Seeking Interpretability and Explainability in Binary Activated Neural Networks

Benjamin Leblanc, Pascal Germain

August 31st, 2023

Seeking Interpretability and Explainability in Binary Activated Neural Networks

Benjamin Leblanc, Pascal Germain

August 31st, 2023

Introduction - Interpretability

We define the degree of interpretability

Introduction - Interpretability

We define the degree of interpretability of a predictor

Introduction - Interpretability

We define the degree of interpretability of a predictor by the capacity of a non-expert

Introduction - Interpretability

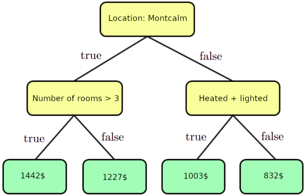
We define the degree of interpretability of a predictor by the capacity of a non-expert to understand its decision process

Introduction - Interpretability

We define the degree of interpretability of a predictor by the capacity of a non-expert to understand its decision process solely by considering the model in itself.

What seem to be the key points of an interpretable model?

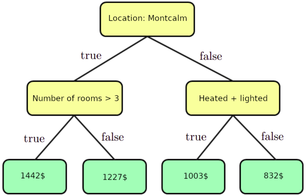
$$\begin{aligned}\hat{y} =\ & 70\$\ + \\ & 90¢ \times \text{number of square foot} + \\ & 58\$ \times \text{number of rooms}\end{aligned}$$



What seem to be the key points of an interpretable model?

- Additive model

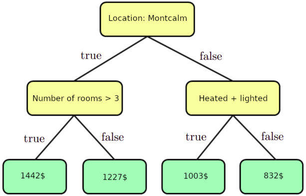
$$\hat{y} = 70\$ + 90¢ \times \text{number of square foot} + 58\$ \times \text{number of rooms}$$



What seem to be the key points of an interpretable model?

- Additive model
- Simple interactions between features

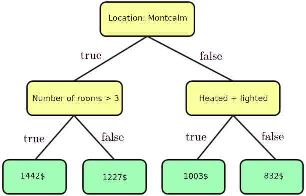
$$\hat{y} = 70\$ + 90¢ \times \text{number of square foot} + 58\$ \times \text{number of rooms}$$



What seem to be the key points of an interpretable model?

- Additive model
- Simple interactions between features
- Simple feature manipulations

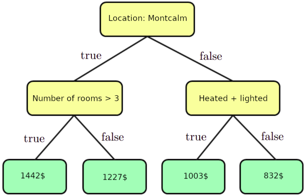
$$\hat{y} = 70\$ + 90¢ \times \text{number of square foot} + 58\$ \times \text{number of rooms}$$



What seem to be the key points of an interpretable model?

- Additive model
- Simple interactions between features
- Simple feature manipulations
- Small amount of features

$$\hat{y} = 70\$ + 90¢ \times \text{number of square foot} + 58\$ \times \text{number of rooms}$$



Neural networks

Neural networks

- Here : fully connected neural networks

Neural networks

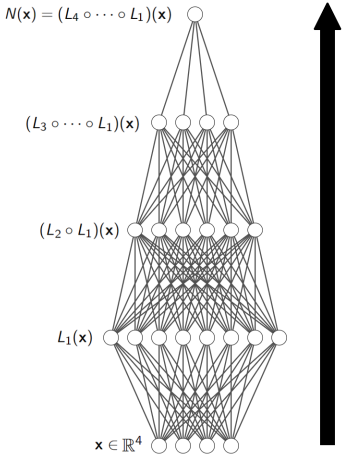
- Here : fully connected neural networks
- $B(\mathbf{x}) = (L_l \circ \cdots \circ L_1)(\mathbf{x}) = L_l(\ldots L_2(\ L_1(\mathbf{x})\ )\ldots)$

Neural networks

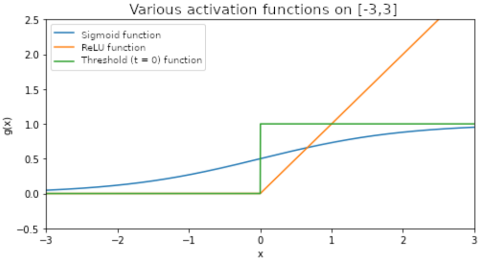
- Here : fully connected neural networks
- $B(\mathbf{x}) = (L_l \circ \cdots \circ L_1)(\mathbf{x}) = L_l(\ldots L_2(\ L_1(\mathbf{x})\ )\ldots)$
- $L_k(\mathbf{x}) = g_k(\mathbf{W}_k\mathbf{x})\ \forall k \in \{1, \ldots, l\}$

