Mathematical Problems of Computer Science 64, 37–46, 2025.
doi:10.51408/1963-0138

UDC 004.855.5

A PDE-Based Convolutional Neural Network with Variational Information Bottleneck: Experimental Evaluation and Generalization Analysis

Gor A. Gharagyozyan

Institute for Informatics and Automation Problems of NAS RA, Yerevan, Armenia
e-mail: gor.gharagyozyan@edu.isec.am

Abstract

We present a hybrid convolutional architecture that combines trainable PDE-based preprocessing with a Variational Information Bottleneck (VIB) to improve generalization in image classification. The PDE stage applies a small number of discretized Laplacian steps with learnable step size and depthwise coupling, injecting physics-inspired inductive bias into early feature maps. A tensor-wise VIB module then parameterizes a Gaussian latent $(\mu, \log \sigma^2)$ via 1×1 convolutions and enforces information compression through a KL penalty to a unit prior, encouraging retention of task-relevant features while discarding nuisance variability. The compressed representation feeds a ResNet-18 backbone adapted for CIFAR-10 inputs. On CIFAR-10, systematic variation of the VIB weight β shows that moderate compression yields improved test performance and training stability relative to both a baseline CNN and a PDE-only variant. Qualitative analysis indicates smoother activations and reduced sensitivity to input noise, consistent with the information-theoretic objective. The results suggest that PDE priors and variational compression act complementarily, offering a principled path to robust and generalizable convolutional models.

Keywords: Information bottleneck, Partial differential equations, Deep learning, Convolutional neural networks, Generalization.
Article info: Received 15 October 2025; sent for review 16 October April 2025; accepted 13 November 2025.
Acknowledgements: The author expresses sincere gratitude to Prof. Mariam Haroutunian for her steadfast guidance, insightful discussions, and generous support throughout this work. Her rigorous feedback and encouragement were instrumental in shaping the research questions, refining the methodology, and bringing this study to completion.

1. Introduction

Convolutional Neural Networks (CNNs) have become the foundation of modern computer vision, demonstrating outstanding performance across various image classification tasks. Ar-

chitectures such as ResNet [1] have shown that deep hierarchical representations can capture complex visual patterns; however, their ability to generalize remains sensitive to data quality, overparameterization, and the presence of irrelevant features. Improving generalization thus requires mechanisms that not only increase model capacity but also regulate the information flow within the network.

Recent studies have explored the incorporation of domain knowledge into CNNs to embed structural priors and reduce reliance on purely data-driven learning. In particular, Partial Differential Equation (PDE)-based layers [2], derived from the discretization of physical processes such as diffusion and wave propagation, have been shown to enhance low-level representations by enforcing spatial smoothness and continuity. These physics-inspired kernels act as a regularizing bias, improving robustness without adding significant computational cost. Yet, such deterministic transformations may also retain redundant information, which can propagate noise through deeper layers.

To address this limitation, this article builds upon a previous study presented at the Conference on Computer Science and Information Technologies (CSIT 2025) [3], where the integration of the Variational Information Bottleneck (VIB) module [4, 5] into the PDE-based CNN framework was first introduced conceptually. In the present work, the approach is evaluated through quantitative experiments, providing empirical evidence for the effectiveness of the PDE-VIB combination. The VIB principle seeks a stochastic latent representation that retains only information relevant to predicting the target while discarding task-irrelevant details. By combining PDE-based structural priors with information-theoretic compression, the proposed PDE-VIB-CNN achieves a balance between inductive bias and adaptive regularization. The resulting model learns compact and task-focused feature maps, leading to improved stability, robustness, and generalization on challenging datasets such as CIFAR-10.

2. Theoretical Background

This section provides a brief overview of the theoretical components underlying the proposed model, the PDE-based convolutional layers, and the Variational Information Bottleneck (VIB) framework. A detailed formulation and motivation can be found in [2] and [3].

