Investigating the Learning Dynamics of CNN using MINE

Nikita Tokovenko, Lars Rigter, Maximilian Ilse

University of Amsterdam

UNIVERSITEIT VAN AMSTERDAM

Theory

Mutual information captures non-linear statistical dependencies between variables and thus can act as a measure of true dependence (Kinney, Atwal, 2014)

$$I(X,Z) = D_{KL}(\mathbb{P}_{X,Z} \| \mathbb{P}_X \otimes \mathbb{P}_Z)$$

Where the KL divergence can be written as its dual representation:

$$D_{KL}(\mathbb{P} \| \mathbb{Q}) = \sup_{T:\Omega \mapsto \mathbb{R}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T])$$

MINE is an estimator for mutual information and is defined as:

$$T_\theta : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$$

Where T is a neural network parameterized with theta in order to find the lower bound statistical network:

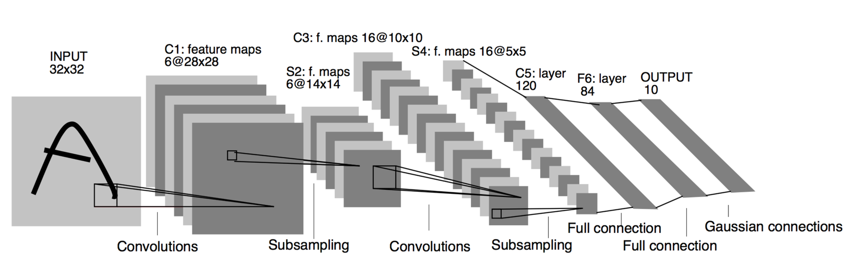
$$I(X,Z) \geq I_\theta(X,Z),$$

Where $I_\theta(X,Z)$ is the mutual information lower bound, defined as:

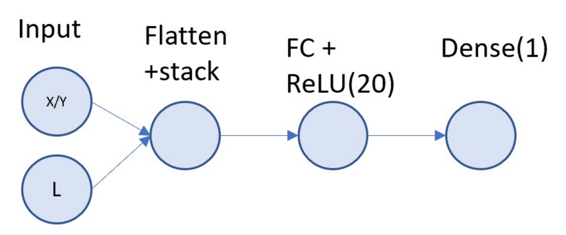
$$I_\theta(X,Z) = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{XZ}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}[e^{T_\theta}]).$$

Architecture

Target Network – LeNet-5



MINE



Training procedure

MINE

Algorithm 1 MINE

$\theta \leftarrow$ initialize network parameters
repeat
Draw b minibatch samples from the joint distribution: $(\boldsymbol{x}^{(1)}, \boldsymbol{z}^{(1)}), \ldots, (\boldsymbol{x}^{(b)}, \boldsymbol{z}^{(b)}) \sim \mathbb{P}_{XZ}$
Draw n samples from the Z marginal distribution: $\overline{\boldsymbol{z}}^{(1)}, \ldots, \overline{\boldsymbol{z}}^{(b)} \sim \mathbb{P}_Z$
Evaluate the lower-bound: $\nu(\theta) \leftarrow$
$\leftarrow \frac{1}{b}\sum_{i=1}^{b} T_\theta(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) - \log(\frac{1}{b}\sum_{i=1}^{b} e^{T_\theta(\boldsymbol{x}^{(i)}, \overline{\boldsymbol{z}}^{(i)})})$
Evaluate gradients:
$\widehat{G}(\theta) \leftarrow \nabla_\theta \nu(\theta)$
Update the statistics network parameters:
$\theta \leftarrow \theta + \widehat{G}(\theta)$
until convergence

