

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 ON THE DECREASE OF GRANDMOTHER-ENTROPY IN AI

Anonymous authors

Paper under double-blind review

ABSTRACT

Machine intelligence is best achieved today through deep learning. We propose a method to quantify the information content of activation values in a neural network layer for a given input. First, we map the activation of each node to a probability that reflects the likelihood that this node is selected among all nodes in the layer, where the more activation, the higher the probability. Then, we measure the information of the layer across channels or features, using conditional entropy, and we refer to it as grandmother-entropy. Empirical evaluation of ResNet-50 shows that the grandmother-entropy decreases as the input image propagates forward through the network layers. Moreover, the grandmother-entropy of the last convolution layer provides a reliable confidence measure to the classification result.

1 INTRODUCTION

1.1 INTELLIGENCE AND DEEP LEARNING

Deep learning LeCun et al. (2015) is today a central tool for building intelligent systems. By deep learning, we mean deep hierarchical neural networks trained with gradient descent. A common technique within this framework — skip connections — have become standard in most state-of-the-art models.

In this work, we explore the role of information in deep learning, paying particular attention to this technique. Our study is both theoretical and empirical. On the theoretical side, we introduce a measure of information for deep learning layers. On the empirical side, we evaluated this measure on ResNet-50 He et al. (2016), a widely used architecture that remains a benchmark in large-scale image classification. Rather than proposing a new architecture or application, our goal is to provide a clearer understanding of deep neural networks operations and to shed light on how classification models convey confidence in their predictions.

The study of information content in neural networks (NN) covers various aspects of data processing.

One direction of research treats the variables updated during learning, namely the weights and biases, as random variables and associates them with a measure of information, a tutorial Hoefler et al. (2021) provides the references.

Another direction considers the input vector x as a random variable originating from a distribution $P(x)$ of all possible inputs for which the machine is learning. For example, images belonging to a given class can be viewed as originating from a specific class distribution. Extending this idea, the activation distribution in each layer, given an input x , is also modeled as a sample of a distribution on all possible activations of that layer. Representative studies include Gabrié et al. (2018); Saxe et al. (2018); Tishby & Zaslavsky (2015) and references therein. The roots of this perspective trace back to the original works of Geiger et al. (1997); Tishby et al. (1999) which proposed that, ideally, the network extracts approximate sufficient statistics of the input with respect to the output, i.e., the network finds a maximally compressed mapping of the input variable that preserves as much of the information on the output variable as possible.

We also point out studies of complexity in a layer, not focusing on the concept of entropy, such as Montufar et al. (2014).

We note that in these previous studies, layers have often been treated as undifferentiated blocks. This perspective overlooks an important limitation: any information measure that remains unchanged under one-to-one rearrangements of these blocks fails to capture the internal organization of a layer. In particular, it ignores how the nodes are structured into feature dimensions or channels. This topic will be an aspect central to our approach.

Another topic addressed in this paper is the confidence measure associated with the output of a classification machine. A confidence measure quantifies how likely a prediction is to be correct. Although the standard softmax output score is often used, it can be misleading and overly confident. Here, we conduct a preliminary study to use the information content that we propose as a possible candidate for the confidence measure.

1.2 OUR WORK: GRANDMOTHER-ENTROPY

The starting hypothesis we make is that the activity of a cell in any layer is informative about the class associated with an input. This hypothesis is inspired by the idea of a "grandmother cell" in neuroscience Barlow (1972); Gross (2002) as well as pioneering work on the sparsity of cell activation values Field (1994); Olshausen & Field (1996). It has the intuition proposed early by Geiger et al. (1997); Tishby et al. (1999) that NNs compress the input to provide selected information to the classification output. Our approach considers the association of the activation value, given an input, with the probability that this node is informative to the class of input. In the extreme case, when a "Jennifer Aniston neuron" Connor (2005) is activated, classification information is immediately inferred. Also, clearly, the last layer of a classification NN is presented in this way, each class being associated with each cell activation. In the best-case scenario, only the correct node is activated in the output layer. In neuroscience the invariance of neuron responses to varying 3D object appearances has been observed Vetter et al. (1995).

Our approach associates with the larger activation of a given node, given an input, a larger probability that such a node is informative towards a classification of such input. Next, we develop an entropy for each layer of a network, associated with an input, that captures the information given within this structure, and we refer to it as the grandmother-entropy, or GM-entropy for simplicity. We then analyze the evolution of this entropy through the layers of the network.