PDE-Based Convolutional Layers

The PDE-based layer incorporates physically inspired priors into early feature extraction. The approach relies on the discretization of parabolic or hyperbolic partial differential equations, such as the two-dimensional diffusion (heat) equation:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}. \tag{1}$$

Using finite differences [6], the update rule can be expressed as:

$$u_{i,j}^{t+1} = u_{i,j}^{t} + \phi\, P(u^t), \tag{2}$$

where P denotes a convolution operator equivalent to the Laplacian kernel and ϕ is a learnable or fixed scaling parameter. These layers act as structural filters, enforcing smoothness and spatial continuity while reducing sensitivity to high-frequency noise [2].

VIB

The VIB framework [4, 5] formulates learning as an optimization of the mutual information trade-off between input compression and predictive relevance. The objective is defined as:

$$\mathcal{L}_{\mathrm{VIB}} = \mathbb{E}_{\mathrm{p(x,y)}}\Big[\mathbb{E}_{\mathrm{q(t|x)}}[-\log \mathrm{p(y \mid t)}]\Big] + \beta\, \mathrm{D_{KL}}(\mathrm{q(t \mid x)} \,\|\, \mathrm{p(t)}), \tag{3}$$

where q(t | x) is a Gaussian encoder producing the latent representation t, p(y | t) is the decoder, and β controls the compression–prediction trade-off. This formulation encourages the representation to retain only task-relevant information while suppressing redundancy.

The *PDE-VIB-CNN* model evaluated in this study extends the theoretical foundation proposed in [3], validating it experimentally on CIFAR-10 [7].

3. Experimental Setup

Dataset

All experiments are conducted on the CIFAR-10 dataset (60,000 color images, 32×32, 10 classes; 50k train, 10k test). Inputs are normalized per channel. We apply standard data augmentation: random horizontal flipping and random cropping with 4-pixel padding.

Architecture

The evaluated model, *PDE-VIB-CNN*, consists of three sequential stages.

(i) PDE stage. We prepend a stack of S PDE-based convolutional layers, each corresponding to one explicit Euler update of a discretized Laplacian step. For an input feature map u^t, a single PDE layer computes

$$u^{t+1} = u^t + \lambda\,(P * u^t),$$

where P is a fixed 3×3 Laplacian stencil and λ is a **learnable per-channel diffusion coefficient**. Thus, the PDE block performs S successive PDE updates (we use $S = 3$ in all experiments unless otherwise stated), injecting physics-inspired priors and encouraging smooth, spatially coherent feature representations [2]. No free-form convolution kernels are learned in this stage; the only learnable parameters are the diffusion coefficients $\{\lambda_c\}$ and batch-normalization parameters.

(ii) VIB module. After the PDE stage, we apply a variational information bottleneck module [4, 5]. The PDE output $f(x)$ is first compressed through a 1×1 bottleneck ($C \to C_b$ channels). Two parallel 1×1 convolutions then produce the parameters of a Gaussian latent distribution:

$$\mu(x) = W_\mu * f(x) + b_\mu, \qquad \log \sigma^2(x) = W_\sigma * f(x) + b_\sigma.$$

A latent tensor is sampled via the reparameterization trick,

$$t = \mu(x) + \sigma(x) \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(\imath, \mathcal{I}),$$

ensuring differentiability during training. This module enforces information compression and reduces overfitting by discouraging the encoding of spurious, high-frequency details.

(iii) CNN backbone. The sampled latent representation t is passed to a ResNet-18 backbone [1] (with the initial stem modified for CIFAR-10), followed by a linear classifier. This journal version extends the conceptual formulation introduced in [3] by providing detailed implementation, quantitative evaluation, and calibration analysis.

Training Protocol

Models are implemented in PyTorch [8] and trained end-to-end using stochastic gradient descent with momentum (SGD with momentum). Batch normalization is used in convolutional blocks, and dropout is employed in deeper layers to reduce overfitting. The VIB trade-off coefficient β is tuned empirically to balance compression and accuracy.

For fairness, all models are trained under **identical optimization and augmentation settings**, including the cosine-decay learning-rate schedule, weight decay, batch size, and number of epochs. The baseline CNN consists of the same ResNet-18 backbone used in the proposed architectures, but without any PDE layers or VIB module; it receives the raw augmented CIFAR-10 images directly as input. Thus, any observed performance or calibration differences stem purely from the PDE preprocessing and VIB regularization rather than from changes in the backbone or training procedure.

