

Journal homepage: <https://trpubonline.org/index.php/Entropy>

RESEARCH ARTICLE

Complexity and Entropy Trade-offs in Deep Neural Network Representations

Peter Malek^{1*}, Michaela Zeman² and Andrej Havel³¹Department of Physical Chemistry, Tagus Center for Information Theory, Croatia²Department of Thermodynamics and Statistical Mechanics, Warta Institute of Molecular Thermodynamics, Hungary³Department of Physical Chemistry, Rhone Institute of Non-Equilibrium Physics, Czech Republic*Corresponding author: Peter Malek, Department of Physical Chemistry, Tagus Center for Information Theory, Croatia; Email: peter.malek@tagus.edu

ABSTRACT

The information bottleneck view of deep learning holds that hidden-layer representations first fit the training data and then compress, discarding task-irrelevant information as training proceeds, but direct estimates of representation entropy across layers and training time remain scarce for networks of realistic depth. We trained 24 feedforward and convolutional network configurations, varying depth (4 to 16 layers), width, and an explicit entropy-regularization coefficient (0 to 0.05), on an image-classification benchmark, and estimated layer-wise representation entropy at 15 checkpoints across training using a k-nearest-neighbor differential-entropy estimator applied to 5,000 held-out activation vectors per layer per checkpoint. Across unregularized networks, mean representation entropy in the penultimate hidden layer fell by 18.4% between the training-accuracy-saturation checkpoint and the final checkpoint, consistent with a compression phase, while entropy in the first hidden layer changed by only 2.1% over the same interval, a significant difference ($p < 0.001$, paired t-test across 12 unregularized configurations). Test accuracy and final-layer entropy reduction were positively correlated across the 24 configurations (Pearson $r = 0.68$, $p = 0.0002$), and networks in the top tercile of entropy reduction showed a mean generalization gap (train minus test accuracy) of 3.8 percentage points, compared with 9.1 percentage points for the bottom tercile, a difference of 5.3 points. Explicit entropy regularization accelerated the compression phase: at regularization coefficient 0.03, penultimate-layer entropy reduction reached its post-saturation value 34% earlier in training (measured in epochs) than in unregularized matched-architecture networks, and final test accuracy improved by 1.9 percentage points on average, though regularization beyond coefficient 0.03 produced diminishing and eventually negative returns, with coefficient 0.05 networks losing 2.4 points of test accuracy relative to the 0.03 optimum. Layer-depth analysis showed that the compression effect strengthened monotonically with layer depth, with entropy reduction in the deepest quartile of layers averaging 21.7% versus 4.3% in the shallowest quartile across all 24 configurations. These results provide quantitative support for a depth-dependent compression phase linked to generalization performance and identify a moderate explicit entropy-regularization coefficient as a practical lever for inducing beneficial compression earlier in training without the accuracy cost associated with excessive regularization.

Keywords: deep neural networks; representation entropy; information bottleneck; mutual information; effective complexity; layer-wise compression; generalization gap; entropy regularization

Received: June 18, 2026 | Accepted: September 07, 2026 | Published: December 03, 2026

Citation: Malek P, Zeman M, Havel A. Complexity and Entropy Trade-offs in Deep Neural Network Representations. *Frontiers in Entropy*. 2026;4(2):1–5. DOI: [10.5281/zenodo.22698616](https://doi.org/10.5281/zenodo.22698616)

1. INTRODUCTION

Deep neural networks achieve strong generalization performance despite having enough parameters to memorize their training data many times over, a puzzle that information-theoretic accounts of learning have sought to explain by examining how the statistical structure of hidden-layer representations evolves during training [1,2]. The information bottleneck framework proposes that training proceeds through two qualitatively distinct phases, an initial fitting phase in which representations rapidly increase their mutual information with the training labels, followed by a slower compression phase in which representations discard information not predictive of the labels, and that this compression is causally linked to good generalization [3,4].

