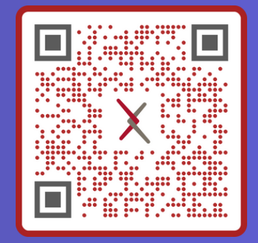
Center of Optical Neural Technologies

SpikeFit: Towards Efficient Deployment of Spiking Networks on Neuromorphic Hardware

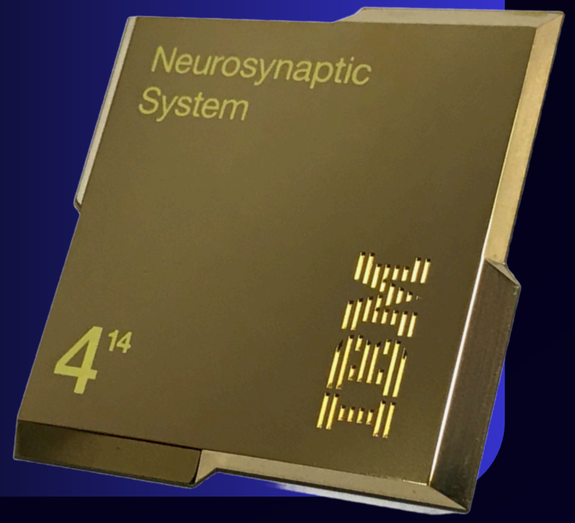


Ivan Kartashov, Mariia Pushkareva, Iakov Karandashev

iakartashov@edu.hse.ru

Center of Optical Neural Technologies

Spiking Neural Networks Compression



Movitation of SpikeFit

Several neuromorphic systems, designed for superior energy efficiency, restrict each synapse to a very small number of available states. IBM TrueNorth, Alt-AI 1 highlighted use of just four integer values to represent synaptic weights. This is a fundamental constraint. Training SNNs to perform well under such extreme constraints is a major challenge. Therefore, further research is required on SNN compression methods that allow full deployment on the hardware.

Center of Optical Neural Technologies

State of SNN Compression Research

Related Works

Model compression is essential for deploying deep neural networks on resource-constrained devices. Quantization and pruning remain key techniques reducing the precision of model weights, its size, and energy use.

QP-SNN architecture (Wei et al. ICLR 2025) achieves state-of-the-art efficiency by integrating hardware-friendly SVS-based structured pruning technique with quantization and weight rescaling strategy applied beforehand. However, QP-SNN is not applicable for neuromorphic processors limiting synaptic discrete weights values.

SpikeFit achieves state-of-the-art results by integrating novel quantization, clusterization, pruning schemes allowing full deployment.

What is SpikeFit?

Key Components of the framework

1. Clusterization-Aware Training (CAT) method finding optimal sets of discrete values to represent weights in low numerical precision where the M unique discrete weight values for each layer are learned end-to-end, allowing the model to discover its own optimal resource-friendly weight representation.

2. Fisher Spike Contribution (FSC) Pruning method approximating the diagonal Fisher Information of channel-wise gates in spiking network.

Center of Optical Neural Technologies

Clusterization-Aware Training Pipeline

For each layer using Clusterization Aware Training (CAT):

$$W_{\text{latent}} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_k},$$

where W_{latent} are full-precision latent weights updated via backpropagation.

$$C = \{c_m \in \mathbb{R} \mid m = 1, 2, \ldots, M\},$$

where C is a learnable codebook defining the discrete quantized values.

Each layer therefore maintains
(W_{latent}, C) as its learnable parameter sets.

Optimized Clustering Codebooks for each layer

Center of Optical Neural Technologies

Clusterization-Aware Training Pipeline

The core idea is to map each latent weight to the closest value in the codebook during the forward pass, while ensuring gradients can flow back to both the latent weights and the codebook during the backward pass.

Each latent weight w^{latent} is quantized to its nearest neighbor in the codebook C. The resulting weight, $w^{\text{quantized}}$, is used for the layer's operation (e.g., convolution).