As an application of this entropy we examine whether the entropy of the last convolution layers in a classification network can serve as a more reliable confidence measure than the output softmax values of the top classification unit. We emphasize that this is a preliminary study restricted to data corruption techniques applied to images and we obtain encouraging results.

1.3 ORGANIZATION

The paper is organized as follows. Section 2 presents our main definition of information, a grandmother-entropy. Section 3 performs an empirical study of the propagation of the grandmother-entropy through the layers of ResNet-50. Section 4 considers the grandmother-entropy of the last convolution layer of ResNet-50 as a candidate measure of confidence in the classification output. We emphasize that this is a preliminary study. Section 5 concludes the paper.

2 GRANDMOTHER-ENTROPY IN A NEURAL NETWORK

We are given a Neural Network (NN) layer, indexed by $l = 1, \dots, L$ and with each feature, or channel, indexed by $d = 1, \dots, D_l$ (d for feature dimension). For each feature, the nodes are indexed by $j = 1, \dots, N_l$, i.e., for all features of a layer l , there are N_l nodes. For an image convolution layers are described by $N_l = w_l \times h_l$, where w_l is the width and h_l is the height of the feature layer. Thus, a node or vertex in an NN is indexed by the triplet $v = (j, d, l)$. Note that the total number of nodes in a layer l is then the product $D_l \times N_l$. Given an input x propagating through an NN, each node of the NN is activated with values $a_v = a(j, d, l; x)$.

A layer space in an NN is indexed by (j, d) . Note that at times the pair (j, d) is to be interpreted as indices and at times as variables. Just as in physics, a position of a particle in a discrete lat-

108 tice can have these two interpretations. Throughout the paper, it should be self-explanatory which
 109 interpretation is the most appropriate.
 110

111 We investigate the information content of the activation values a_v in a layer, given an input x . We
 112 first must clarify what we mean by information. Our approach is to associate an activation of a node
 113 with a probability of a node. Here we stress that the activation is not to be interpreted as the outcome
 114 of a probability distribution but rather associated with the probability that a node, and not another
 115 node, is an outcome of a probability distribution over all nodes. Our approach is based on the idea
 116 that a node’s activation value reflects the probability that the node is informative for the classification
 117 task associated with the input.
 118

119 In our empirical studies, we consider the activation values as nonnegative quantities (such is the
 120 case with state-of-the-art ReLu activation functions, and it was also the case for early models using
 121 sigmoid functions). We associate the activation of a node $a(j, d, l; x) \in \mathbb{R}_+$ in a layer l with a
 122 probability as follows.
 123

$$a(j, d, l; x) \geq 0 \quad \Rightarrow \quad P(j, d; l, x) = \frac{1}{Z_l} a(j, d, l; x). \quad (1)$$

124 where
 125

$$Z_l = \sum_{d=1}^{D_l} \sum_{j=1}^{N_l} a(j, d, l), \quad (2)$$

126 is a normalization constant that transforms all node activity into a probability distribution
 127 $P(j, d; l, x) \geq 0$ of a layer probability space l . Interpreting (j, d) as random variables in a layer
 128 probability space indexed by l and given an input x implies a random process that selects integer
 129 values (j, d) with probability $P(j, d; l, x)$. Note that for negative activation values, we would first
 130 map them to their magnitude or square values, so they become non-negative, but we have not per-
 131 formed an empirical study for this case.
 132

133 We could now define the layer information as the Shannon entropy associated with the joint distri-
 134 bution $P(j, d; l, x)$. This entropy would be invariant to any one-to-one transformation of the layer
 135 variables (j, d) to another same size set of two integer variables, as the probabilities in this new
 136 set of integer variables would not change. However, we argue that the information content of an
 137 NN needs to be based on the feature (or channel) architecture. Given an input x to a network, it is
 138 informative that a feature, in a given layer l , responds well or not to a node in position j . Thus, we
 139 look for a layer information definition that reflects the feature structure and thus, it is not invariant
 140 to changes in the NN architecture.
 141

142 We then proceed to factorize the layer probability space and, for simplicity and clarity, drop the input
 143 x and the layer l temporarily from the probability notation. We compute the marginal probability of
 144 each feature dimension of a layer
 145