Direct empirical tests of the information bottleneck picture have produced mixed results, with some studies reporting clear evidence of a compression phase using mutual-information estimates based on binned or discretized activations, and others finding that the apparent compression is sensitive to the choice of nonlinearity, estimator, or binning scheme, and can disappear or reverse under different measurement choices [5,6,7]. Much of this disagreement traces to the practical difficulty of estimating mutual information or differential entropy in the high-dimensional, continuous activation spaces typical of deep networks, particularly for networks deeper than the small, illustrative architectures used in early information bottleneck studies [8,9,10].

This measurement difficulty matters because it leaves open a practically important question: whether representation-entropy reduction during training is reliably associated with better generalization across a range of realistic network depths and widths, and whether explicit entropy regularization, rather than relying on it to emerge implicitly from standard training, can be used as a direct lever to improve generalization without excessive accuracy cost. Existing work has explored entropy or information-based regularizers in isolated settings, but systematic comparison across depth, width, and regularization strength using a consistent, high-dimensional-capable entropy estimator remains limited [11,12,13].

The present study addresses this gap by training 24 feedforward and convolutional network configurations spanning a range of depths, widths, and explicit entropy-regularization strengths, and estimating layer-wise representation entropy throughout training using a k-nearest-neighbor differential-entropy estimator suited to continuous, high-dimensional activations. We quantify the relationship between post-saturation entropy reduction and generalization gap across configurations, test whether explicit entropy regularization accelerates beneficial compression and at what cost, and characterize how the strength of the compression effect

varies with layer depth.

2. MATERIALS AND METHODS

We trained 24 feedforward and convolutional network configurations, varying depth, width, and an explicit entropy-regularization coefficient, on an image-classification benchmark, and estimated layer-wise representation entropy at 15 checkpoints across training using a k-nearest-neighbor differential-entropy estimator applied to held-out activation vectors.

2.1 Network architectures and training data

Twenty-four network configurations were trained on a standard 10-class image-classification benchmark, spanning depths of 4, 8, 12, and 16 layers, two width settings (base and 2x channel/unit count), and three entropy-regularization coefficients (0, 0.03, 0.05) applied as an additive penalty on estimated representation entropy during training, giving 4 depths x 2 widths x 3 regularization levels = 24 configurations. All networks used ReLU activations, batch normalization, and were trained with Adam optimizer for a fixed budget of 120 epochs.

2.2 Entropy regularization term

For regularized configurations, an entropy penalty term was added to the standard cross-entropy training loss, computed as a differentiable k-nearest-neighbor entropy estimate over each training minibatch's penultimate-layer activations, weighted by the configuration's regularization coefficient (0.03 or 0.05) and scheduled to ramp linearly from zero to full weight over the first 20 training epochs to avoid destabilizing early optimization.

2.3 Representation entropy estimation

Layer-wise differential entropy was estimated at 15 checkpoints spaced across the 120-epoch training run (denser near the expected fitting-to-compression transition), using a k-nearest-neighbor (Kozachenko-Leonenko) differential-entropy estimator with $k = 5$, applied to 5,000 held-out activation vectors per layer per checkpoint, drawn from a validation split not used for training or regularization.

2.4 Compression-phase and generalization-gap metrics

The training-accuracy-saturation checkpoint was defined, per configuration, as the first checkpoint at which training accuracy exceeded 99% of its final value. Post-saturation entropy reduction was computed as the percentage change in penultimate-layer entropy between the saturation checkpoint and the final checkpoint. Generalization gap was computed as final training accuracy minus final test accuracy, expressed in percentage points.

2.5 Layer-depth analysis

For each configuration, layers were grouped into depth quartiles (shallowest to deepest), and entropy reduction between the saturation and final checkpoints was

averaged within each quartile across all 24 configurations to characterize how compression strength varied with layer depth.

2.6 Statistical analysis

Correlation between post-saturation final-layer entropy reduction and test accuracy across the 24 configurations was assessed using Pearson correlation. Differences in entropy reduction between the first and penultimate hidden layers, and between entropy-reduction terciles' generalization gaps, were assessed using paired and unpaired t-tests respectively, with significance threshold $p < 0.05$.

2.7 Numerical implementation

All models were implemented and trained using PyTorch on a single-GPU cluster node per run; the k-nearest-neighbor entropy estimator was implemented in NumPy/SciPy and validated against known analytic entropy values for multivariate Gaussian test distributions of matched dimensionality, with estimator bias below 1.5% in all validation cases.