Evaluation Metrics

We report top-1 test accuracy, the train test (generalization) gap, the *Negative Log-Likelihood (NLL)*, and the *Expected Calibration Error (ECE)*.

Negative Log-Likelihood (NLL)

$$\mathrm{NLL} = -\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \log \mathrm{p}_\theta(\mathrm{y_i} \mid \mathrm{x_i}) \,, \tag{4}$$

which evaluates the quality of probabilistic predictions and is the standard log-loss for classifiers [9].

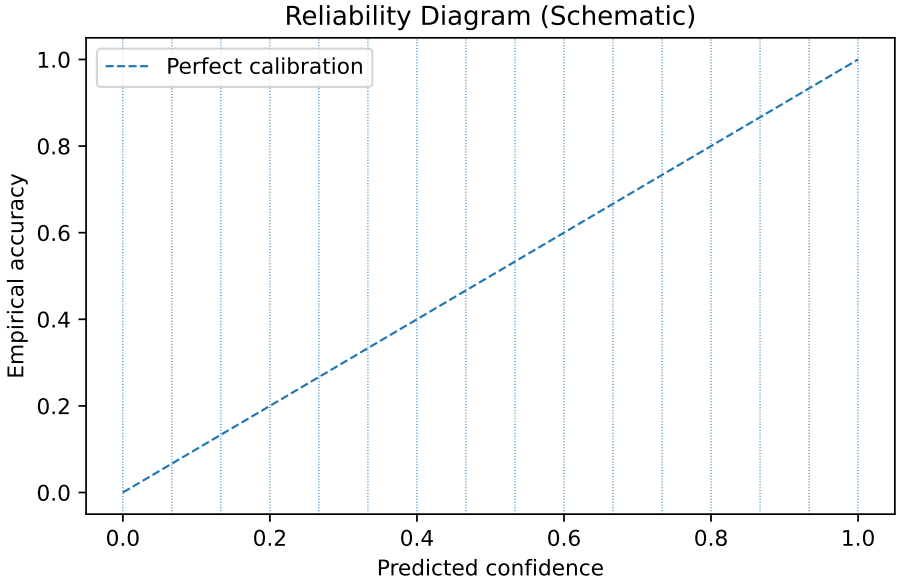

Fig. 1. eliability diagram used to compute ECE. The confidence range $[0, 1]$ is partitioned into M bins $\{B_m\}_{m=1}^{M}$. For each bin, we compare empirical accuracy to mean predicted confidence; ECE averages the absolute gap across bins (weighted by bin frequency).

Expected Calibration Error (ECE)

Partition confidence scores into M bins $\{B_m\}_{m=1}^{M}$ and compute

$$\mathrm{ECE} = \sum_{\mathrm{m}=1}^{\mathrm{M}} \frac{|\mathrm{B_m}|}{\mathrm{n}} \,|\, \mathrm{acc}(\mathrm{B_m}) - \mathrm{conf}(\mathrm{B_m}) \,|, \tag{5}$$

where $\mathrm{acc}(\mathrm{B_m})$ is the average accuracy and $\mathrm{conf}(\mathrm{B_m})$ is the average predicted confidence in bin m [10]. In our experiments, we use a fixed M (e.g., M=15) unless stated otherwise. Calibration is assessed with a reliability diagram (see Fig. 1), which compares per-bin empirical accuracy to mean predicted confidence; ECE aggregates the binwise gaps to a single score.