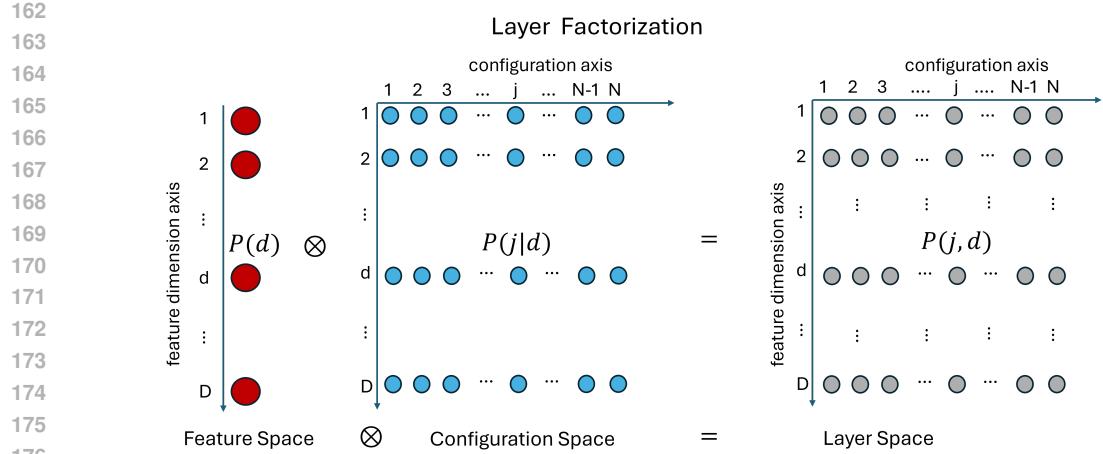
$$P(d) = \sum_{j=1}^{N_l} P(j, d) = \frac{1}{Z_l} \sum_{j=1}^{N_l} a(j, d, l; x). \quad (3)$$

146 From this marginal distribution we obtain the conditional probability distribution for each node
 147 within a layer as follows.
 148

$$P(j|d) = \frac{P(j, d)}{P(d)} = \frac{a(j, d, l; x)}{\sum_{j=1}^{N_l} a(j, d, l; x)}. \quad (4)$$

149 Thus, by Bayes theorem, the layer probability space is factorized into a feature probability space,
 150 with $P(d)$ — prior probabilities, and a configuration probability space, with $P(j|d)$ — conditional
 151 probabilities given a feature dimension (see figure 1 for a visualization of the factorization of the
 152 layer probability space.)
 153

154 This factorization exposes the information structure of the layer; General one-to-one correspondence
 155 transformations of the set $\{j, d\}$ that mix nodes from different feature dimensions would cause
 156 different factorization probabilities of the joint distribution. We are now equipped to define the
 157 information content of a layer associated with this particular layer structure, given an input x , as the
 158



178 Figure 1: A visualization of the probability spaces associated with a layer of the NN. The layer
179 captures the joinXt?X probability $P(j, d)$ and can be factorized as a product of the feature
180 probability space with the configuration probability space per feature dimension. This product of
181 probability spaces requires the operation of one unit probability of the feature space to be multiplied
182 by each unit probability of the configuration space per feature. While the entropy of $P(j, d)$ would
183 be invariant under all one-to-one correspondence transformations of the variables j, d into other two
184 integer variable, the conditional entropy would not. The conditional entropy respects the particular
185 factorization shown above. Note that the conditional entropy is invariant under a global scale of the
186 layer activations.

187
188 conditional entropy

189
190
191
192
193
194
195
196

$$S_l^{\text{GM}}(x) = - \sum_d^{D_l} \sum_{j=1}^{N_l} P(j, d; l, x) \log_2 P(j|d; l, x)$$

$$= \sum_{d=1}^{D_l} P(d; l, x) S_{l,d}^{\text{configuration}}(x) \quad (5)$$

where the configuration entropy, per feature dimension, is given by

197
198
199

$$S_{l,d}^{\text{configuration}}(x) = - \sum_{j=1}^{N_l} P(j|d; l, x) \log_2 P(j|d; l, x), \quad (6)$$

200 and we return the inclusion of the variables l, x to the probabilities notation. The label GM stands
201 for grandmother-entropy. Conditional probabilities $P(j|d; l, x)$ assign the likelihood that a node j , a
202 layer l and feature d , being activated from an input x , and not another node among the ones belonging
203 to the feature layer d, l . Thus, conditional entropy assigns the expected information content to
204 these likelihoods on all features of the layer l . Note that the information of a network layer is
205 invariant under a global scale transformation of the activation, since the scale will reset the value Z
206 in equation 2 while the probabilities remain invariant.

