10 year	
PLCO	INVASE	.637±.007	.329±.013	.673±.007	.506±.006
	Without INVASE	.629±.008	.324±.011	.657±.006	.485±.008

Feature Selection, Interpretation, Prediction

	Key ideas	Experiments shown	Global/ Instance-wise	Model agnostic	# of relevant features
SCFS [11]	Max-dependency min-redundancy criteria with Pearson correlations	Feature selection	Global	Yes	Not needed
MIFS [21]	Max-dependency min-redundancy criteria with Mutual Information	Feature selection	Global	Yes	Not needed
LASSO [31]	Linear regression with l_1 -norm penalty	Feature selection Prediction	Global	Yes	Not needed
Knock-off [3]	Comparison between knock-off variables and real variables	Feature selection Hypothesis test	Global	Yes	Not needed
L2X [4]	Mutual Information maximization with Gumbel-softmax	Interpretation	Instance-wise	Yes	Should be given
LIME [24]	Locally linear approximation	Interpretation	Instance-wise	Yes	Should be given
Shapley [18]	Shapley value estimation to quantify feature importance	Feature selection	Instance-wise	Yes	Should be given
DeepLIFT [27]	Decompose the output of NN on a reference input	Interpretation	Instance-wise	No	Should be given
Saliency [29]	Backpropagation from the output of the NN to the input	Interpretation	Instance-wise	No	Should be given
Tree SHAP [19]	Shapley value estimation only for tree-ensemble models	Interpretation	Instance-wise	No	Should be given
Pixel-wise [1]	Measuring the effects on the output using input perturbation	Interpretation	Instance-wise	No	Should be given
INVASE (Ours)	Minimize KL divergence using deep NN influenced by actor-critic models	Feature selection Interpretation Prediction	Instance-wise	Yes	Not needed

