Deep learning based Feature Selection Study

Sanghyuck Lee

2021. 12. 28.

To DO List

➢Last Week

- Related Works Study and Implementation
- 갑상선 연구 보고 : Active vs Inactive 실험 보고

➢This Week

- Related Works Study and Implementation
- 갑상선 연구 보고 : 세그멘테이션 정보 활용 실험 보고

Papers

Chen, Jianbo, et al. "Learning to explain: An information-theoretic perspective on model interpretation." *International Conference on Machine Learning*. PMLR, 2018.

Yoon, Jinsung, James Jordon, and Mihaela van der Schaar. "INVASE: Instance-wise variable selection using neural networks." *International Conference on Learning Representations*. 2018.

Schwab, Patrick, and Walter Karlen. "Cxplain: Causal explanations for model interpretation under uncertainty." *arXiv preprint arXiv:1910.12336* (2019).

Jethani, Neil, et al. "Have We Learned to Explain?: How Interpretability Methods Can Learn to Encode Predictions in their Interpretations." *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021.

Panda, Pranoy, Sai Srinivas Kancheti, and Vineeth N. Balasubramanian. "Instance-wise Causal Feature Selection for Model Interpretation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Learning to Explain (L2X) – ICML 2018

Problem Setting

$$\max_{\mathcal{E}} I(X_S; Y) \quad \text{subject to} \quad S \sim \mathcal{E}(X). \Rightarrow I(X;Y) = \mathbb{E}_{X,Y}\left[\log \frac{p_{XY}(X,Y)}{p_X(X)p_Y(Y)}\right]$$

$$I(X_S;Y) = \mathbb{E}\Big[\log \frac{\mathbb{P}_m(X_S,Y)}{\mathbb{P}(X_S)\mathbb{P}_m(Y)}\Big] = \mathbb{E}\Big[\log \frac{\mathbb{P}_m(Y|X_S)}{\mathbb{P}_m(Y)}\Big]$$

$$\Rightarrow \quad = \mathbb{E}\Big[\log \mathbb{P}_m(Y|X_S)\Big] + \text{Const.}$$

$$= \mathbb{E}_X \mathbb{E}_{S|X} \mathbb{E}_{Y|X_S}\Big[\log \mathbb{P}_m(Y|X_S)\Big] + \text{Const.}$$

Idea: Tractable Variational Lower Bound

$$\mathcal{Q} := \Big\{\mathbb{Q} \mid \mathbb{Q} = \{x_S \to \mathbb{Q}_S(Y|x_S), S \in \wp_k\}\Big\}. \Rightarrow \mathbb{E}_{Y|X_S}[\log \mathbb{P}_m(Y|X_S)] \geq \int \mathbb{P}_m(Y|X_S) \log \mathbb{Q}_S(Y|X_S)$$

$$= \mathbb{E}_{Y|X_S}[\log \mathbb{Q}_S(Y|X_S)],$$

Final Objective

$$\max_{\mathcal{E},\mathbb{Q}} \mathbb{E}\Big[\log \mathbb{Q}_S(Y \mid X_S)\Big] \quad \text{such that } S \sim \mathcal{E}(X). \Rightarrow \mathbb{E}_{X,\zeta}\Big[\sum_{y=1}^{c}[\mathbb{P}_m(y \mid X) \log g_\alpha(V(\theta,\zeta) \odot X, y)\Big].$$

INVASE – ICLR 2019

1) In L2X, they try to maximize a lower bound of the **mutual information** between the target Y and the selected input variables X_S. In contrast, we try to minimize the **KL divergence** between the conditional distributions $Y|X$ and $Y|X_S$.

2) In order to be able to **backpropagate through subset sampling**, L2X use the **Gumbel-softmax trick** [13] to approximately discretize the continuous outputs of the neural network. In our work, we use methods from actor-critic models [22] to bypass backpropagation through the sampling and instead use **the predictor network to provide a reward to the selector network.**

3) Due to the Gumbel-softmax used in L2X, **the number of variables to be detected must be fixed** in advance and is necessarily the same for every sample. The actor-critic methodology used in our model has no such limitations and so we are able to flexibly select a different number of relevant variables for each sample and instead **induce sparsity using an l0 penalty term**. In fact, using the actor-critic methodology allows us to directly use the l0 penalty term (which is not differentiable and therefore not practical to use in general).