207 Also note that it is common for NN's to have some special layers reduced in the configuration dimension
208 while increasing in the feature dimension, as one moves forward to the output. When such
209 changes occur, the conditional entropy has a bias towards reduction. For example, it is straightfor-
210 ward to verify that a uniform distribution across the entire layer will have the conditional entropy to
211 be reduced with the configuration sizes reduction, regardless of if the feature dimension increases.
212 This effect at these special layers deserves attention when we analyze the GM-entropy.

213 Figure 2 illustrates an example that compares, across the layers of a neural network, the behavior of
214 the entropy of the joint distribution $P(j, d; l, x)$ with the conditional entropy $S_l^{\text{GM}}(x)$. The results
215 support the empirical usefulness of GM-entropy as a measure of the information content of a network
layer.

216 Next, we conduct an empirical study of how the information content $S_l^{\text{GM}}(x)$ evolves as input x
 217 propagates through the ResNet-50 layers. We also analyze the effect of residual connections on the
 218 information content at each layer. In this paper, our experiments focus on networks with pretrained
 219 weights and biases. Extending the study to examine how information evolves during the learning
 220 process is an important direction for future research.

221

222 3 AN EMPIRICAL STUDY OF THE EVOLUTION OF GM-ENTROPY THROUGH 223 RESNET-50

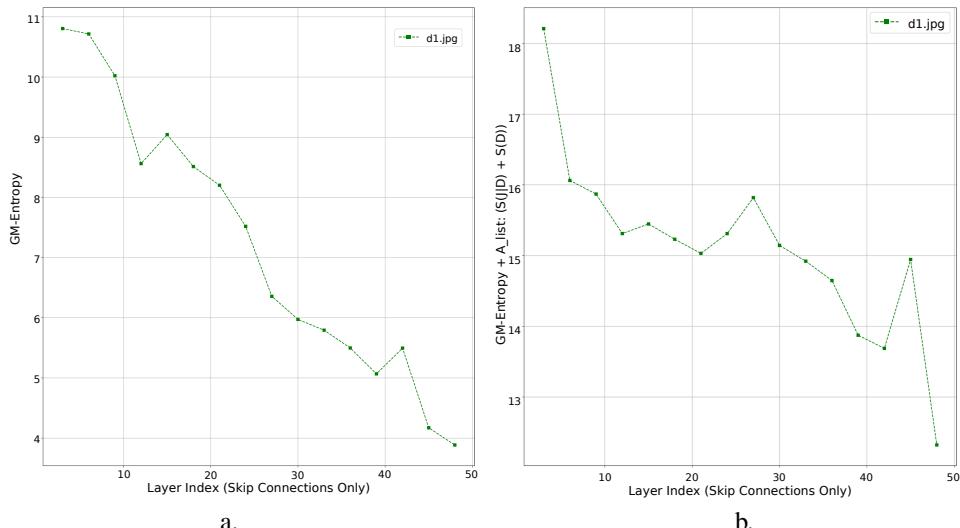
224

225 Consider a trained classification network, specifically ResNet-50, which exemplifies state-of-the-art
 226 performance. ResNet is a convolutional neural network (CNN) that embodies a foundational layer
 227 technique that is now ubiquitous in deep learning: skip connections.

228 The input image x , is a color image comprising three channels (R, G, B). This image is processed
 229 through successive convolution layers in the ResNet architecture, incorporating ReLU activations,
 230 skip connections, and batch normalization, until an average pooling across the spatial dimension
 231 flattens the information and a fully connected layer produces the output where a softmax outputs a
 232 probability distribution $P(c|x)$, over the C output nodes/classes.

233 Note that there are four layers where the spatial size is reduced and the feature dimension is in-
 234 creased, namely the layers 12, 14, 25, 27, 44, 46. This reduction creates a bias effect to reduce the
 235 GM-entropy.

236 In the first empirical study, we investigate the GM-entropy through each of the convolution layers
 237 after skip connections for ResNet-50 (see Figure 2 a). We observe that the GM-entropy decreases
 238 through most of the layers. This illustrative result is also verified by an experiment with the Im-
 239 ageNet21K (Winter 2021 Release) data set Deng et al. (2009); Deng et al., where we randomly
 240 sample 10 classes and 100 images per class, that is, 1000 images test. Table 1 indicates that the
 241 GM-entropy is decreasing.



261 Figure 2: a. GM-entropy evolution through the convolution layers, for an input image shown in fig-
 262 ure 7a d1.jpg. The GM-entropy is shown for all convolution layers, after skip connections and batch
 263 normalization. The GM-entropy decreases for all layers evolution with the exception of two layers.
 264 b. The entropy of the joint distribution $P(j, d; l, x)$. Note that while the GM-entropy decreases con-
 265 sistently through layers the joint distribution does not. The fact that GM-entropy is smaller than the
 266 joint entropy is a mathematical property, as their difference is the entropy of the marginal distribu-
 267 tion $P(d; l, x)$.