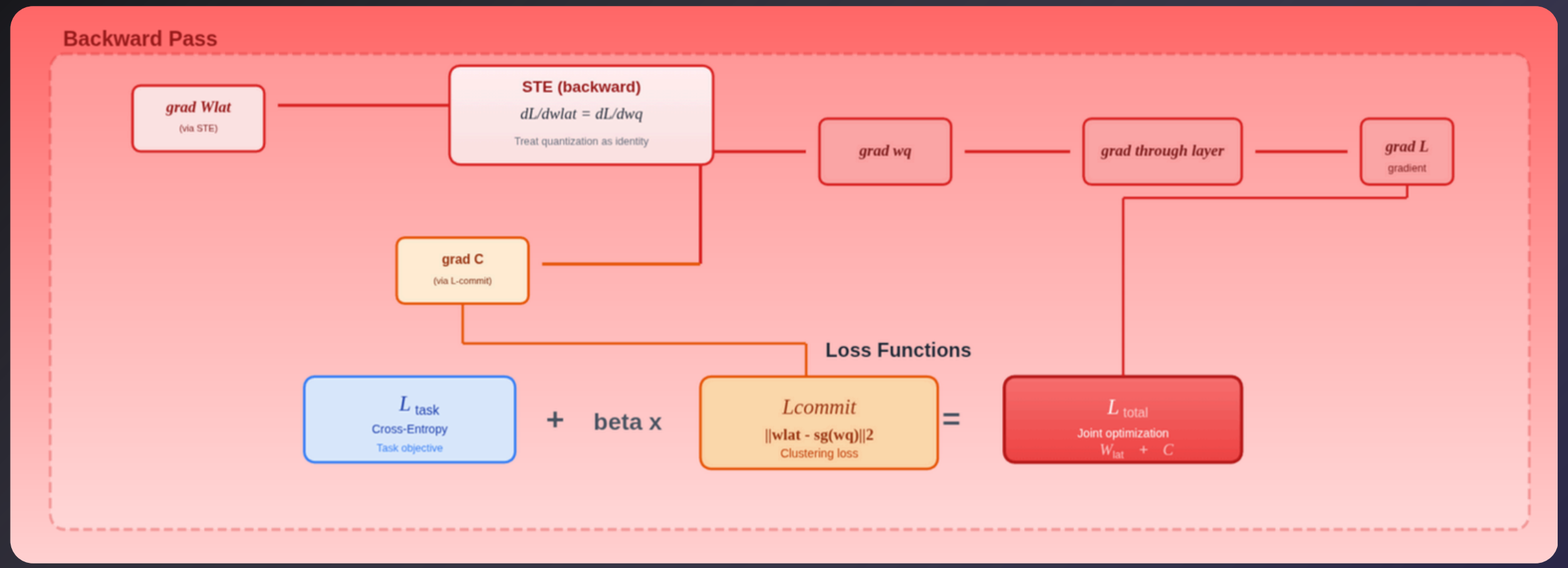
$$k^* = \arg \min_{k \in \{1,\ldots,M\}} \|w^{\text{latent}} - c_k\|^2 w^{\text{quantized}} = c_{k^*}$$

Optimized Clustering Codebooks for each layer

Center of Optical Neural Technologies

Clusterization-Aware Training Pipeline

Training is guided by a composite loss function.
The first component is the standard task loss, $\mathcal{L}_{\text{task}}$ (e.g., Cross-Entropy).
The second component is a **commitment loss**, $\mathcal{L}_{\text{commit}}$,
which enables the "smart clustering" behavior.
It penalizes the distance between the latent weights
and their corresponding chosen codebook values:

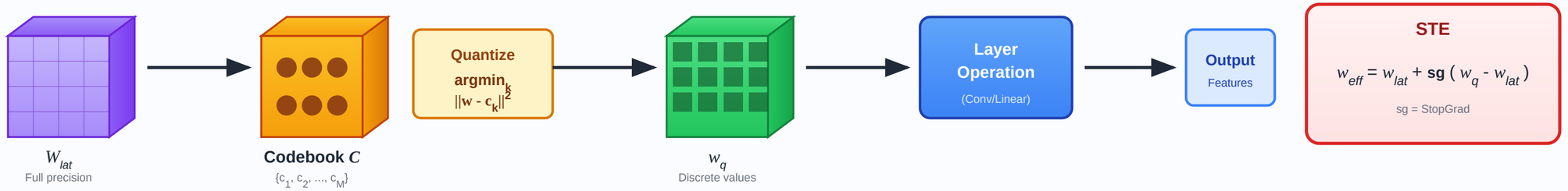


Optimized Clustering Codebooks for each layer

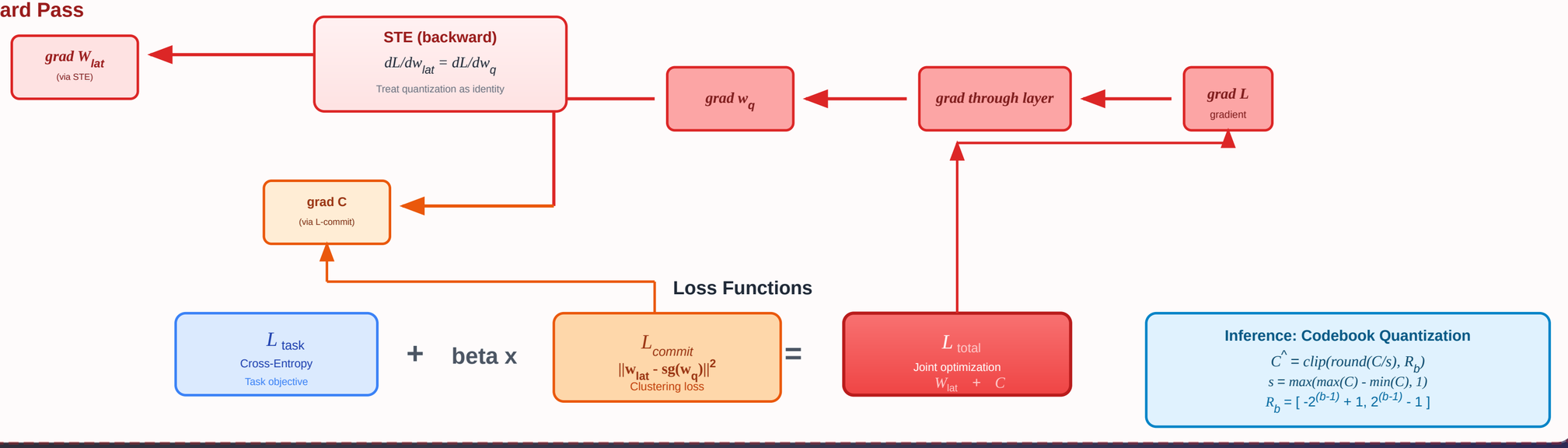
Center of Optical Neural Technologies

Clusterization-Aware Training (CAT)

Forward Pass



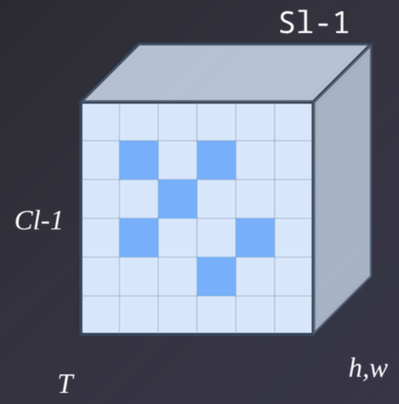
Backward Pass



Center of Optical Neural Technologies

We define the Fisher Spike Contribution score for channel c as

$$S_c^{\mathrm{FSC}} = \frac{1}{N} \sum_{b=1}^{N} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(\delta_{b,t,c,h,w}\right)^2 \left(y_{b,t,c,h,w}\right)^2.$$



Intuitively, S_c^{FSC} is large only when a channel spikes (large $|y|$) exactly where/when the loss is sensitive (large $|\delta|$).

Fisher Spike Contribution Pruning

Spiking Activity based & Loss Aligned Pruning criterion

Center of Optical Neural Technologies

Hessian Approximation

Fisher Information & Hessian

For modern neural networks, computing the full Hessian is not feasible, since the stoted matrices scale quadratically with the number of parameters.