CXPlain – NeurIPS 2019

Problem Setting

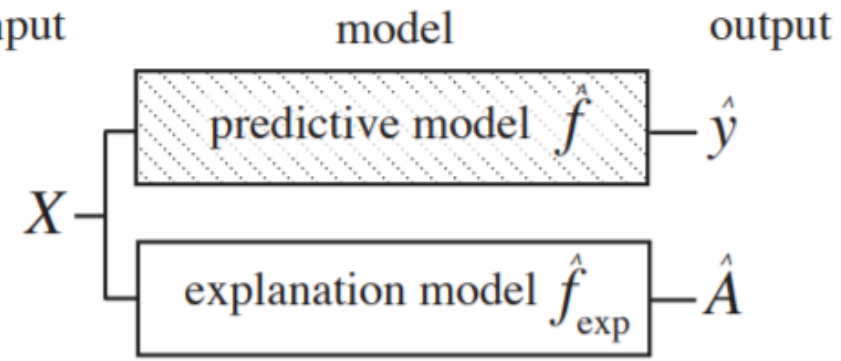
Goal: Train an explanation model $\hat{f}_{exp}$ that produces accurate estimates $\hat{A}$ with elements $\hat{a}_i$ corresponding to the importances assigned to each of the p input features x_i to the predictive model $\hat{f}$.

Idea: Causal Objective

$$\hat{y}_{X\setminus\{i\}} = \hat{f}(X \setminus \{i\}) \qquad \hat{y}_X = \hat{f}(X)$$

$$\varepsilon_{X\setminus\{i\}} = \mathcal{L}(y, \hat{y}_{X\setminus\{i\}}) \qquad \varepsilon_X = \mathcal{L}(y, \hat{y}_X)$$

$$\Delta\varepsilon_{X,i} = \varepsilon_{X\setminus\{i\}} - \varepsilon_X \Rightarrow \omega_i(X) = \frac{\Delta\varepsilon_{X,i}}{\sum_{j=0}^{p-1}\Delta\varepsilon_{X,j}} \Rightarrow \Omega \text{ with } \Omega(i) = \omega_i(X) \Rightarrow \textbf{Supervised Learning}$$



Real-X – NeurIPS 2020, AISTATS 2021,

Algorithm 1 REAL-X Algorithm

Input: $\mathcal{D} := (\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x} \in \mathbb{R}^{N \times D}$, feature matrix; $\boldsymbol{y} \in \mathbb{R}^{N}$, labels

Output: $q_{\text{sel}}(\cdot; \beta)$, function that returns feature selections given an instance of $\mathbf{x}$

Select: λ, regularization constant; α, learning rate; M, mini-batch size, T, training-steps

for $1, ..., T$ **do**

  Randomly sample mini-batch of size M, $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})_{i=1}^{M} \sim \mathcal{D}$

  for $i = 1, ..., M$ **do**

    Sample Selections:

      $\boldsymbol{r}^{(i)} \sim \text{Bernoulli}(0.5)$

      Sample $\boldsymbol{s}^{(i)}$, $\boldsymbol{z}^{(i)}$, and $\tilde{\boldsymbol{z}}^{(i)}$ using $q_{\text{sel}}(\cdot; \beta)$ as in eqs. (8) to (10)

  end

  Optimize Models:

    $\theta = \theta + \alpha \nabla_{\theta} \left[\frac{1}{M} \sum_{i=1}^{M} \log q_{\text{pred}}(\boldsymbol{y}^{(i)} | m(\boldsymbol{x}^{(i)}, \boldsymbol{r}^{(i)}; \theta) \right]$