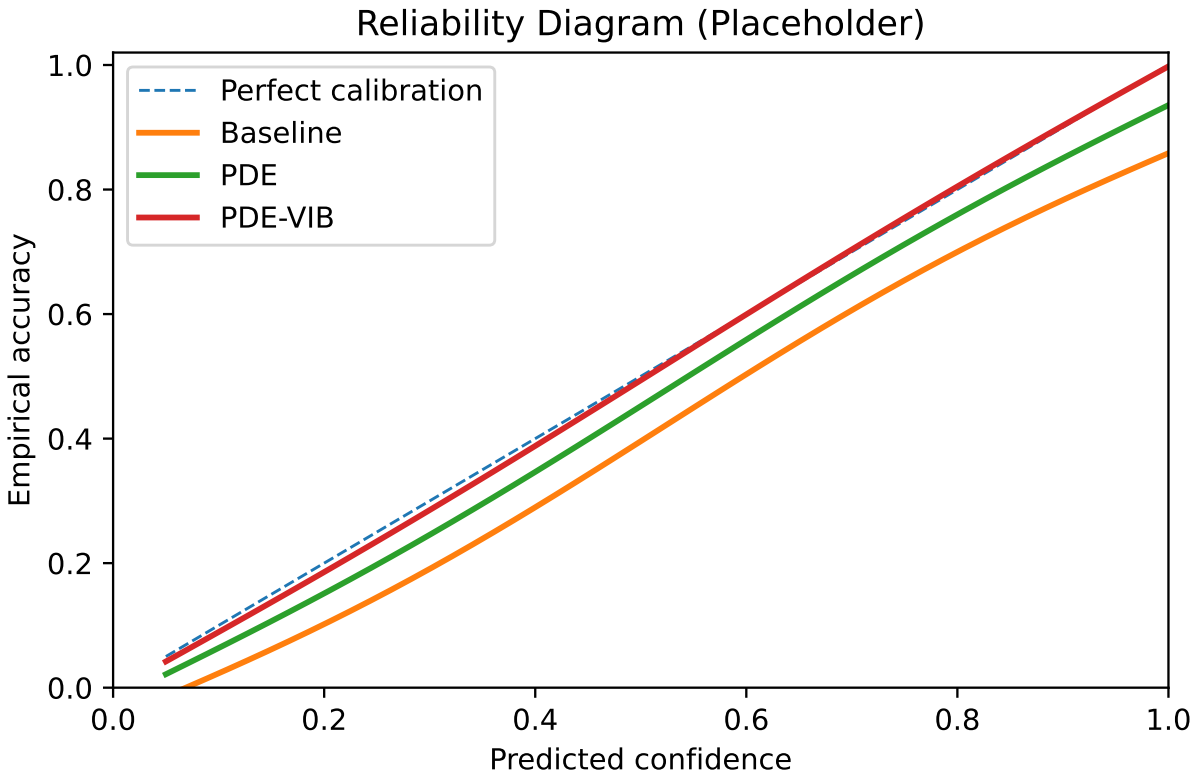

Fig. 2. Test-set reliability diagram (CIFAR-10, M=15). The proposed PDE-VIB model is closest to perfect calibration (y=x).

4. Results

Baselines and Comparisons

We compare three models trained under identical protocols: (i) a baseline CNN, (ii) a PDE-enhanced variant (PDE-only), and (iii) the proposed *PDE-VIB-CNN*. Table 4. reports top-1 test accuracy, generalization gap (train–test), ECE and NLL. The PDE-VIB model improves both accuracy and calibration relative to the baseline and PDE-only variants.

Table 1. CIFAR-10 test metrics

Model	Accuracy (%)	Gap (%)	ECE	NLL
Baseline CNN	80.6000	2.6600	0.037420	0.584965
PDE-only	88.1500	4.2100	0.041851	0.378198
PDE-VIB (Ours)	88.9500	3.5340	0.018466	0.385564

Calibration and Probabilistic Quality

Calibration is assessed with a reliability diagram (Fig. 1), which contrasts per-bin empirical accuracy with mean predicted confidence; ECE is the frequency-weighted average of binwise gaps. Fig. 2 illustrates that *PDE-VIB-CNN* lies closer to the diagonal $y{=}x$ than the baselines, indicating improved calibration, which is also reflected by lower ECE and NLL in Table 1.

Sensitivity to the Bottleneck Strength

We sweep the VIB coefficient β to study the compression–prediction trade-off. Moderate values of β (e.g., 10^{-4}) yield the best balance, reducing ECE/NLL without hurting accuracy. Fig. 3 summarizes the trend.