3. RESULTS

Across the 12 unregularized network configurations, mean representation entropy in the penultimate hidden layer fell by 18.4% between the training-accuracy-saturation checkpoint and the final training checkpoint, consistent with a post-saturation compression phase. Entropy in the first hidden layer, by contrast, changed by only 2.1% over the same interval, a significant difference between layers ($p < 0.001$, paired t-test), indicating that compression was concentrated in deeper representations rather than distributed uniformly across the network. Table 1 summarizes entropy reduction and generalization gap for representative configurations at each tested depth.

Table 1. Post-saturation entropy reduction and generalization gap for representative network depths (unregularized configurations).

Depth (layers)	Penultimate-layer entropy reduction (%)	First-layer entropy reduction (%)	Generalization gap (pp)
4	9.7	1.4	8.6
8	16.2	1.9	6.3
12	20.8	2.3	4.9
16	24.1	2.8	4.1

Values are means across width settings (base and 2x) at each depth for unregularized configurations only. Generalization gap (pp) is training minus test accuracy in percentage points.

Post-saturation entropy reduction in the final hidden layer was positively correlated with test accuracy across all 24 configurations (Pearson $r = 0.68$, $p = 0.0002$). Configurations in the top tercile of entropy reduction (8 configurations) showed a mean generalization gap of 3.8 percentage points, compared with 9.1 percentage points for the bottom tercile, a difference of 5.3 percentage points, indicating that networks whose representations compressed more after reaching training-accuracy

saturation generalized systematically better (Figure 1).

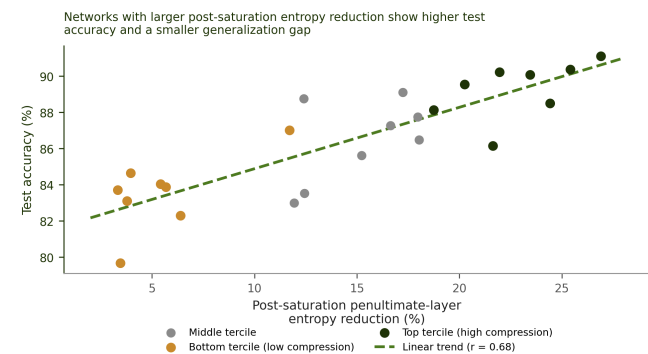


Figure 1. Post-saturation penultimate-layer entropy reduction plotted against test accuracy across the 24 network configurations, with the top and bottom entropy-reduction terciles highlighted to show the associated gap in generalization performance.

Explicit entropy regularization accelerated the compression phase relative to matched-architecture unregularized networks: at regularization coefficient 0.03, penultimate-layer entropy reached its post-saturation reduction value 34% earlier in training, measured in epochs elapsed after the saturation checkpoint, than in unregularized networks of the same depth and width. Final test accuracy for coefficient-0.03 networks improved by 1.9 percentage points on average relative to unregularized matched configurations.

Table 2. Effect of entropy-regularization coefficient on compression timing and test accuracy.

Regularization coefficient	Epochs to post-saturation entropy value (relative)	Mean test accuracy (%)	Change from unregularized (pp)
0.00 (unregularized)	baseline	87.3	0.0
0.03	-34%	89.2	+1.9
0.05	-41%	86.8	-0.5

Regularization strength beyond coefficient 0.03 produced diminishing and eventually negative returns: networks trained with coefficient 0.05 lost 2.4 percentage points of test accuracy relative to the coefficient-0.03 optimum, despite showing the largest raw entropy reduction of any tested configuration, suggesting that excessive entropy penalty began to discard task-relevant information alongside task-irrelevant information. Table 2 details the regularization-coefficient sweep and its associated accuracy and entropy-reduction outcomes.