Neural networks

- Here : fully connected neural networks
- $B(\mathbf{x}) = (L_l \circ \cdots \circ L_1)(\mathbf{x}) = L_l(\ldots L_2(\ L_1(\mathbf{x})\ )\ldots)$
- $L_k(\mathbf{x}) = g_k(\mathbf{W}_k\mathbf{x})\ \forall k \in \{1, \ldots, l\}$

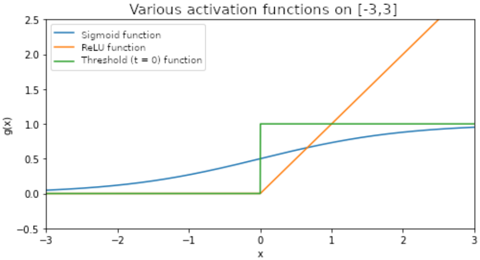


Why binary activated neural networks (BANNs)?



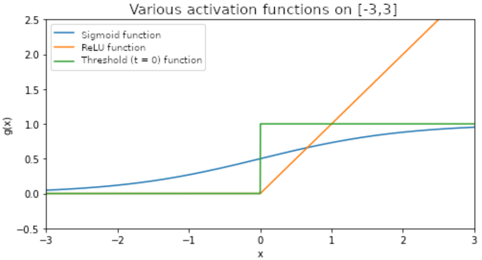
Why binary activated neural networks (BANNs)?

- Savings on computing time



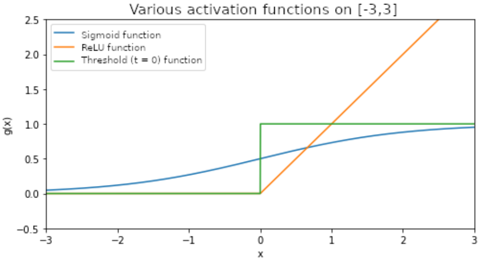
Why binary activated neural networks (BANNs)?

- Savings on computing time
- Predictors are less prone to overfitting [1-2]



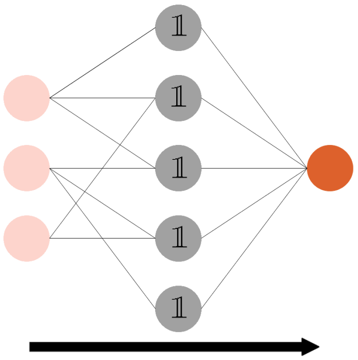
Why binary activated neural networks (BANNs)?

- Savings on computing time
- Predictors are less prone to overfitting [1-2]
- **Interpretability possibilities**



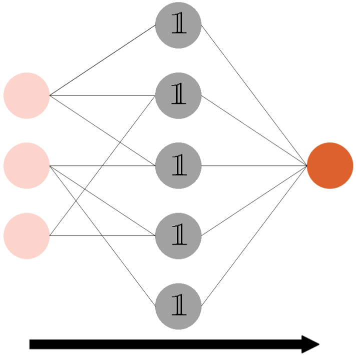
Simple neural networks

Simple neural networks



Simple neural networks

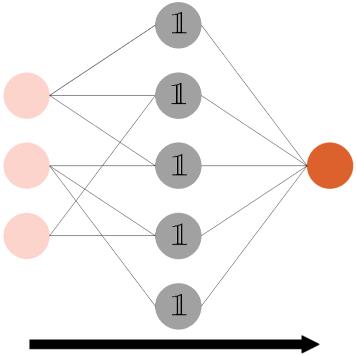
- Binary activations (threshold)