268 We test the hypothesis that GM-entropy decreases with depth. For each image, we fit an OLS
 269 (ordinary least squares) regression of GM-Entropy vs. layer index (0–48) and test $H_0 : \beta_1 = 0$ vs.

270 $H_1 : \beta_1 < 0$. A one-sample t -test on slopes yields $t(999) = -1395.6$, $p < 10^{-300}$. A regression
 271 on the layer-wise mean curve (49 points) also shows a strong negative trend (slope -0.1617 , $r =$
 272 -0.976 , $p = 1.27 \times 10^{-32}$), providing strong evidence that GM-entropy decreases with depth in
 273 ResNet-50.

274

275

276

277

278

279

280

Table 1: Per-image OLS fit of GM-entropy vs. layer index (0–48) over $N = 1000$ images.

281

	Mean slope	Intercept	SD	95% CI (slope)
Per-image best-fit line	-0.161708	11.449416	0.00366	-0.161708 ± 0.00023

283

284

285

286

287

288

289

290

3.1 SKIP CONNECTIONS

291

292

293 In skip connection, the input is directly passed through to the next layer along with the output of the
 294 intermediate convolution layers. We hypothesize that, during the intermediate convolution layers,
 295 less information needs to be learned since the aggregation of the input will restore the information of
 296 the previous layer. To test this hypothesis, we investigate the GM-entropy during these intermediate
 297 layers and after the addition of the input, when the information is restored. The results seen in
 298 Figure 3 suggest that this hypothesis is empirically verified in ResNet-50. It suggests that between
 299 skip connections less information is learned, possibly providing an explanation for the efficiency of
 300 skip connection architecture.

301

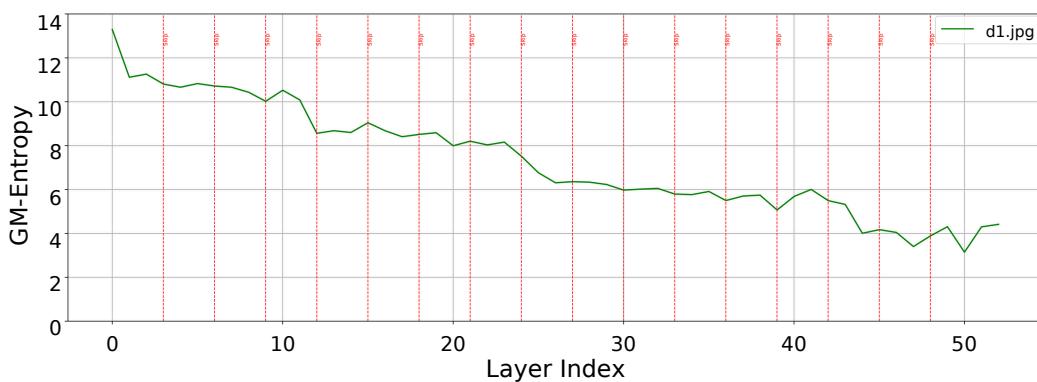
302

303

304

305

306



318

319

320

321

322

323

320 Figure 3: GM-entropy of activations after BatchNorm for each of the 53 layers in ResNet-50. The
 321 input image is the one shown in figure 7a d1.jpg. Vertical red dashed lines indicate skip-connection
 322 aggregation layers. It is noticeable that between the skip connections the GM-entropy many times
 323 increases to only decrease after the addition of the input. The vertical lines at layers 12, 14, 25, 27,
 44, 46 indicate when a reduction in spatial size occurs.