Prediction

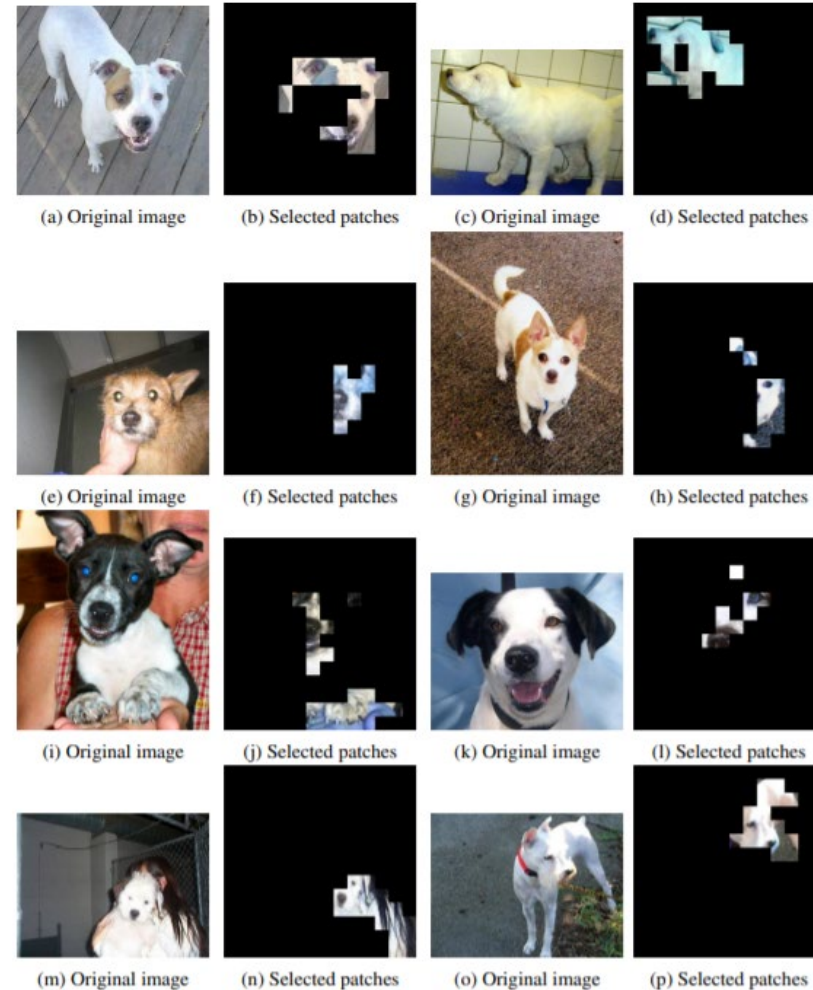
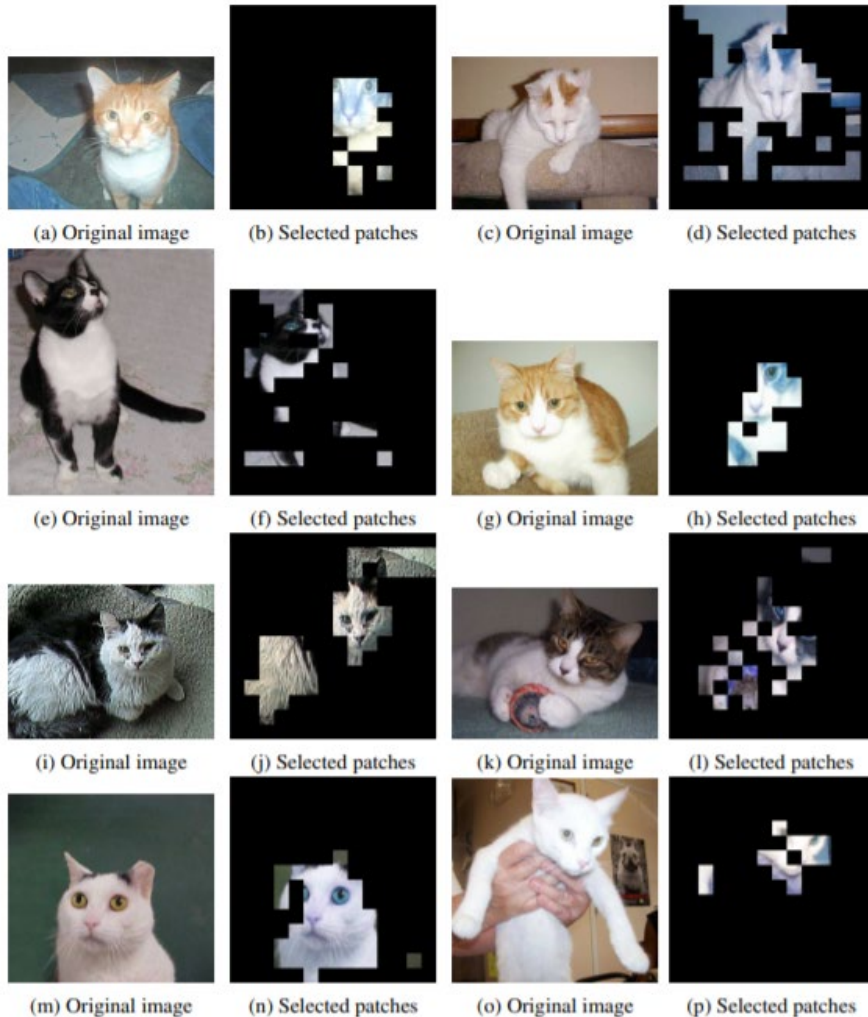
Datasets	MAGGIC				PLCO			
Labels	3-year		5-year		5-year		10-year	
Metrics	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
INVASE	0.722	0.655	0.740	0.867	0.637	0.329	0.673	0.506
L2X	0.609	0.529	0.607	0.794	0.558	0.170	0.583	0.365
LIME	0.637	0.5596	0.634	0.808	0.597	0.183	0.601	0.374
Shapley	0.641	0.557	0.617	0.797	0.614	0.194	0.615	0.381
Knockoff	0.686	0.614	0.711	0.853	0.619	0.230	0.658	0.475
Tree	0.678	0.604	0.708	0.850	0.632	0.269	0.655	0.469
SCFS	0.683	0.623	0.723	0.857	0.632	0.231	0.632	0.444
LASSO	0.692	0.615	0.709	0.847	0.623	0.218	0.656	0.467

Regression & Computer Vision

- EXTENDING INVASE TO REGRESSION
- COMPUTER VISION

U Net as **Selector Network**, VGG Net as **Predictor & Baseline Network**

Regression & Computer Vision



Next Step

- INVASE based Prediction Performance Improvement
 - 갑상선 연구 : Computer Vision, Regression (부피 추론)
 - Choose Predictor Network and Experiment Dataset
- Next Week: Basic Implementation with some specific decisions related to
above various direction.