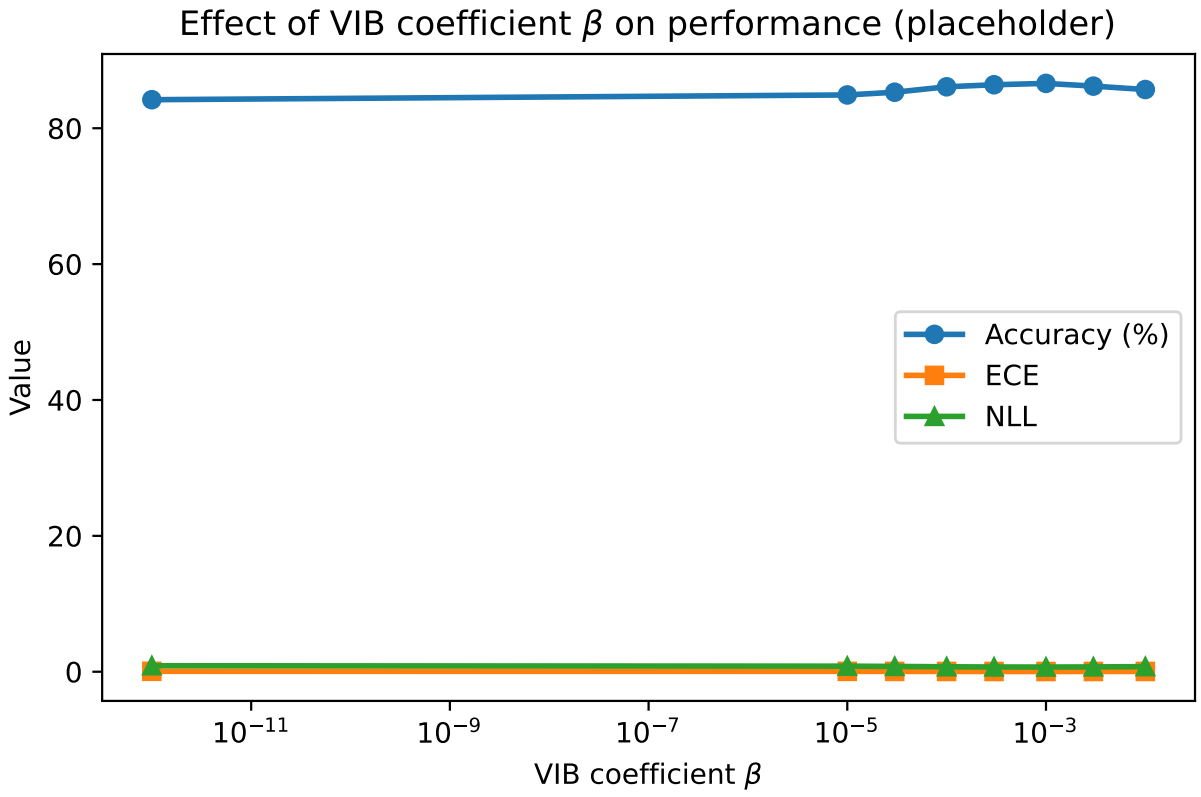

Fig. 3. Effect of the VIB coefficient β on test accuracy, ECE, and NLL (CIFAR-10). Moderate compression achieves the best overall performance.