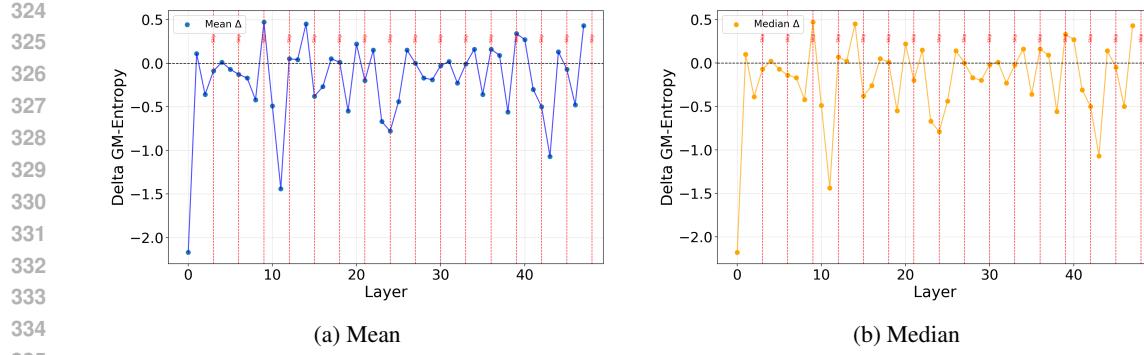


Figure 4: The change of GM-entropy between layers in ResNet-50 tested on 1000 images from ImageNet21K, selected randomly from 10 classes and 100 images per class. a. The y-axis is the mean of the changes (Mean Δ) over 1000 images vs the x-axis are each layer. b. The y-axis is the median of the changes (Median Δ) over 1000 images vs the x-axis are each layer. Clearly the changes alternate with positive changes between skip connections.

4 TOWARDS A CONFIDENCE MEASURE

We start by examining to what extent the GM-entropy of the last convolutional layer of ResNet-50 varies as the input is corrupted by blur or noise. Note that the last convolutional layer is the layer before the penultimate layer in the sense that after that there is a global average pooling followed by a fully connected layer to the output layer.

So, the first set of experiments we did was to consider effects on the GM-entropy by blurring an image (see figure 5 a.) or by adding noise to an image (see figure 5 b.) and examining the GM-entropy evolution (see figure 6). For all layers, the GM-entropy increases as the blur increases or as the

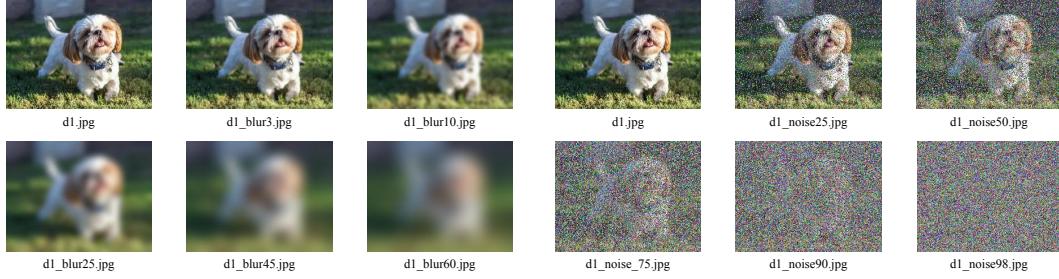
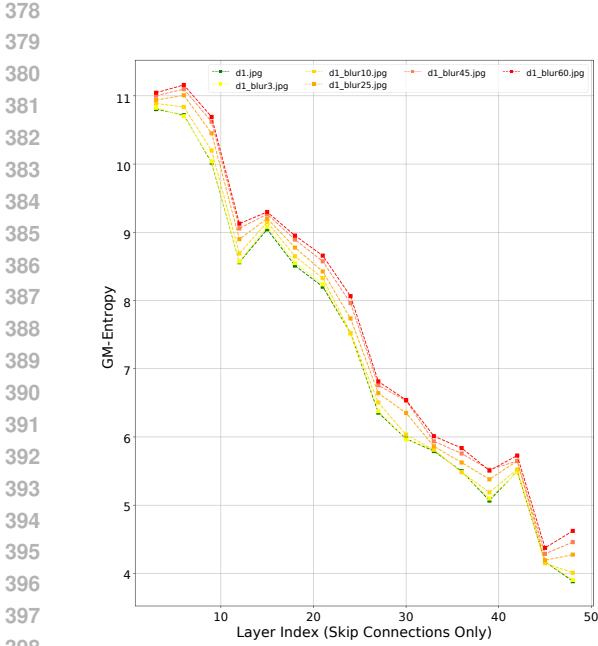
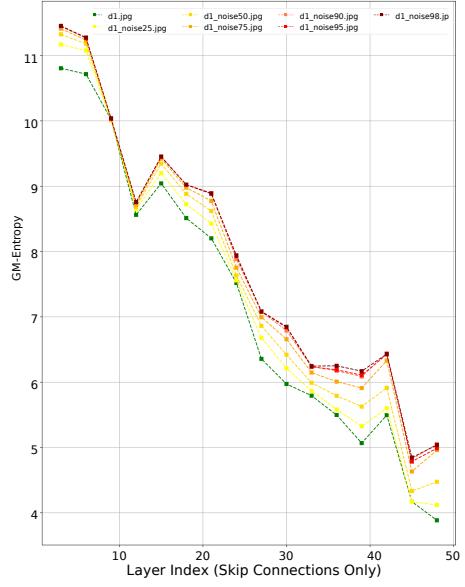