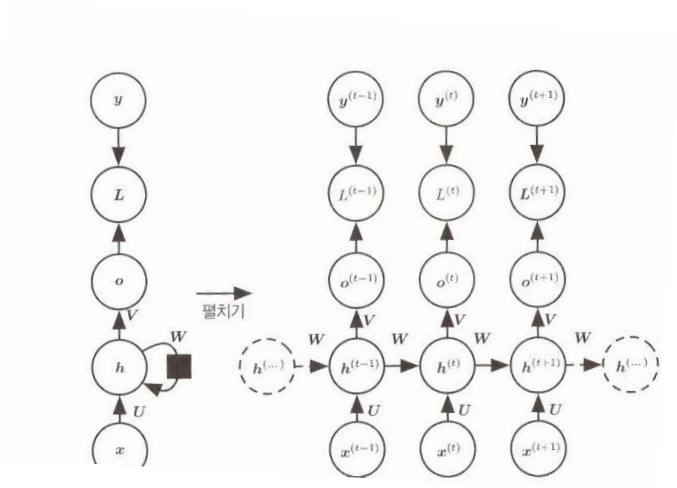
시계열 자료 연구 관련

10.2 순환 신경망

§10.1에서 설명한 그래프 펼치기와 매개변수 공유 개념에 기초해서 다양한 종류의 순환 신경망을 설계할 수 있다.

다음은 순환 신경망의 설계에서 자주 볼 수 있는 중요한 패턴 몇 가지이다.

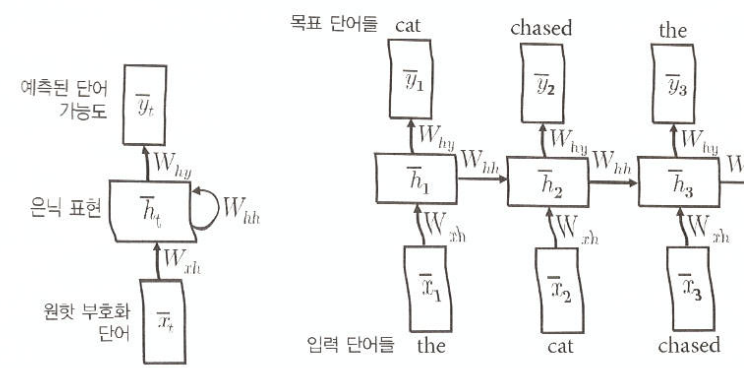
- 각 시간 단계에서 하나의 출력을 산출하며, 은닉 단위들 사이에 순환 연결들이 존재하는 순환 신경망. 그림 10.3이 그러한 예이다.
- 각 시간 단계에서 하나의 출력을 산출하며, 한 시간 단계의 출력과 그다음 시간 단계의 은닉 단위들 사이에만 순환 연결들이 존재하는 순환 신경망. 그림 10.4가 그러한 예이다.
- 은닉 단위들 사이에 순환 연결들이 존재하고, 순차열 전체를 읽어서 하나의 출력을 산출하는 순환 신경망. 그림 10.5가 그러한 예이다.



들로부터 만들어진 은닉 상태들과 직접 상호작용할 수 있다. 그림 1.17(a)는 순환 신경망의 기본적인 구조를 나타낸 것이다. 여기서 핵심은 각 시간에 하나의 입력 $\bar{x}_t$ 가 있다는 점과 각 은닉 상태(hidden state) $\bar{h}_t$ 는 각 시간에서 새 자료점이 입력됨에 따라 변한다는 것이다. 각 시간에는 또한 출력값 $\bar{y}_t$ 도 있다. 예를 들어 시계열 자료의 예에서 출력 $\bar{y}_t$ 는 $\bar{x}_{t+1}$ 을 미리 예측한 값일 수 있다. 텍스트 자료에서 다음 단어를 예측하는 것을 가리켜 언어 모형화(language modeling)라고 부른다. 응용에 따라서는 매 시간에서 출력 $\bar{y}_t$ 를 산출하는 것이 아니라 순차열의 끝에서만 출력을 산출하는 경우도 있다. 예를 들어 주어진 문장에 담긴 정서가 ‘긍정적’인지 ‘부정적’인지 분류하는 응용의 경우 마지막 시간(문장의 끝 단어)에서만 분류 결과를 출력한다.

시간 t 에서의 은닉 상태는 시간 t 에서의 입력 벡터와 시간 $(t-1)$ 에서의 은닉 벡터의 함수이다.

$$\bar{h}_t = f(\bar{h}_{t-1}, \bar{x}_t) \tag{1.36}$$



Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
Aggarwal, Charu C. "Neural networks and deep learning." *Springer* 10 (2018): 978-3.

중견 연구 제안서 작성

- 심미성 제안서 기반으로 재작성 진행중

갑상선 연구: 안구 부피 추론

총 76개 특징 중 2개의 특징 선택 완료, 3번째 특징을 위한 후보 특징들 성능 확인중, 현재 48/74번째 특징 확인 완료
이를 위해, 14 시간 정도 소요되었음.