Simple neural networks

- Binary activations (threshold)

$$\mathbb{1}_{\{x\}} = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

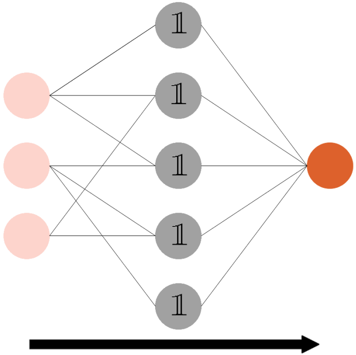


Simple neural networks

- Binary activations (threshold)

$$\mathbb{1}_{\{x\}} = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

- Shallow networks (1 hidden layer)

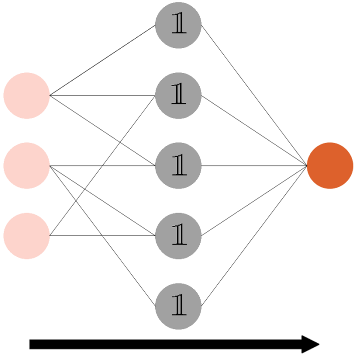


Simple neural networks

- Binary activations (threshold)

$$\mathbb{1}_{\{x\}} = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

- Shallow networks (1 hidden layer)
- Narrow networks

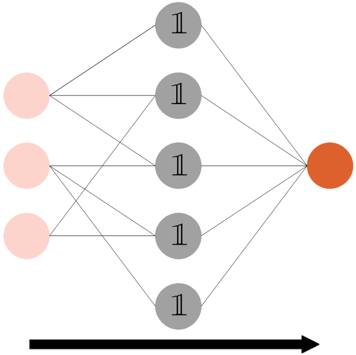


Simple neural networks

- Binary activations (threshold)

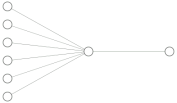
$$\mathbb{1}_{\{x\}} = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

- Shallow networks (1 hidden layer)
- Narrow networks
- Sparse (2)



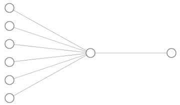
Building the networks: BGN

Characteristics



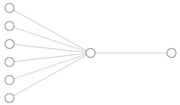
Characteristics

- For tackling regression tasks



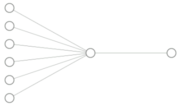
Characteristics

- For tackling regression tasks
- Greedy algorithm (inspired from Adaboost)



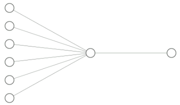
Characteristics

- For tackling regression tasks
- Greedy algorithm (inspired from Adaboost)
- Convergence guarantees on the train mean squared error



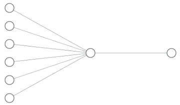
Characteristics

- For tackling regression tasks
- Greedy algorithm (inspired from Adaboost)
- Convergence guarantees on the train mean squared error
- No fixed architecture



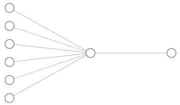
Characteristics

- For tackling regression tasks
- Greedy algorithm (inspired from Adaboost)
- Convergence guarantees on the train mean squared error
- No fixed architecture
- Uses a L1 norm regularization



Characteristics

- For tackling regression tasks
- Greedy algorithm (inspired from Adaboost)
- Convergence guarantees on the train mean squared error
- No fixed architecture
- Uses a L1 norm regularization
- No hyperparameter (learning rate, batch size, ...)



SHAP values

SHAP values

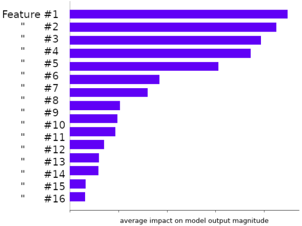
- Goal: quantifying the contribution (magnitude, impact) of each feature to a prediction

SHAP values

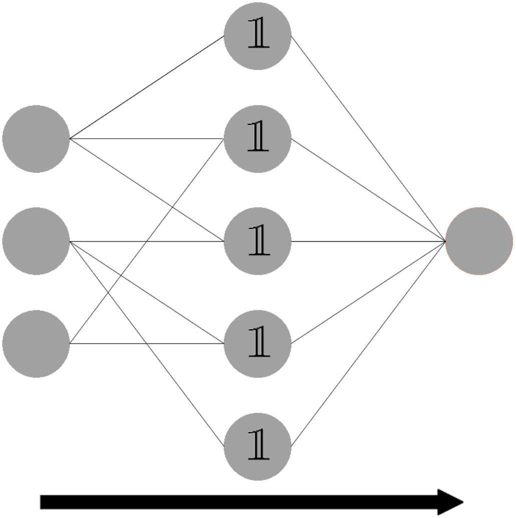
- Goal: quantifying the contribution (magnitude, impact) of each feature to a prediction
- At a prediction or a dataset level of aggregation