The most popular methods include the generalized **GaussNewton (GGN) matrix** and the **Fisher information matrix**, as well as block-diagonal factorizations such as the **Kronecker-factored approximate curvature (K-FAC).**

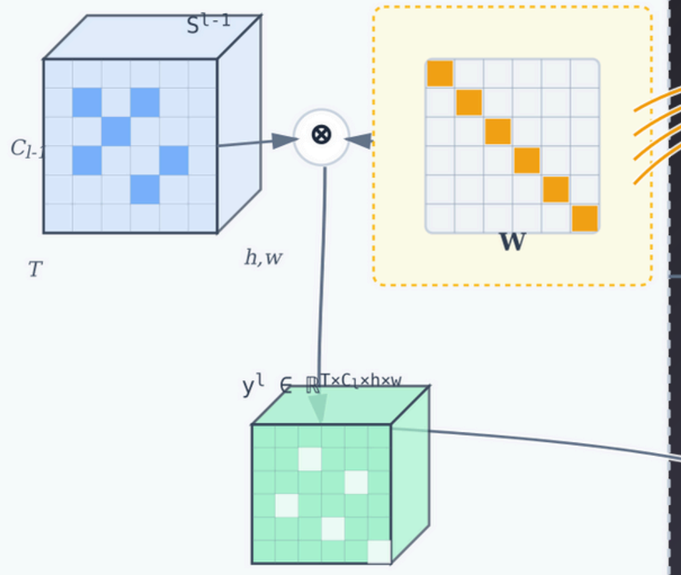
Fisher is the same as GGN when utilizing distributions from the exponential family (Martens et. al), making them both well-justified approximations of the Hessian.

Under standard conditions for negative log-likelihood losses, the expected *Hessian equals the Fisher information matrix (Amari et. al).*

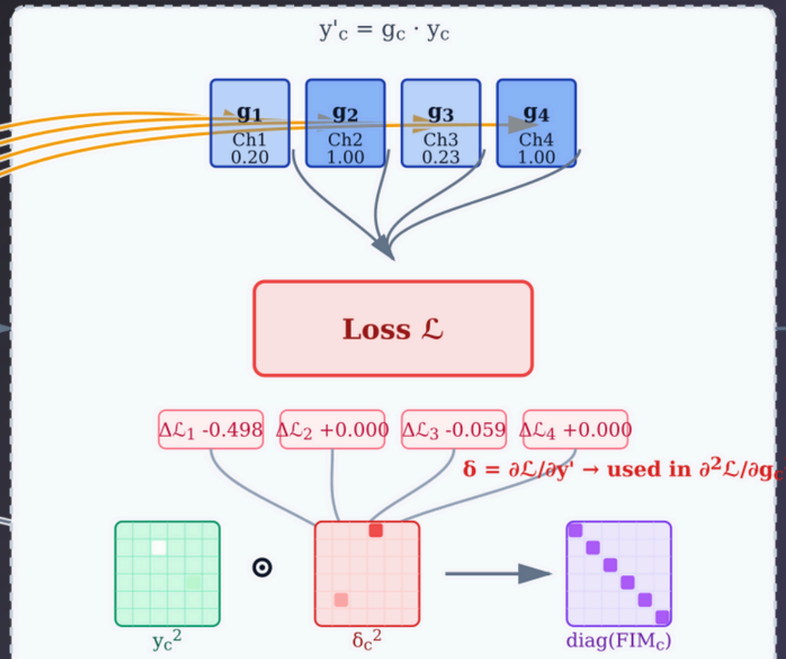
SpikeFit uses Fischer Spike Contribution pruning scheme.
It calculates the curvature of a loss function (2nd derivitive) to estimate parameters contribution.