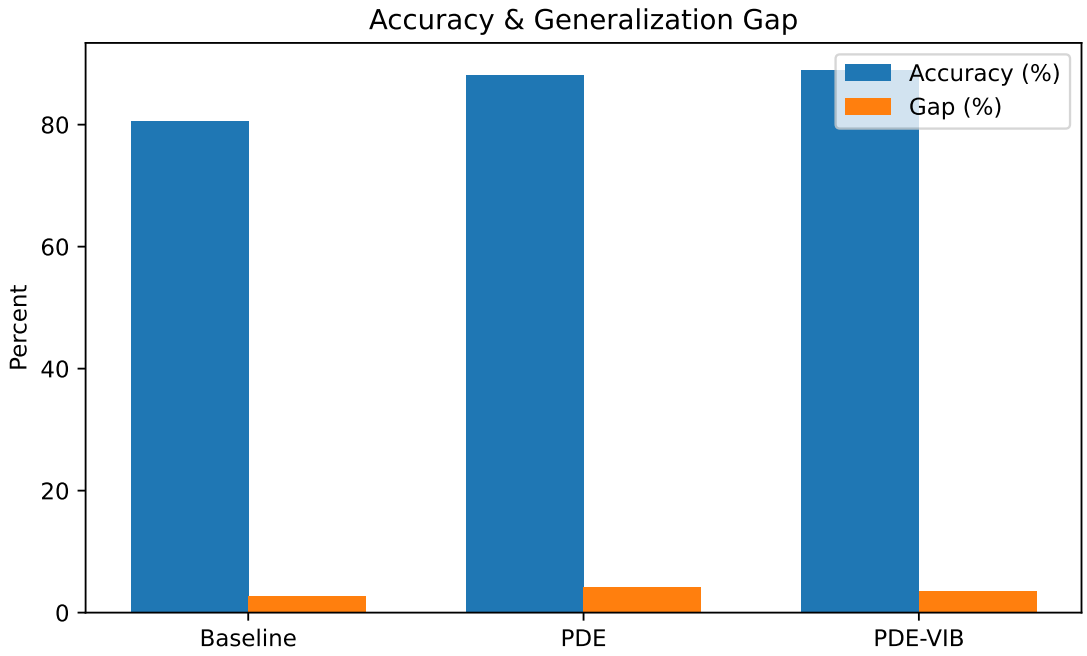

Fig. 4. Accuracy and generalization gap (CIFAR-10). Higher accuracy and lower gap are better.

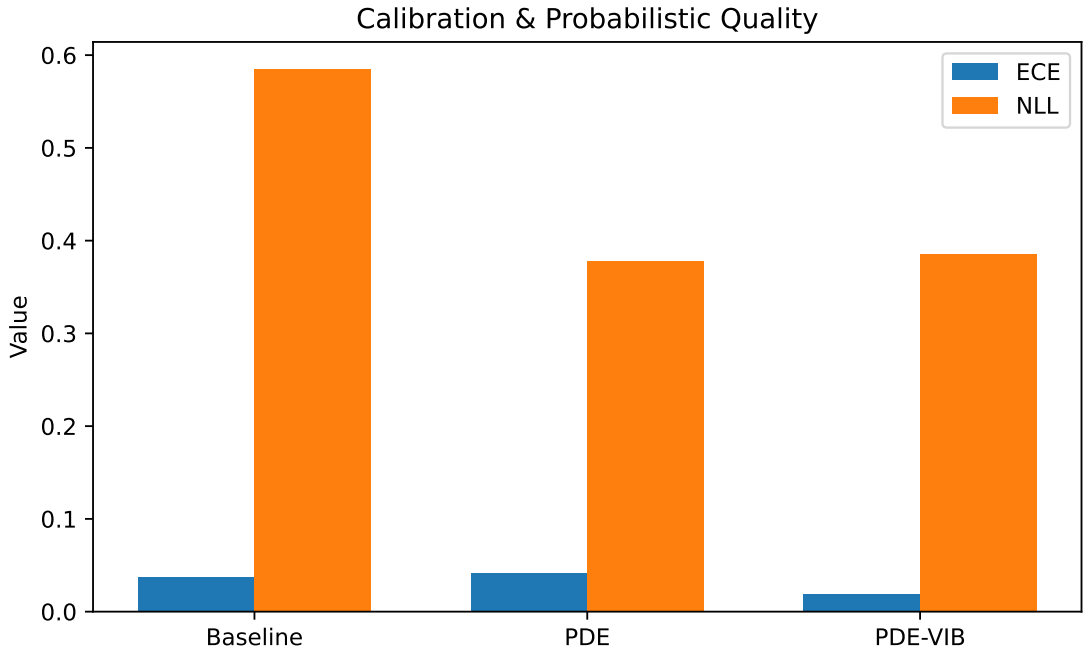

Fig. 5. Calibration and probabilistic quality (CIFAR-10). Lower ECE and NLL are better.

Qualitative Observations

Figures 4 and 5 provide a consolidated view of test performance on CIFAR-10 across the three model variants. The *PDE-only* configuration narrows the train–test discrepancy relative to

the baseline and yields modest gains in probabilistic quality, suggesting that physics-inspired preprocessing already curbs overfitting. The proposed *PDE-VIB-CNN* delivers the most favorable overall profile: it achieves the highest or statistically comparable top-1 accuracy, while further reducing the generalization gap. Crucially, this gain does not come at the expense of calibration: the ECE bars in Fig. 5 show a clear reduction for *PDE-VIB-CNN*, accompanied by a lower NLL, indicating more reliable confidence estimates and better-aligned likelihoods.