SHAP values

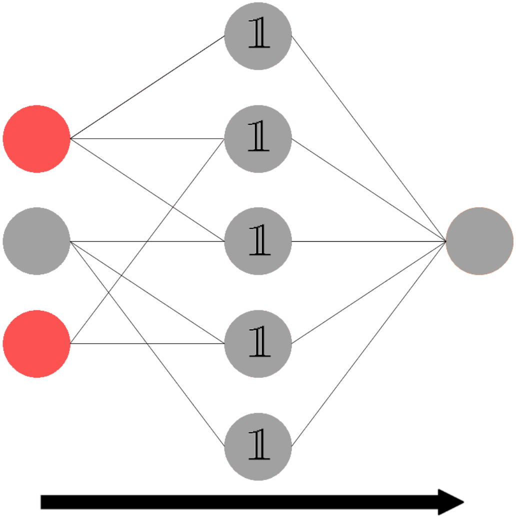
- Goal: quantifying the contribution (magnitude, impact) of each feature to a prediction
- At a prediction or a dataset level of aggregation



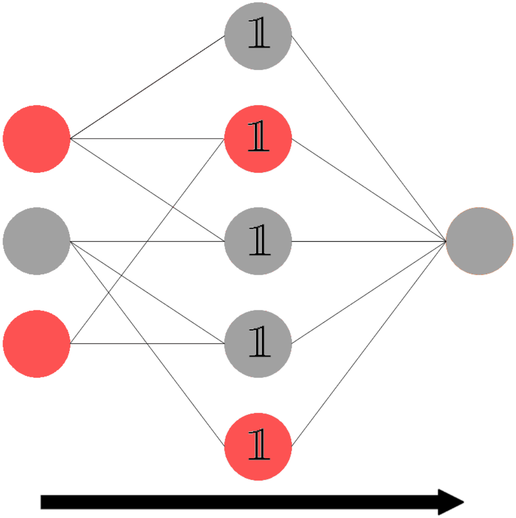
Explaining the networks: SHAP values



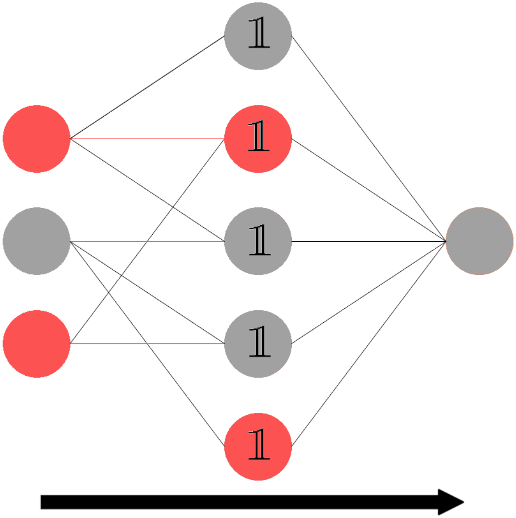
Explaining the networks: SHAP values



Explaining the networks: SHAP values



Explaining the networks: SHAP values



Explaining the networks: 1-BANN SHAP

Algorithm 1-BANN SHAP

Input : $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, $\mathbf{x} \in \mathbb{R}^d$, the features of dataset
B, a BANN; $\{\mathbf{W}_1, \ldots, \mathbf{W}_l\}$, its weights
$\mathbf{R} = \mathbf{0}_{d \times d \times |L_1|}$
$\mathbf{C} = \mathbf{0}_{1 \times d}$
For $g \in \{1, \ldots, |L_1|\}$:
    $\mathbf{a} = \mathbb{1}_{\{\mathbf{w}_g \neq 0\}}$
    $\mathbf{C} = \mathbf{C} \cup comb(\mathbf{a})$
For $i \in \{1, \ldots, d\}$ such that $(\exists j \mid c_{j,i} = 1)$:
    For $j \in \{1, \ldots, |\mathbf{C}|\}$ such that $c_{j,i} = 1$:
        For $\mathbf{x}, \mathbf{x}' \in S$:
            If $L_1\left(\mathbf{x}_{\mathbf{c}\setminus\{f\}} \cup \mathbf{x}'_{\overline{\mathbf{c}\setminus\{f\}}}\right) \neq L_1(\mathbf{x}_{\mathbf{c}} \cup \mathbf{x}'_{\overline{\mathbf{c}}})$:
                $\mathbf{r}_{i,|c_{j,i}|_1} = \mathbf{r}_{i,|c_{j,i}|_1} + \frac{\theta_{\mathbf{x},f}}{m} \odot \left|\sum_{k=1}^{d_l} \mathbf{w}_k\right|$,
with $\theta_{\mathbf{x},f} = \left|L_1\left(\mathbf{x}_{\mathbf{c}\setminus\{f\}} \cup \mathbf{x}'_{\overline{\mathbf{c}\setminus\{f\}}}\right) - L_1(\mathbf{x}_{\mathbf{c}} \cup \mathbf{x}'_{\overline{\mathbf{c}}})\right|$
Return R

Testing our *interpretable* approach

The task

Testing our *interpretable* approach

The task

- Task: predict the cost of a house (measured in 100k USD)

Testing our *interpretable* approach