Figure 5: A picture of a Shih Tzu dog, d1.jpg, corrupted by various degrees of a. blur or b. noise.

noise increases. The last convolution layer, layer 48, is where the effect is most accentuated. Thus, we now explore further the hypothesis that the GM-entropy of layer 48 can provide a confidence measure focusing on the GM-entropy values at the last convolution layer. The next experiment we consider five pictures of dogs from ImageNet21K. We apply blur at various levels to all pictures and evaluate GM-entropy of layer 48 as a candidate for a confidence measure. We then compute the softmax value of the classification unit, often considered a confidence measure of the classification, and create a graph of these two confidence measures (see figure 7 a.). Similarly we apply noise at various levels and produce the same graph (see figure 7 b.) This comparison suggests that for these corruptions of images the GM-entropy outperforms the softmax measure.

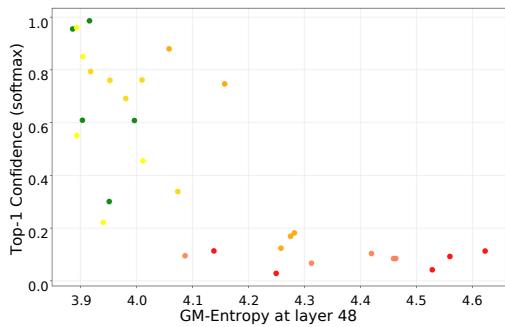


(a) Blurred images. Color encodes blur level on the python code applied-blur-to-image with blur levels (0, , 10, 25, 45, 60).

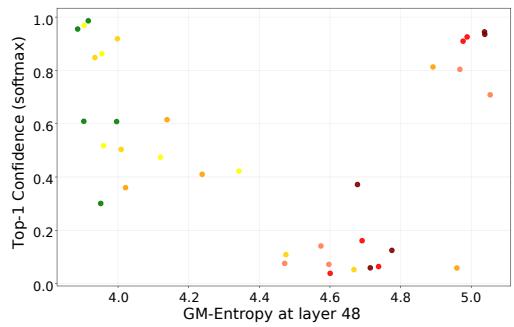


(b) Noisy images. Color encodes noise level (0%, 25%, 50%, 75%, 90%, 95%, 98%), where this numbers are the percentage of pixels replaced by random values.

Figure 6: GM-Entropy at skip-connection after BatchNorm layers. Clearly, the GM-entropy increases at all layers the more the images are corrupted either by blur or noise.



(a) Color coded blurred images displayed in a graph of confidence of Top-1 Classification vs GM-Entropy



(b) Color coded noisy images displayed in a graph of confidence of Top-1 Classification vs GM-Entropy

Figure 7: We considered five pictures of dogs from ImageNet21K, so each color in the graph above is shown five times. We applied to these five pictures the same levels and same color coded as figure 6 in a. blur at five levels and in b. noise at six levels. We then evaluate GM-entropy of layer 48 as a candidate for a confidence measure. For all these data, a. and b. we then compute the softmax value of the top-1 classification unit, often considered a confidence measure of the classification, and create a graph with axis of these two confidence measure (see figure 7 a. and b.). Clearly, the higher is the corruption of the image the higher is the GM-entropy while the top-1 softmax produces high values for large amount of corruption (and wrong classification results). The GM-entropy outperforms the softmax measure for these corruptions.

432 Table 2: We study the impact on the GM-entropy of blurring or adding noise to an image over 1000
 433 images from ImageNet21k (10 random classes and 100 images per class). In this case, we fixed at
 434 the third level of the previous experiments, namely the blur of python code applied-blur-to-image
 435 at blur level 25 and the noise level at 75%. We focused on layer 48 (the last convolution layer) and
 436 evaluated the GM-entropy. We compare with the GM-entropy of the original image.

437

Image Corruption Method	mean	median	std
original image	4.024042	4.017071	0.182951
blur level 25	4.221479	4.216569	0.137546
noise added 75%	4.614819	4.615324	0.199378

442

443

444 5 CONCLUSION

445

446 We introduced GM-entropy as a measure of information associated with each layer of a deep neural
 447 network for a given input. GM-entropy interprets each activation unit as contributing to an infor-
 448 mation probability: higher activations correspond to a greater likelihood of conveying information.
 449 Unlike the entropy of the joint distribution over a layer, GM-entropy is defined as a conditional en-
 450 tropy over the feature dimension, and thus it is sensitive to transformations that mix feature units
 451 rather than invariant to them.