    $\beta = \beta + \alpha \frac{1}{M} \sum_{i=1}^{M} \hat{g_{\beta}}$ ($\hat{g_{\beta}}$ as in eq. (11))

end

Structural Causal Model (SCM) – CVPR2021

Problem Setting and Idea (CS)
Causal Strength (CS) of a subset s going from input X to the response variable of the model Y,

$$CS_s = I\left(X_s; Y | X_{\bar{s}}\right)$$
$$CS_s = -H(Y|X) + H(Y|X_{\bar{s}})$$

$$\Rightarrow \quad \max_s CS_s \equiv \min_s E_{Y,X_{\bar{s}}}[\log(P(Y|X_{\bar{s}})]$$

We simply use the output of the black-box model F when X_S is given as input, **to estimate $\boldsymbol{P(Y|X_{\bar{S}})}$. $\boldsymbol{X_{\bar{S}}}$** is represented as follows: if $s_i = 0$, $X_{\bar{S}i} = \boldsymbol{X_i}$, else $X_{\bar{S}i} = 0$.

Final Objective

$$\min_\theta \; E_{X,Y,\zeta}[\log(\mathcal{F}(Z(\theta,\zeta) \odot X))]$$

$$= \min_\theta \; E_{X,\zeta}\left[\sum_{y=1}^{c} P(y|X)\log(\mathcal{F}(Z(\theta,\zeta) \odot X))\right]$$

To-do

- L2X, INVASE, CXPlain, Real-X, SCM -> Complex Image Datasets

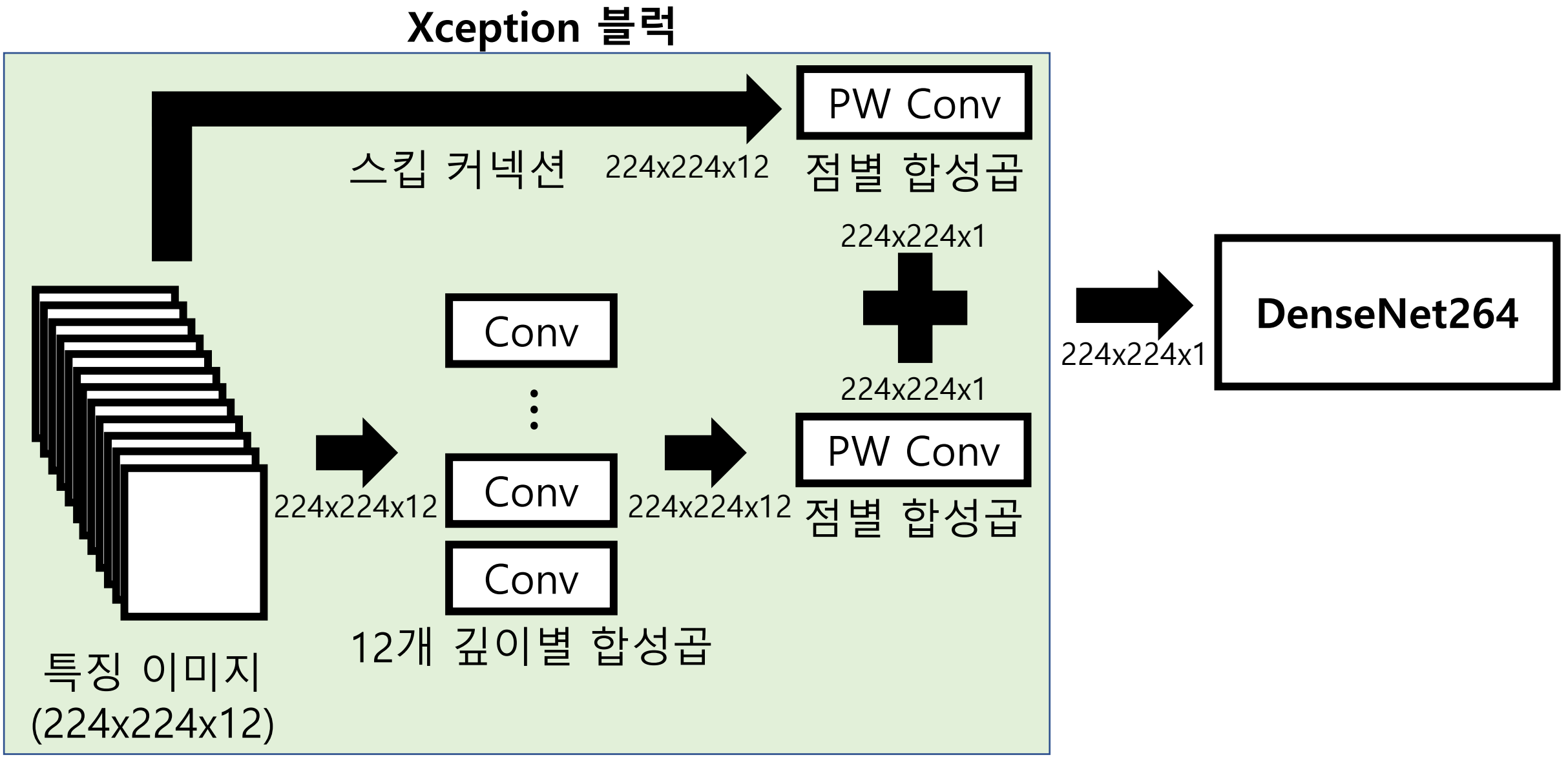
갑상선 연구 : 실험 1 (베이스 라인 재선정)