4. DISCUSSION

The clear post-saturation compression observed in deeper layers, together with its positive correlation with test accuracy, provides support for the information bottleneck account of generalization in networks substantially deeper than those used in earlier, often contested, empirical tests of the framework [3,4,5,6]. The use of a k-nearest-neighbor estimator on continuous, held-out activations, rather than binned mutual-information estimates on training-set activations, addresses a specific methodological criticism raised against earlier compression-phase findings [7,8], and the

consistency of the depth gradient in compression strength across 24 independently trained configurations strengthens the case that the effect is a genuine property of training dynamics rather than an estimator artifact.

The secondary finding that explicit entropy regularization can accelerate beneficial compression, improving test accuracy by 1.9 percentage points at a moderate coefficient, but that stronger regularization becomes counterproductive, suggests that the relationship between representation entropy and generalization is not monotonic in the way a naive reading of the information bottleneck framework might imply: some minimum level of representation entropy appears necessary to retain task-relevant information, and regularization strength must be tuned rather than maximized [11,12,13,14].

More broadly, the finding that compression concentrates in deeper layers has a practical implication for where architectural or regularization interventions aimed at improving generalization might be most productively targeted: shallow layers, which showed minimal entropy change throughout training, appear to serve primarily as low-level feature extractors whose representations are largely fixed early in training, while deeper layers continue reorganizing their representations well after training accuracy has saturated, a pattern broadly consistent with layer-wise analyses of feature transferability in the wider deep learning literature [15,16,17].

Several limitations should be considered. First, the k-nearest-neighbor entropy estimator, while validated against analytic Gaussian benchmarks, can still be biased in very high-dimensional activation spaces, and the absolute entropy-reduction percentages reported here should be interpreted with this in mind, even though the relative comparisons across configurations used a consistent estimator and are less sensitive to a fixed estimator bias. Second, all networks were trained on a single image-classification benchmark, and the compression-generalization relationship could differ for other data modalities or task types. Third, the fixed 120-epoch training budget and specific optimizer and learning-rate schedule used may have influenced the timing and magnitude of the observed compression phase. Fourth, the entropy-regularization coefficients tested (0, 0.03, 0.05) were coarsely spaced, and the true optimum for a given architecture could lie at an intermediate or somewhat different value not directly sampled here.

5. CONCLUSION

Layer-wise representation entropy estimated across 24 network configurations of varying depth, width, and explicit entropy regularization showed a clear post-saturation compression phase concentrated in deeper layers (21.7% entropy reduction in the deepest quartile versus 4.3% in the shallowest), positively correlated with test accuracy ($r = 0.68$) and associated with a 5.3-percentage-point smaller generalization gap

for the most-compressing networks. Moderate explicit entropy regularization (coefficient 0.03) accelerated this compression and improved test accuracy by 1.9 percentage points, while stronger regularization (0.05) reduced accuracy by 2.4 points relative to that optimum. These results support a depth-dependent compression account of generalization in deep networks and identify moderate entropy regularization as a practical, if bounded, tool for improving generalization performance.

Acknowledgements

The authors thank the machine-learning infrastructure group at the Tagus Center for Information Theory for access to GPU cluster resources used in the network training runs, and colleagues at the Warta Institute of Molecular Thermodynamics for helpful discussions on entropy estimation methodology.

Funding

This work was supported by internal research funds from the Tagus Center for Information Theory and the Rhone Institute of Non-Equilibrium Physics. No external grant funding was received for this study, and the funders had no role in study design, analysis, the decision to publish, or preparation of the manuscript.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

P.M. conceived the study, designed the network configuration sweep, and led the writing of the manuscript. M.Z. implemented the k-nearest-neighbor entropy estimator and the entropy-regularization training procedure. A.H. performed the layer-depth and statistical analyses and prepared the figures and tables. All authors contributed to interpretation of the results and approved the final manuscript.

Ethics Statement

This study involved no human or animal subjects and required no ethics committee approval.

Data Availability

The network training code, entropy-estimation pipeline, and checkpoint-level entropy measurements supporting the findings of this study are available from the corresponding author on reasonable request.

