# Lightweighted Sparse Autoencoder based on Explainable Contribution

**Joohong Rheey** and **Hyunggon Park**

Graduate Program in Smart Factory, Department of Electronic and Electrical Engineering
Ewha Womans University, Seoul, Republic of Korea

## Introduction

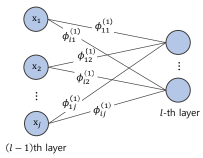
- **Motivations & Challenges**
  - **Performance improvement leads to heavy autoencoders**
  - **Heavy autoencoders require high computing powers**
    **→ Cannot be implemented into power-limited devices**
- **Objective: Design lightweight autoencoder while maintaining performance**

## Proposed Algorithm : SHAP-SAE

- **Link importance (LI) based on Shapley value**

$$\phi_{ij}^{(l)} = \sum_{J \subseteq I \setminus \{j\}} \frac{|J|!\,(|I|-|J|-1)!}{|I|!}\left(v(J \cup \{j\}) - v(J)\right)$$

- $I$ : set of links in $l$-th layer
- $J$ : subset excluding $j$-th link connected to $i$-th unit in $l$-th layer
- $v(J)$ : value of subset $J$




- **LI of link that connects $j$-th unit in $(l-1)$th layer and $i$-th unit in $l$-th layer**
- **Measurement of LI based on their contributions to output of layer**

- **Unit importance (UI)**

$$\bar{v}_j^{(l-1)} = \frac{1}{n^{(l)}} \sum_{i=1}^{n^{(l)}} \left|\phi_{ij}^{(l)}\right|$$

- **UI of $j$-th unit in $(l-1)$th layer**
- **Average impact on output of $l$-th layer**

- $n^{(l)}$ : number units in $l$-th layer

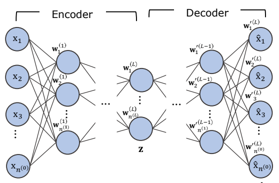
### Step 1: Training Stage

1. **Determine the set of optimal parameters for autoencoder**

$$\{\theta^*, \theta'^*\} = \underset{\theta,\theta'}{\operatorname{argmin}}\ \mathcal{L}(\mathbf{x}, \hat{\mathbf{x}})$$

- **Encoder:** $\theta = \{\mathbf{W}^{(l)}, \mathbf{b}^{(l)} | 1 \le l \le L\}$
- **Decoder:** $\theta' = \{\mathbf{W}'^{(l)}, \mathbf{b}'^{(l)} | L+1 \le l \le 2L\}$

- Weight matrix : $\{\mathbf{W}^{(l)}, \mathbf{W}'^{(l)}\} \in \mathbb{R}^{n^{(l)} \times n^{(l-1)}}$
- Bias vector : $\{\mathbf{b}^{(l)}, \mathbf{b}'^{(l)}\} \in \mathbb{R}^{n^{(l)}}$
- $2L$ : number of layers in autoencoder




Autoencoder

### Step 2: Sparsification Stage

1. **Total LI $\phi_T^{(l)}$ as sum of individual LIs in $l$-th layer**
2. **Set of descending ordered Shapley values in $l$-th layer**

$$\mathbf{\Phi}^{(l)} = \left[\Phi^{(l)}(1), \Phi^{(l)}(2), \cdots, \Phi^{(l)}(n^{(l-1)}n^{(l)})\right]$$

3. **Support set $\Gamma^{(l)}$**

$$\Gamma^{(l)} = \left\{(i,j) \,\middle|\, \sum_{k=1}^{k^*} \Phi^{(l)}(k) \ge m \cdot \phi_T^{(l)}\right\}$$

- $m$ : importance level ($0 < m \le 1$)
- $k, k^*$: integer

- **Set of the pairs $(i,j)$ of units $i$ and $j$ that have $k^*$ largest contribution**

**Example**

$\phi_T^{(l)} = 16$

| $\phi_{11}^{(l)}$ | $\phi_{12}^{(l)}$ | $\phi_{21}^{(l)}$ | $\phi_{22}^{(l)}$ |
|---|---|---|---|
| 1 | 8 | 4 | 3 |
| $\Phi^{(l)}(4)$ | $\Phi^{(l)}(1)$ | $\Phi^{(l)}(2)$ | $\Phi^{(l)}(3)$ |

➡ **Let $m = 0.75$**

$\therefore m \cdot \phi_T^{(l)} = 0.75 * 16 = 12$

➡ $\Gamma^{(l)} = \{(1,2), (2,1)\}$

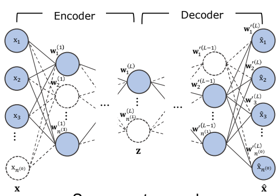
| $\phi_{12}^{(l)}$ | $\phi_{21}^{(l)}$ |
|---|---|
| 8 | 4 |
| $\Phi^{(l)}(1)$ | $\Phi^{(l)}(2)$ |