AX3의 특징 이미지 12개를 입력으로 사용 (12channels)

2회 반복 평균 : DenseNet264가 유망하다고 판단

Model Name	AUROC	Accuracy	AUPRC	Macro F1	Sensitivity	Specificity	Precision
DenseNet169	0.621	0.477	**0.549**	0.346	**0.794**	0.449	0.271
DenseNet264	**0.714**	0.729	0.526	**0.374**	0.676	0.752	0.324
InceptionV2	0.618	**0.865**	0.428	0.35	0.283	**0.952**	**0.479**
VGGNet16	0.595	0.696	0.38	0.283	0.449	0.741	0.236
VGGNet19	0.608	0.743	0.479	0.317	0.444	0.773	0.449

갑상선 연구 : 실험 2


Xception 블럭
PW Conv
스킵 커넥션
224x224x12
점별 합성곱
224x224x1
224x224x1
DenseNet264
224x224x1
Conv
Conv
Conv
224x224x12
224x224x12
PW Conv
점별 합성곱
12개 깊이별 합성곱
특징 이미지
(224x224x12)

갑상선 연구 : 실험 2

5회 반복 평균 : DenseNet264+Xception Block 실험

Table . Metrics with 95% confidence interval (▼/△ indicates that the corresponding method is **significantly worse** than the best model written in bold based on paired t-test at 5% significance level).

Model	AUROC			AUPRC			Accuracy			F1 Score			Sensitivity TP/(TP+FN)			Specificity TN/(TN+FP)			Precision TP/(TP+FP)		
	95% CI			95% CI			95% CI			95% CI			95% CI			95% CI			95% CI		
	Average	Lower Upper	p-value	Average	Lower Upper	p-value	Average	Lower Upper	p-value	Average	Lower Upper	p-value	Average	Lower Upper	p-value	Average	Lower Upper	p-value	Average	Lower Upper	p-value
DenseNet264 (1channel)	0.624	0.543 0.705	0.301	0.445	0.316 0.574	0.953	0.708	0.491 0.926	0.382	0.339	0.183 0.496	0.662	0.509	0.289 0.729	0.971	0.739	0.473 1.005	0.453	0.317	0.031 0.604	0.851
DenseNet264 (12channels)	0.664	0.561 0.768	0.807	0.442	0.302 0.582	0.884	0.772	0.662 0.882	0.354	0.352	0.267 0.437	0.862	**0.515**	0.174 0.857		0.814	0.649 0.978	0.444	0.305	0.172 0.439	0.22
DenseNet264 +Xception Block (12channels)	**0.673**	0.535 0.812		**0.449**	0.269 0.630		**0.807**	0.717 0.898		**0.362**	0.200 0.524		0.494	0.116 0.872	0.834	**0.853**	0.714 0.991		**0.341**	0.208 0.474	

CI, confidence interval; *: p-value < 0.05; **: p-value < 0.01

기타 이슈 사항 및 추후 진행 방향

- 224x224 size -> 512x512 size, 메모리 이슈
- CO 1, 2, 3으로 실험 예정