The task

- Task: predict the cost of a house (measured in 100k USD)
- Retained features (out of eight):

Testing our *interpretable* approach

The task

- Task: predict the cost of a house (measured in 100k USD)
- Retained features (out of eight):
  - The median age of a house within a block (MedAge)

Testing our *interpretable* approach

The task

- Task: predict the cost of a house (measured in 100k USD)
- Retained features (out of eight):
    - The median age of a house within a block (MedAge)
    - The total number of bedrooms within a block (TotalBed)

Testing our *interpretable* approach

The task

- Task: predict the cost of a house (measured in 100k USD)
- Retained features (out of eight):
  - The median age of a house within a block (MedAge)
  - The total number of bedrooms within a block (TotalBed)
  - The median income for households within a block (measured in 1k USD) (MedInc)

Testing our *interpretable* approach

The task

- Task: predict the cost of a house (measured in 100k USD)
- Retained features (out of eight):
  - The median age of a house within a block (MedAge)
  - The total number of bedrooms within a block (TotalBed)
  - The median income for households within a block (measured in 1k USD) (MedInc)

$$
\begin{aligned}
&B(\text{MedInc, MedAge, TotalBed}) = \\
&1.00\ + \\
&1.27 \cdot \mathbb{1}_{\{0.08 \cdot \text{TotalBed} + \text{MedInc} > 65.46\}} + \\
&1.01 \cdot \mathbb{1}_{\{0.59 \cdot \text{TotalBed} + \text{MedInc} > 63.56\}} + \\
&0.60 \cdot \mathbb{1}_{\{\text{MedInc} > 28.20\}} + \\
&0.36 \cdot \mathbb{1}_{\{\text{TotalBed} > 622.0\}} + \\
&0.27 \cdot \mathbb{1}_{\{\text{MedAge} > 20.0\}}
\end{aligned}
$$

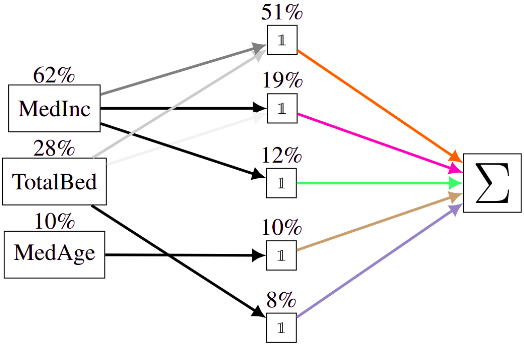
Testing our *interpretable* approach

The criterion

- Additive model
- Simple interactions between features
- Simple feature manipulations
- Small amount of features