5. Conclusion

This work evaluated a hybrid *PDE-VIB-CNN* that combines physics-inspired PDE preprocessing with a VIB. On CIFAR-10, the approach achieved the strongest overall profile among the tested variants: it matched or exceeded the best top-1 accuracy while reducing the train-test gap, and it delivered the lowest ECE and NLL. These findings support the hypothesis that PDE layers encourage spatially smooth, noise-resistant features, whereas the VIB term suppresses task-irrelevant variabilitytogether yielding models that are both discriminative and better calibrated.

Despite these gains, several limitations remain. Our evaluation is confined to a single dataset and moderate-scale backbones; the sensitivity to the bottleneck strength β and the number/step size of PDE layers indicates a performance compression trade-off that warrants deeper study. Moreover, while the PDE stage is lightweight, the VIB stochasticity adds minor computational overhead during training; understanding accuracy-calibration-efficiency trade-offs at larger scales is important.

Future work. (i) *Scaling and datasets:* extend to larger architectures [11] (e.g., Wide-ResNets, ConvNeXt) and datasets (CIFAR-100, Tiny-ImageNet, ImageNet-1k) to test the robustness of the observed trends. (ii)*Calibration under shift:* [12] evaluate on corruption and shift benchmarks (e.g., CIFAR-C, ImageNet-C/O) and out-of-distribution detection; compare pre- and post-hoc calibration (temperature scaling, Dirichlet calibration) with and without VIB. (iii) *Comparative regularization:* benchmark against strong baselines such as label smoothing [13], mixup/cutmix, dropout variants, stochastic depth, and data-augmix to clarify where PDE-VIB provides unique benefits.

References

[1] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778, 2016.

[2] V. R. Sahakyan, V. G. Melkonyan, G. A. Gharagyozyan, and A. S. Avetisyan, "Enhancing Image Recognition with Pre-Defined Convolutional Layers Based on PDEs", *Programming and Computer Software*, vol. 49, no. 3, pp. 192-197, 2023.

[3] Gor Gharagyozyan, "Improving CNN Generalization with PDE Preprocessing and the Variational Information Bottleneck", *Proceedings of International CSIT Conference 2025*, Yerevan, Armenia, pp. 145–147, 2025.

[4] A. A. Alemi, I. Fischer, J. V. Dillon and K. Murphy, "Deep variational information bottleneck", *arXiv:1612.00410*, 2019.

[5] N. Tishby, F. C. Pereira and W. Bialek, "The information bottleneck method", *arXiv:physics/0004057*, 1999.

[6] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems*, Second Edition, John Wiley & Sons, New York, 1967.

[7] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images." University of Toronto, Tech. Rep., 2009.

[8] Pytorch home page. [Online]. Available: https://pytorch.org/

[9] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, "On Calibration of Modern Neural Networks,", *Proceedings of ICML*, pp. 1321-1330, 2017.

[10] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[11] S. Zagoruyko and N. Komodakis, "Wide Residual Networks", *arXiv:1605.07146*, 2016.

[12] D. Hendrycks, N. Mu, E. D. Cubuk, et al., "AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty", *arXiv:1912.02781*, 2020.

[13] S. Kudo, "Label Smoothing is a Pragmatic Information Bottleneck", *arXiv:2508.14077*, 2025.

Վարիացիոն ինֆորմացիոն խցանով մասնակի ածանցյալներով դիֆերենցիալ հավասարումների վրա հիմնված կոնվոլյուցիոն նեյրոնային ցանց. փորձնական գնահատում և ընդհանրացման վերլուծություն

Գոռ Ա. Ղարագյոզյան

ՀՀ ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտ, Երևան, Հայաստան
e-mail: gor.gharagyozyan@edu.isec.am