4. **Prune autoencoder using mask function $\mathcal{M}$**
   - **Mask function $\mathcal{M}$: $\mathbf{W}^* = \mathcal{M}(\mathbf{W})$ and $\mathbf{W}'^* = \mathcal{M}(\mathbf{W}')$**

$$\begin{cases} w_{ij}^{*(l)} = 0, & \text{if } (i,j) \in \Gamma^{(l)c} \quad \text{for } 1 \le l \le L \\ w_{ij}'^{*(l)} = 0, & \text{if } (i,j) \in \Gamma^{(l)c} \quad \text{for } L+1 \le l \le 2L \end{cases}$$

- $\Gamma^{(l)c}$ : complementary set of $\Gamma^{(l)}$
- $\mathbf{W}^*$ : pruned weight matrix of encoder
- $\mathbf{W}'^*$ : pruned weight matrix of decoder
- $w_{ij}^{*(l)}$ : element of weight matrix $\mathbf{W}^*$
- $w_{ij}'^{*(l)}$ : element of weight matrix $\mathbf{W}'^*$

- **The rest of weights remain unchanged**

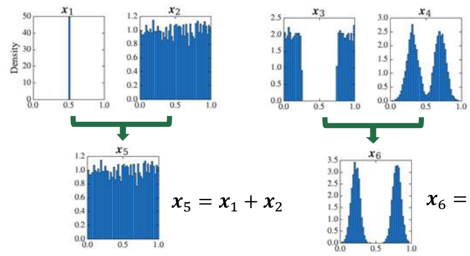



Sparse autoencoder

## Experiment Results

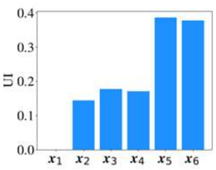
### SHAP-SAE with Synthetic dataset

- **Dataset**




$x_1, x_2, x_3, x_4$ **: independent**

- **Explainability**
  - **Impact of UIs on latent vector z**
  - $\bar{v}_5^{(0)} > \bar{v}_1^{(0)}$ or $\bar{v}_2^{(0)}$
    $\bar{v}_6^{(0)} > \bar{v}_3^{(0)}$ or $\bar{v}_4^{(0)}$
    - $x_5, x_6$: Include information of $x_1, x_2, x_3, x_4$
  - **Contribution of $x_1, x_2, x_3, x_4$ are marginal**
    - Redundant to $x_5, x_6$
  - $\bar{v}_1^{(0)} = 0$
    - $x_1$: Set of constant values → Irrelevant to **z**

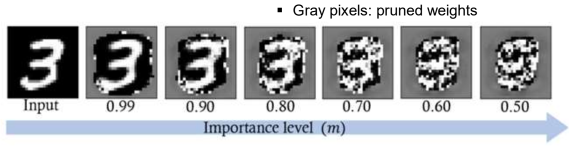



| | Variance of LIs |
|---|---|
| $x_1$ | 0 |
| $x_2$ | 0.0003 |
| $x_3$ | 0.0027 |
| $x_4$ | 0.0012 |
| $x_5$ | 0.0013 |
| $x_6$ | 0.0095 |

### SHAP-SAE with MNIST dataset

- **Performance Analysis**

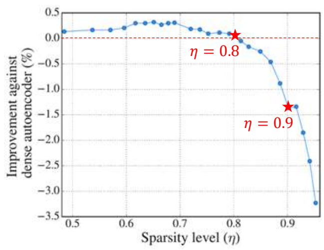
- Gray pixels: pruned weights



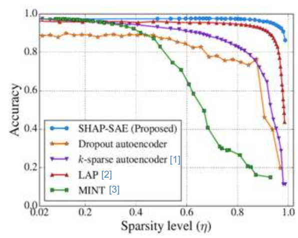

- **As importance level $m$ increases, mask function $\mathcal{M}$ removes less important weights**
- **Outperforms dense autoencoder up to $\eta = 0.8$**
- **Only 1.37% performance degradation with $\eta = 0.9$**

- Sparsity level $\eta$ : how many links are deactivated among all the links in autoencoder

$$\eta = \frac{\sum_{l=1}^{2L} |\Gamma^{(l)c}|}{\sum_{l=1}^{2L} |\Gamma^{(l)} \cup \Gamma^{(l)c}|}$$




- **Performance Comparisons**




- **Outperforms other benchmarks across all sparsity level**
- **Remarkably robust against high sparsity level**

## Conclusions

**SHAP-SAE algorithm can**

- **Explicitly measure unit and link importance based on Shapley value**
  **→ Activate only important units and links**
- **Allow sparse autoencoder to be explainable and robust against high sparsity level**
- **Be implemented into power-limited devices**

[1] Makhzani, A. and Frey, B., "K-sparse autoencoders. *International Conference on Learning Representations*," 2014.
[2] Park, S., Lee, J., Mo, S., and Shin, J., "Lookahead: A farsighted alternative of magnitude-based pruning," *International Conference on Learning Representations*, 2020.
[3] Ganesh, M. R., Corso, J. J., and Sekeh, S. Y, "Mint: Deep network compression via mutual information-based neuron trimming," *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 8251–8258. IEEE, 2021.