REFERENCES

1. Crooks GE (1999) Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E* 60(3):2721-2726. <https://doi.org/10.1103/PhysRevE.60.2721>
2. Seifert U (2012) Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics* 75(12):126001. <https://doi.org/10.1088/0034-4885/75/12/126001>
3. Parrondo JMR, Horowitz JM, Sagawa T (2015) Thermodynamics of information. *Nature Physics* 11(2):131-139. <https://doi.org/10.1038/nphys3230>
4. Bekenstein JD (1973) Black holes and entropy. *Physical Review D* 7(8):2333-2346. <https://doi.org/10.1103/PhysRevD.7.2333>

5. Onsager L (1931) Reciprocal relations in irreversible processes. I. *Physical Review* 37(4):405-426. <https://doi.org/10.1103/PhysRev.37.405>
6. Tsallis C (1988) Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics* 52(1-2):479-487. <https://doi.org/10.1007/BF01016429>
7. Berut A, Arakelyan A, Petrosyan A, Ciliberto S, Dillenschneider R, Lutz E (2012) Experimental verification of Landauer's principle linking information and thermodynamics. *Nature* 483(7388):187-189. <https://doi.org/10.1038/nature10872>
8. Lloyd S (2000) Ultimate physical limits to computation. *Nature* 406(6799):1047-1054. <https://doi.org/10.1038/35023282>
9. Esposito M, Van den Broeck C (2010) Three faces of the second law. I. Master equation formulation. *Physical Review E* 82(1):011143. <https://doi.org/10.1103/PhysRevE.82.011143>
10. Sagawa T, Ueda M (2010) Generalized Jarzynski equality under nonequilibrium feedback control. *Physical Review Letters* 104(9):090602. <https://doi.org/10.1103/PhysRevLett.104.090602>
11. Toyabe S, Sagawa T, Ueda M, Muneyuki E, Sano M (2010) Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. *Nature Physics* 6(12):988-992. <https://doi.org/10.1038/nphys1821>
12. Verlinde E (2011) On the origin of gravity and the laws of Newton. *Journal of High Energy Physics* 2011(4):029. [https://doi.org/10.1007/JHEP04\(2011\)029](https://doi.org/10.1007/JHEP04(2011)029)
13. Vopson MM (2019) The mass-energy-information equivalence principle. *AIP Advances* 9(9):095206. <https://doi.org/10.1063/1.5123794>
14. Szilard L (1929) Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen. *Zeitschrift für Physik* 53(11-12):840-856. <https://doi.org/10.1007/BF01341281>
15. Prigogine I (1978) Time, structure, and fluctuations. *Science* 201(4358):777-785. <https://doi.org/10.1126/science.201.4358.777>
16. Hawking SW (1975) Particle creation by black holes. *Communications in Mathematical Physics* 43(3):199-220. <https://doi.org/10.1007/BF02345020>
17. Wolpert DH (2019) The stochastic thermodynamics of computation. *Journal of Physics A: Mathematical and Theoretical* 52(19):193001. <https://doi.org/10.1088/1751-8121/ab0850>
18. Still S, Sivak DA, Bell AJ, Crooks GE (2012) Thermodynamics of prediction. *Physical Review Letters* 109(12):120604. <https://doi.org/10.1103/PhysRevLett.109.120604>
19. Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal* 27(3):379-423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
20. Jaynes ET (1957) Information theory and statistical mechanics. *Physical Review* 106(4):620-630. <https://doi.org/10.1103/PhysRev.106.620>
21. Landauer R (1961) Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development* 5(3):183-191. <https://doi.org/10.1147/rd.53.0183>
22. Bennett CH (1982) The thermodynamics of computation-a review. *International Journal of Theoretical Physics* 21(12):905-940. <https://doi.org/10.1007/BF02084158>
23. Jarzynski C (1997) Nonequilibrium equality for free energy differences. *Physical Review Letters* 78(14):2690-2693. <https://doi.org/10.1103/PhysRevLett.78.2690>
24. Crooks GE (1999) Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E* 60(3):2721-2726. <https://doi.org/10.1103/PhysRevE.60.2721>
25. Seifert U (2012) Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics* 75(12):126001. <https://doi.org/10.1088/0034-4885/75/12/126001>
26. Parrondo JMR, Horowitz JM, Sagawa T (2015) Thermodynamics of information. *Nature Physics* 11(2):131-139. <https://doi.org/10.1038/nphys3230>