Distilling Spikes: Knowledge Distillation in Spiking Neural Networks

Ravi Kumar Kushawaha, Saurabh Kumar, Biplab Banerjee, Rajbabu Velmurugan

Indian Institute of Technology Bombay, India

Abstract

- Spiking Neural Networks (SNN) are energy-efficient computing architectures that exchange spikes for processing information, unlike classical Artificial Neural Networks (ANN). SNNs are better suited for real-life deployments and benefit from deeper architectures to obtain improved performance.
- The memory, compute and power requirements of SNNs also increase with model size, and model compression becomes a necessity. Knowledge distillation is a model compression technique that enables transferring the learning of a large machine learning model to a smaller model with minimal loss in performance.

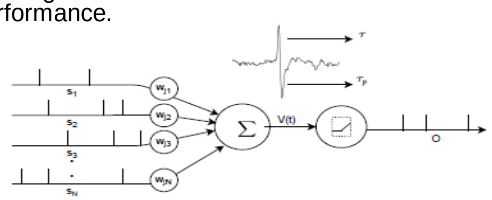
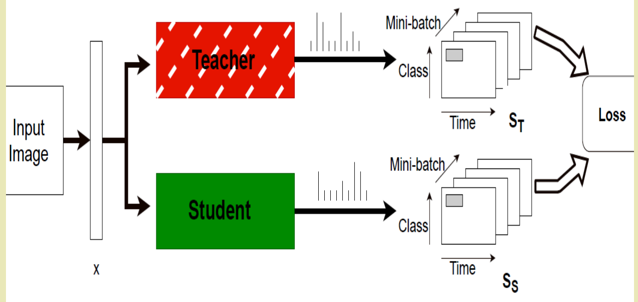


Fig: Working of a Spiking Neuron

Our Contribution

- We propose techniques for knowledge distillation in spiking neural networks for the task of image classification. We present ways to distill spikes from a larger SNN, also called the teacher network, to a smaller one, also called the student network
- We demonstrate the effectiveness of the proposed method with detailed experiments on three standard datasets while proposing novel distillation methodologies and loss functions
- We also present a multi-stage knowledge distillation technique for SNNs using an intermediate network to obtain higher performance from the student network
- Our approach is expected to open up new avenues for deploying high performing large SNN models on resource-constrained hardware platforms

Training Methodology



1. We first train a teacher SNN which is then used in Knowledge Distillation for a student network.
2. Given an input image, the weights of teacher SNN are frozen while the student SNN is trained.
3. The KD process involves training of this two-stream setup with the proposed loss functions on the post-synaptic spike patterns of the Teacher and Student SNN models

Loss Function

- The 3-D tensor (time x classes x mini-batch size) is referred as spiking activation tensor (SAT)
- Losses are calculated by comparing the SATs of both teacher and student model
- L1, L2, KL loss computed on entire tensors and sliding window losses for L1, L2

$$L_{sLm} = \sum_{k \epsilon b} \sum_{j \epsilon c} \sum_{i \epsilon t} ||\mathcal{S}_T[i:i+\Delta;j;k] - \mathcal{S}_S[i:i+\Delta;j;k]||_m$$

Results

TABLE I: Baseline classification performances of individual networks when trained separately on the three datasets.

Dataset	MNIST	F-MNIST	CIFAR10
Teacher	98.35	89.72	45.43
TA	98.17	89.4	45.98
Student	98.00	88.64	42.9

TABLE II: Performance comparison of Student SNNs with knowledge distilled from the Teacher model using individual components of the proposed loss function.

Dataset	MNIST	F-MNIST	CIFAR10
Teacher	98.35	89.72	45.43
Full L1 ($\mathcal{L}_{L1}$)	96.20	86.99	37.90
Full L2 ($\mathcal{L}_{L2}$)	96.80	87.50	38.70
Full KL ($\mathcal{L}_{KL}$)	97.36	88.15	39.21
Sliding L1 ($\mathcal{L}_{sL1}$)	96.09	87.28	38.31
Sliding L2 ($\mathcal{L}_{sL2}$)	96.29	87.08	38.89
Proposed	**97.46**	**88.30**	**41.28**

TABLE III: Classification performance when using an intermediate TA network for KD from teacher to student.

Dataset	MNIST	F-MNIST	CIFAR10
Teacher	98.35	89.72	45.43
T → TA	98.36	89.82	45.33
T → S	97.46	88.30	41.28
T → TA → S	**97.56**	**88.74**	**42.38**

Conclusion

- We demonstrated distilling knowledge from a large SNN model trained for image classification
- Multistep distillation strategy offers further improvement in performance by using an intermediate TA network

email id– rkkush2397@gmail.com LinkedIn: https://www.linkedin.com/in/ravi-kumar-kushawaha-224950121/