Center of Optical Neural Technologies

Input Spikes & Activations

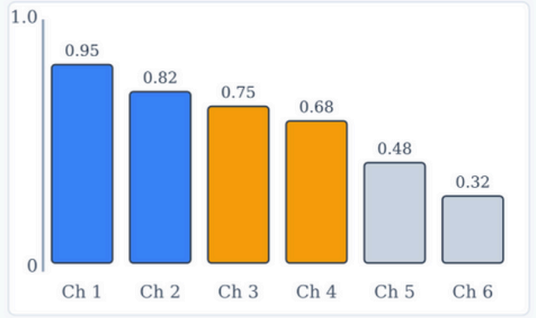


Channel Gates (per-channel) & Gradients



Fisher Spike Contribution

$$S^{FSC}_c \approx \partial^2\mathcal{L}/\partial g_c^2 = \sum_{b,t,h,w} |\delta|^2 \cdot |y|^2$$



Channel Importance Scores

→ Prune channels with lowest S^{FSC}

Integrates temporal spike activity $|y|^2$ with task-sensitive gradients $|\delta|^2$ across time steps

Center of Optical Neural Technologies

DeplyRatio Metric

$$\mathrm{DR}_x = \frac{\mathrm{Perf}_x}{L\,E\,S}, \quad x \in \{\mathrm{acc}, \mathrm{f1}\},$$

$$\mathrm{Perf}_{\mathrm{acc}} = \frac{\mathrm{Accuracy}[\%]}{100}, \; \mathrm{Perf}_{\mathrm{f1}} = \frac{\mathrm{F1}[\%]}{100}.$$

where L is inference latency (s/input),
E energy per inference (mJ),
and S model size (MB).

SpikeFit provides state-of-the-art deployment

Table 1: Comparison of pruning and compression methods across networks and strategies. Bold blue indicates our proposed FSC method.

Method	Network	Pruning	Size (MB)	Energy (mJ)	Acc (%)	Prec (%)	Rec (%)	F1 (%)
QP-SNN	VGG-16	SVS (36)	14.723	**0.03**	86.656±0.116	86.842±0.107	86.656±0.116	86.628±0.118
		SCA (38)	14.723	**0.03**	83.275±0.37	84.927±0.27	83.278±0.37	83.373±0.39
		FSC (ours)	14.725	**0.03**	**87.546±0.059**	**87.621±0.073**	**87.546±0.059**	**87.553±0.065**

Table 2: Performance comparison of compression methods. Accuracy is reported as **mean ± std** over five runs. **DR (DeployRatio)** indicates the deployment cost-performance ratio. Bold blue indicates our proposed **SpikeFit** results.

Dataset	Network	Method	Precision	Size (MB)	Energy (mJ)	Timestep	Accuracy (%)	F1 (%)	DR (%)
CIFAR-10	VGG-16	Clustered Baseline	8	14.723	**0.004**	4	77.57±0.49	77.54±0.44	68.649
		Ternary Baseline	8	14.723	0.014	4	76.244±0.12	76.51±0.04	22.051
		QP-SNN	8	14.723	0.032	4	86.58±0.32	86.91±0.32	10.883
		SpikeFit (ours)	8	14.725	**0.004**	4	**89.14±0.82**	**89.99±0.84**	**89.640**

Results

Center of Optical Neural Technologies

Ablation Studies

Table 3: Ablation study on VGG-16 SNN architecture on the effectiveness of proposed CAT method.

Dataset	Network	Method	Precision	Size (MB)	Timestep	Accuracy (%)	F1 (%)
CIFAR-10	VGG-16	Full-precision	32	58.91	4	**90.97**	**90.92**
		ReScaW (36)	8	14.72	4	89.14	89.73
		QP-SNN (36)	8	14.72	4	86.94	86.92
		CAT (M=2)	8	14.72	4	87.56	85.24
		CAT (M=4)	8	14.72	4	**90.00**	**89.99**

To further prove the effectiveness of SpikeFit training methodology, we conduct extensive ablation studies. We perform thorough ablation experiments to validate the effectiveness of the proposed CAT weight clusterization strategy.

Center of Optical Neural Technologies

HSE University,
Faculty of Biology and Biotechnology

Thank you!

Research Paper

Ivan Kartashov, Mariia Pushkareva,
Iakov Karandashev

-

Corresponding author's email:
iakartashov@edu.hse.ru