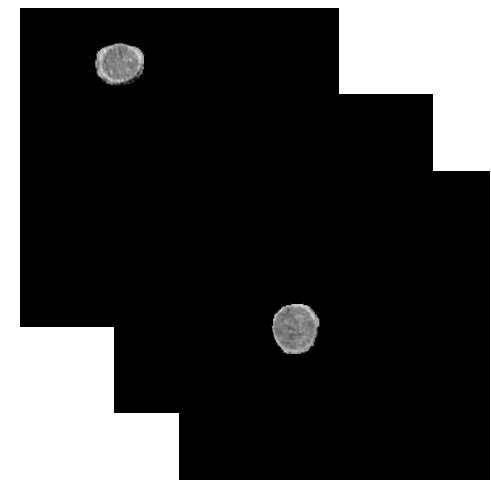
순위	선택된 구조물	R^2
1	LT 눈, AX2 눈 (left), AX1 내직근 (right)	0.630
2	LT 눈	0.545
3	LT 눈, AX2 눈 (left), AX1 시신경 (left)	0.540
4	LT 눈, AX2 눈 (left), AX5 눈 (right)	0.525
5	RT 눈	0.524
6	LT 눈, AX2 눈 (left)	0.518
7	LT 눈, AX2 시신경 (right)	0.501
8	LT 눈, AX2 눈 (left), AX4 시신경 (right)	0.498
9	LT 눈, AX2 눈 (left), AX3 내직근 (right)	0.486
10	LT 눈, AX4 눈 (right)	0.482

순위	선택된 구조물	R^2
11	AX2 눈 (left)	0.476
12	LT 눈, CO1 눈 (right)	0.466
13	LT 눈, AX2 외직근 (left)	0.465
14	LT 눈, CO2 내직근 (right)	0.465
15	LT 눈, AX5 외직근 (left)	0.447
16	AX3 눈 (left)	0.443
17	AX3 눈 (right)	0.442
18	LT 눈, AX2 눈 (right)	0.425
19	LT 눈, AX1 눈 (left)	0.424
20	LT 눈, AX2 눈 (left), AX2 외직근 (right)	0.420

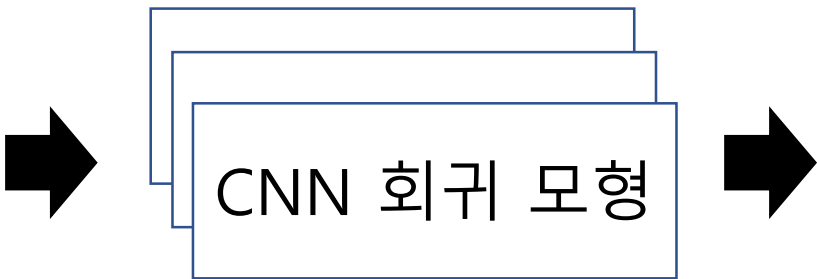
갑상선 연구: 안구 부피 추론 알고리즘

데이터 분할 : 훈련셋, 검증셋, 테스트셋, 최종 테스트셋

1번째 반복



76개 유형의 특징



훈련셋, 검증셋을 이용해 학습

76개의 모형 학습

순위	특징 집합	성능
1	LT 눈	0.xxx
	~	
76	AX5 눈 (right)	0.xxx

테스트셋을 이용해 성능 확인

76개의 성능

2번째 반복

1위 랭크 특징+75개 유형의 특징 ➡ 75개의 모형 학습 ➡ 75개의 성능

갑상선 연구: 안구 부피 추론 알고리즘

- 3번째 반복 : 2번째 반복에서, 1위 실험 특징 집합을 가져와, (76-2)번의 실험을 진행
~
- N번째 반복 : (N-1) 반복에서, 1위 실험 특징 집합을 가져와, (76-(N-1))번의 실험의 진행
- 종료 조건 : 성능 증가 중지 확인, or 전체 반복 종료 (76번째 반복)
- 최종 성능 평가 : 최종 테스트셋으로 최종 예측 성능 검증
- CNN 회귀 모형 : Chaganti et al. (2019) 참고함. CT 시신경의 부피를 측정한다는 점이 유사하나, 전처리 과정이 달라, 동일한 네트워크 디자인이 어렵고, 유사하게 해도, 성능이 나오지 않아, 다른 회귀 모형을 함께 참고해서, 변형 진행함.

Chaganti, Shikha, et al. "Contextual deep regression network for volume estimation in orbital CT." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2019.

갑상선 연구: 안구 부피 추론

- 전달 받은 테이블 파일 관련 : **왼쪽 눈의 부피** – 전문의들과 저희 측의 왼쪽, 오른쪽 지칭이 반대로 되어 있어, 확인 필요함.

산학 협력 프로젝트

- 결과보고서 작성 완료, 제출 계획 중