MINE + LeNet-5

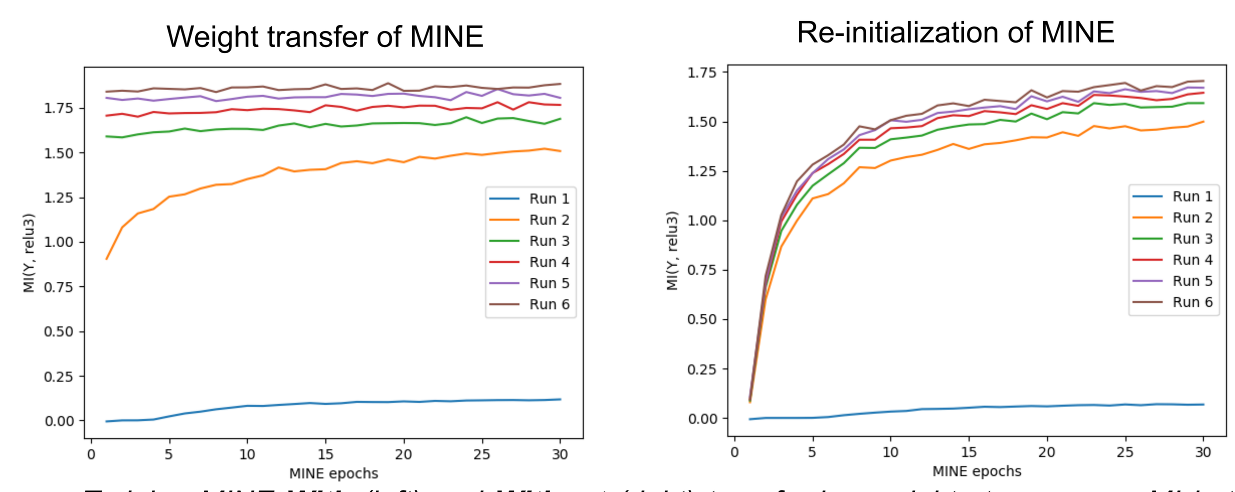
For every run of MINE**:**

- Train LeNet-5 for 4 epochs with low learning rate
- Train MINE until convergence between input (X) and target label (Y) and Conv1 (.), Conv2 (.), ReLU3 (.) and Softmax (.)

end

Transferring weights of MINE

- Every 4th epoch of training LeNet-5, MINE will be initialized with weights from previous epoch



Training MINE ***With*** *(left) and* ***Without*** *(right) transferring weights to measure MI between target labels and first FC layer for different epochs of LeNet-5 training*

Conclusion

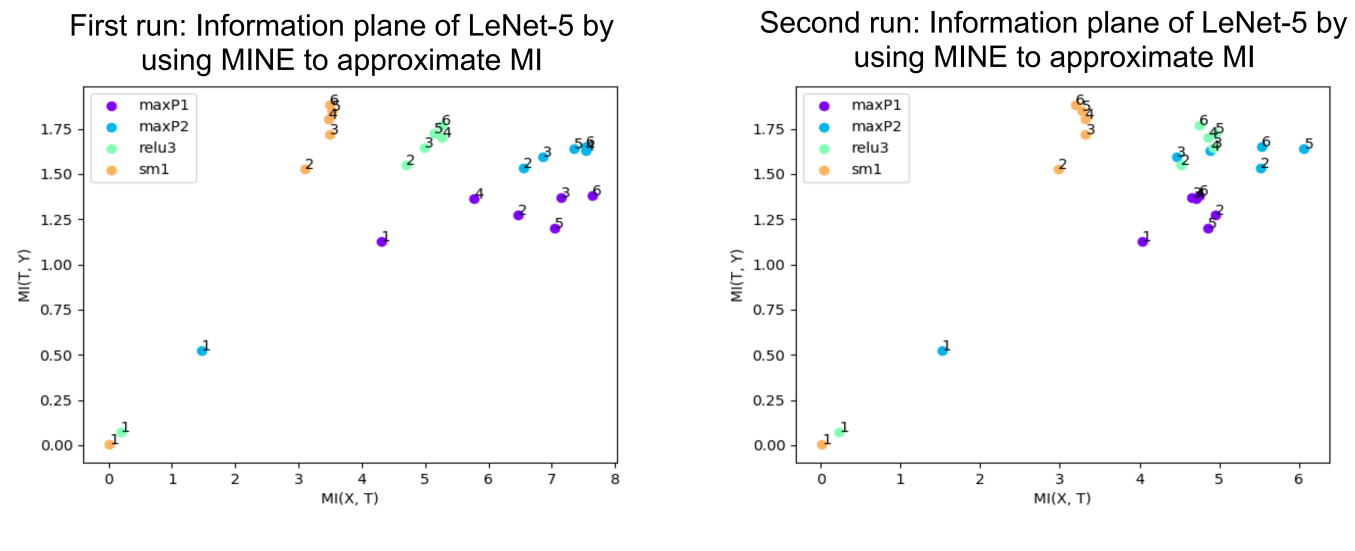
- Transferring weights speeds up the convergence of MINE

Information Plane

The information plane consist of MI between input and layer output I(X, L) on one axis and MI between target and layer output I(Y, L) on the other axis. The information path in information plane satisfies:

$$H(X) \geq I(X, L_1) \geq I(X, L_2) \geq \ldots \geq I(X, L_k) \geq I(X, \widehat{Y})$$

$$I(X, Y) \geq I(L_1, Y) \geq I(L_2, Y) \geq \ldots \geq I(L_k, Y) \geq I(\widehat{Y}, Y)$$



Training LeNet-5 and estimating MI between input and layer output (X, L) and target and layer output (Y, L) for every 4 epochs

Conclusion

- MINE is sensitive for different neural network architectures
- After fine tuning MINE is consistent for 1D shaped date and less consistent for 2D shaped data

Bibliography

1. *Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Devon Hjelm, and Aaron Courville. Mutual information neural estimation. In International Conference on Machine Learning (ICML), pp. 530–539, 2018.*
2. *Kinney, J. B. and Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. Proceedings of the National Academy of Sciences, 111(9):3354–3359, 2014.*