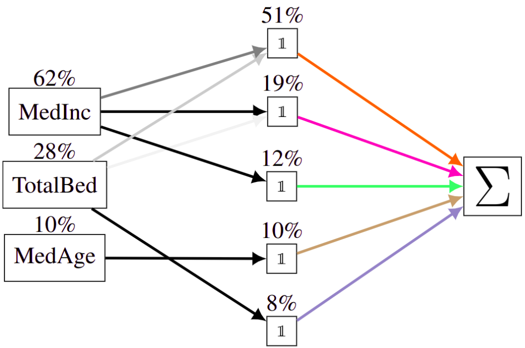
$$B(\text{MedInc, MedAge, TotalBed}) =$$
$$1.00 +$$
$$1.27 \cdot \mathbb{1}_{\{0.08 \cdot \text{TotalBed} + \text{MedInc} > 65.46\}} +$$
$$1.01 \cdot \mathbb{1}_{\{0.59 \cdot \text{TotalBed} + \text{MedInc} > 63.56\}} +$$
$$0.60 \cdot \mathbb{1}_{\{\text{MedInc} > 28.20\}} +$$
$$0.36 \cdot \mathbb{1}_{\{\text{TotalBed} > 622.0\}} +$$
$$0.27 \cdot \mathbb{1}_{\{\text{MedAge} > 20.0\}}$$

Testing our *interpretable* approach



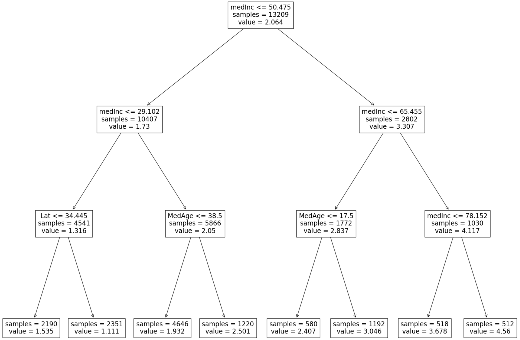
$$B(\text{MedInc, MedAge, TotalBed}) =$$
$$1.00 +$$
$$1.27 \cdot \mathbb{1}_{\{0.08 \cdot \text{TotalBed} + \text{MedInc} > 65.46\}} +$$
$$1.01 \cdot \mathbb{1}_{\{0.59 \cdot \text{TotalBed} + \text{MedInc} > 63.56\}} +$$
$$0.60 \cdot \mathbb{1}_{\{\text{MedInc} > 28.20\}} +$$
$$0.36 \cdot \mathbb{1}_{\{\text{TotalBed} > 622.0\}} +$$
$$0.27 \cdot \mathbb{1}_{\{\text{MedAge} > 20.0\}}$$

Testing our *interpretable* approach



$$B(\text{MedInc, MedAge, TotalBed}) =$$
$$1.00 +$$
$$1.01 \cdot \mathbb{1}_{\{0.59 \cdot \text{TotalBed} + \text{MedInc} > 63.56\}} +$$
$$1.27 \cdot \mathbb{1}_{\{0.08 \cdot \text{TotalBed} + \text{MedInc} > 65.46\}} +$$
$$0.60 \cdot \mathbb{1}_{\{\text{MedInc} > 28.20\}} +$$
$$0.27 \cdot \mathbb{1}_{\{\text{MedAge} > 20.0\}} +$$
$$0.36 \cdot \mathbb{1}_{\{\text{TotalBed} > 622.0\}}$$

Testing our *interpretable* approach



$$B(\text{MedInc, MedAge, TotalBed}) =$$
$$1.00 +$$
$$1.01 \cdot \mathbb{1}_{\{0.59 \cdot \text{TotalBed} + \text{MedInc} > 63.56\}} +$$
$$1.27 \cdot \mathbb{1}_{\{0.08 \cdot \text{TotalBed} + \text{MedInc} > 65.46\}} +$$
$$0.60 \cdot \mathbb{1}_{\{\text{MedInc} > 28.20\}} +$$
$$0.27 \cdot \mathbb{1}_{\{\text{MedAge} > 20.0\}} +$$
$$0.36 \cdot \mathbb{1}_{\{\text{TotalBed} > 622.0\}}$$

Conclusion

Artificial neural networks can be **interpretable** predictors...

Conclusion

Artificial neural networks can be **interpretable** predictors...
... When trained with such a goal in mind!

References

[1] J. Lin, C. Gan, and S. Han, "Defensive quantization: When efficiency meets robustness," CoRR, vol. abs/1904.08444, 2019.

[2] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1," 2016.

Thank you for your attention :)