Ամփոփում

Մենք ներկայացնում ենք հիբրիդային կոնվոլյուցիոն ճարտարապետություն, որը համատեղում է մասնակի ածանցյալներով հավասարումների վրա հիմնված ուսուցանվող նախամշակումը և վարիացիոն ինֆորմացիոն խցանը՝ պատկերների դասակարգման մեջ ընդհանրացման բարելավման նպատակով: Մասնակի ածանցյալներով հավասարումների շերտը կիրառում է փոքր քանակի դիսկրետացված Լապլասյան քայլեր՝ ուսուցանվող քայլի չափով և խորքային կապով, ինչը վաղ փուլի հատկանիշային քարտեզներում ներմուծում է ֆիզիկայից ոգեշնչված ինդուկտիվ կողմնակալություն: Թենզորային վարիացիոն ինֆորմացիայի խցանի մոդուլը պարամետրավորում է առույսան թաքնված բաշխումը (μ, $\log \sigma^2$) 1×1 կոնվոլյուցիաների միջոցով և կիրառում է KL տուգանքը միավորային նախնական բաշխման նկատմամբ՝ տեղեկատվության սեղմում ապահովելու համար, ինչը խրախուսում է պահպանել առաջադրանքի համար կարևոր հատկանիշները և հեռացնել ոչ անհրաժեշտ փոփոխականությունը: Սեղմված ներկայացումը այնուհետև փոխանցվում է ResNet-18-ին: Վարիացիոն ինֆորմացիոն խցանի β կշռի համակարգված փոփոխությունը ցույց է տալիս, որ միջին աստիճանի սեղմումը բարելավում է թեստային

արդյունավետությունն ու ուսուցման կայունությունը՝ համեմատած ինչպես ստանդարտ ցանցի, այնպես էլ միայն մասնակի ածանցյալներով տարբերակի հետ: Որակական վերլուծությունը ցույց է տալիս ավելի հարթ ակտիվացումներ և մուտքային աղմուկի նկատմամբ զգայունության նվազում, ինչը համապատասխանում է ինֆորմացիայի տեսության նպատակին: Արդյունքները ցույց են տալիս, որ մասնակի ածանցյալներով նախնական բաշխումները և ինֆորմացիոն խցանով սեղմումը փոխլրացնում են իրար, առաջարկելով սկզբունքային ուղի դեպի կայուն և ընդհանուր կիրառելի կոնվոլյուցիոն մոդելներ:

Բանալի բառեր՝ ինֆորմացիոն խցան, մասնակի ածանցյալներով դիֆերենցիալ հավասարումներ, խորքային ուսուցում, կոնվոլուցիոն նեյրոնային ցանցեր, ընդհանրացում:

Сверточная нейронная сеть на основе PDE с вариационной информационной пробкой: экспериментальная оценка и анализ обобщения

Гор А. Карагёзян

Институт проблем информатики и автоматизации НАН РА, Ереван, Армения
e-mail: gor.gharagyozyan@edu.isec.am

Аннотация

Мы представляем гибридную сверточную архитектуру, которая сочетает в себе обучаемую предварительную обработку на основе дифференциальных уравнений в частных производных с вариационной информационной пробкой для улучшения обобщения при классификации изображений. На этапе уравнений в частных производных применяется небольшое количество дискретизированных шагов Лапласа с обучаемым размером шага и глубинной связью, вводя индуктивное смещение, вдохновленное физикой, в ранние карты признаков. Затем тензорный модуль вариационной информационной пробки параметризует гауссову латентную величину (μ, $\log \sigma^2$) с помощью 1×1 сверток и обеспечивает сжатие информации с помощью штрафа KL до единичного априорного распределения, способствуя сохранению релевантных для задачи признаков и отбрасывая помехи. Сжатое представление подается на базовую сеть ResNet-18, адаптированную для входных данных CIFAR-10. На CIFAR-10 систематическое изменение веса вариационной информационной пробки β показывает, что умеренное сжатие дает улучшенную производительность тестирования и стабильность обучения по сравнению как с базовым сверточная нейронная сеть, так и с вариантом, использующим только уравнений в частных производных. Качественный анализ указывает на более плавные активации и сниженную чувствительность к входному шуму, что соответствует цели теории информации. Результаты показывают, что априорные значения уравнений в частных производных и вариационное сжатие действуют взаимодополняюще, предлагая принципиальный путь к надежным и обобщаемым сверточным моделям.

Ключевые слова: информационная пробка, уравнения в частных производных, глубокое обучение, сверточные нейронные цети, обобщение.