452

453 On the empirical side, we evaluated the GM-entropy on ResNet-50, a widely used benchmark in
 454 large-scale image classification, and showed that it decreases consistently between layers. It was
 455 also observed that during the skip connection, before the aggregation of the input, the GM-entropy
 456 increased to later decrease with the input aggregation. This suggests that a possible explanation for
 457 the efficiency of skip connections is the management of the information. We conducted preliminary
 458 experiments on how GM-entropy responds to image corruption, such as increasing levels of blur or
 459 noise. Our results indicate that the last convolutional layer (layer 48) is particularly informative:
 460 GM-entropy rises systematically with the severity of corruption, effectively discriminating between
 461 degrees of degradation. In contrast, softmax confidence values can remain high even under heavy
 462 corruption, often leading to incorrect classifications.

461

462 In future work, we think that more studies of the GM-entropy for all other AI domains would be in-
 463 teresting, as well as a more thorough study of the GM-entropy values at the last layer of convolution
 464 may lead to better confidence measures for classification.

465

466 AUTHOR CONTRIBUTIONS

467

468 Equal contribution of the authors.

469

470 REPRODUCIBILITY

471

472 We plan to make the code available in GitHub and the dataset is already publicly available.

473

474 REFERENCES

475

476 H. B. Barlow. Single units and sensation: a neuron doctrine for perceptual psychology? *Per-
 477 ception*, 1(4):371–394, 1972. doi: 10.1068/p010371. URL [https://doi.org/10.1068/
 478 p010371](https://doi.org/10.1068/p010371).

478

479 Charles E. Connor. Friends and grandmothers. *Nature*, 435(7045):1036–1037, 2005. doi: 10.1038/
 480 4351036a. URL <https://doi.org/10.1038/4351036a>.

481

482 J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical
 483 Image Database. In *CVPR09*, 2009.

484

485 Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet21k (winter 2021
 486 release). URL <https://www.image-net.org/>.

486 David J. Field. What is the goal of sensory coding? *Neural Computation*, 6(4):559–601, 07 1994.
 487 ISSN 0899-7667. doi: 10.1162/neco.1994.6.4.559. URL <https://doi.org/10.1162/neco.1994.6.4.559>.

488

489 M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborová. Entropy
 490 and mutual information in models of deep neural networks. In *Advances in Neural Information
 491 Processing Systems*, 2018.

492

493 D. Geiger, A. Rudra, and L. Maloney. Features as sufficient statistics. In *Advances in Neural
 494 Information Processing Systems 10 (NIPS '97)*, pp. 794–800, 1997.

495

496 Charles G. Gross. Genealogy of the “grandmother cell”. *Neuroscientist*, 8(5):512–518, 2002. doi:
 497 10.1177/107385802237175. URL <https://doi.org/10.1177/107385802237175>.

498

499 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recog-
 500 nition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
 501 (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL <https://doi.org/10.1109/CVPR.2016.90>.

502

503 T. Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste. Sparsity in deep learning: Pruning
 504 and growth for efficient inference and training in neural networks. *Journal of Machine Learning
 505 Research*, 22(241):1–124, 2021. doi: 10.3929/ethz-b-000519202. URL <https://jmlr.org/papers/v22/21-0366.html>.

506

507 Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444,
 508 2015. doi: 10.1038/nature14539. URL <https://doi.org/10.1038/nature14539>.

509

510 G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural
 511 networks. In *Advances in Neural Information Processing Systems*, pp. 2924–2932, 2014.

512

513 B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a
 514 sparse code for natural images. *Nature*, 381:607–609, 1996.

515

516 Andrew M. Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D.
 517 Tracey, and David D. Cox. On the information bottleneck theory of deep learning. In *International
 518 Conference on Learning Representations (ICLR)*, 2018.

519

520 Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In
 521 *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2015.

522

523 Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In
 524 *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*,
 525 pp. 368–377, Urbana, Illinois, 1999.

526

527 Thomas Vetter, Anya Hurlbert, and Tomaso Poggio. View-based models of 3d object recognition:
 528 Invariance to imaging transformations. *Cerebral Cortex*, 5(3):261–269, 05 1995. ISSN 1047-
 529 3211. doi: 10.1093/cercor/5.3.261. URL <https://doi.org/10.1093/cercor/5.3.261>.

530

531

532

533

534

535

536

537

538

539