Characterizing Task Difficulty Using Spatial Entropy

Anonymous Author(s)
Affiliation
Address
email

Abstract

Our study highlights the use of *spatial entropy* as a means to characterize the difficulty of learning tasks. We show how the mutual information of class co-occurrences with regions in the feature space provides an informative curve profile to estimate the degree of difficulty in classification tasks. Empirical results demonstrate the feasibility of employing spatial entropy to quantify the quality of new representations in deep neural networks; results show how spatial entropy can act as a powerful meta-feature to enrich the current family of dataset characterizations.

1 Introduction

Understanding the relation between dataset properties and model performance is a central topic in meta-learning [1,2]; a key topic is the quantification of the difficulty of a learning task to understand the relation between model performance, model complexity, and data distributions. While there have been multiple studies advancing quantitative approaches to capture task difficulty [3-7], few studies have used such metrics in a meta-learning setting, to understand model performance, or to incorporate such metrics as meta-features.

In this paper, we follow an information-theoretical approach to task difficulty and show how incorporating the notion of space when computing class entropy sheds more light on the difficulty (or simplicity) of a learning task. The use of *spatial entropy* enables us to differentiate tasks with marked differences in difficulty that otherwise would have remained alike. Our study shows how computing entropy on a joint space that combines spatial and class-distribution information leads to a powerful tool to assess the quality of new representations.

Our experiments show how each layer in a deep neural network evolves as a function of our task-difficulty metric. Results point to utilizing spatial entropy as a measure of task complexity over the training period of a neural network. In section 3 we introduce spatial entropy as a measure of task difficulty. In section 4 we describe our experimental design and report our results. Finally, we conclude with our conclusions and offer future directions to explore spatial entropy in-depth.

2 Preliminaries

We assume a training set, $T = \{(X_i, Y_i)\}_{i=1}^{N}$, where $X = (x_1, x_2, \cdots, x_P)$ is a an instance (vector) of the input space $\mathcal{X}$, and $Y \in \{y_1, y_2, ..., y_K\}$ is an instance (nominal or categorical value) of the output space $\mathcal{Y}$. We assume T contains i.i.d. examples from a fixed but unknown joint probability distribution, $P(X, Y)$, in $\mathcal{X} \times \mathcal{Y}$. The output of the learning algorithm is a function $f_\theta(X)$, $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$, and $f_\theta \in \mathcal{F}$. The goal is to search for the function that minimizes the expectation of a loss $L(Y, f(X|\theta))$, a.k.a. the risk, $R(\theta, P(X, Y)) = E_{\sim P}[L(Y, f(X|\theta))$. Here we employ the zero-one loss function: $L(Y, f(X|\theta)) = I(Y \neq f(X|\theta))$, where $I(\cdot)$ is an indicator function, and Y is the class of X.

Submitted to the 5th Workshop on Meta-Learning at NeurIPS 2021, Sydney, Australia. Do not distribute.

We are primarily interested in the computation of class entropy $H(Y)$; we advocate the use of spatial information (Section 3) to provide a more clear picture of the role of neighborhoods in the instance space. We follow Shannon's definition of entropy over a probability mass function $[p(y_1), p(y_2), ..., p(y_K)]^T : H(Y) = \sum_{k=1}^{K} p(y_k) \log\left(\frac{1}{p(y_k)}\right)$.

3 Spatial Entropy and Task Difficulty

Initial studies characterized task difficulty as a function of the number of alternating peaks in the class distribution across the input space [3]. Other approaches include computing the cross-entropy between the output labels and a "trivial" output that always shows a constant value [4]; and defining task difficulty as a function of meta-features capturing the geometrical complexity of the class boundary [5]. Besides the design of task-difficulty metrics, other studies have tried to explain theoretically the complexity of learning conditioned on properties of the learning algorithm [6,7].

In contrast to previous work, our metric for task difficulty approximates the entropy of Y over different neighborhoods, following closely the definition of spatial entropy proposed by Altieri [8,9]. Specifically, for a neighborhood $\mathcal{N}(X^*)$ centered on point X^*, we are interested in the entropy of the distribution of Y, $H(Y)$, in $\mathcal{N}(X^*)$.

To introduce the notion of space, we define concentric hyperspheres around X^* of varying width. Each hypersphere refers to a region in the input space with some associated probability density. We refer to the space variable as W and to the space between hyperspheres as $w_1, w_2, ..., w_M$. Each region w_m has an associated probability $P(w_m)$, estimated as the fraction of training examples falling in that region. Figure 1 illustrates these ideas. Note that, for the first hypersphere, such region is the entire space filled by the sphere; the second region is the space between the border of the second hypersphere and the border of the first hypersphere; regions are then mutually exclusive.

Rather than working with Y directly, and for efficiency, we define a new variable Z referring to the possible combination of values of Y for a pair of neighbor examples lying close to each other in the input space (within a hypersphere). For example, in the two-class problem where $Y \in \{0, 1\}$, the new variable Z is defined as $Z \in \{(0, 0), (0, 1), (1, 1)\}$, with an associated probability mass $[p(z_1), p(z_2), ..., p(z_L)]^T$.

Our focus has now been redirected to the entropy of Z, $H(Z)$, with space, W, playing an important role. Spatial entropy is defined as

$$H(Z) = I(Z; W) + E[H(Z|W)] \tag{1}$$

The first term, I(Z;W), is the mutual information between Z and W. Since, $P(W, Z) = P(W)P(Z|W)$, mutual information can be defined defined as

$$I(Z; W) = \sum_{m=1}^{M} P(w_m) D_{\text{KL}}(P(Z|W)||P(Z)) = \sum_{m=1}^{M} P(w_m) \left[\sum_{l=1}^{L} P(z_l|w_m) \log\left(\frac{P(z_l|w_m)}{P(z_l)}\right)\right] \tag{2}$$

The right term in brackets is the relative entropy (Kullback-Leibler divergence) of $P(Z|W)$ and $P(Z)$, named *spatial partial information*.

The second term in equation 1, $E[H(Z|W)]$ is named *spatial global residual entropy*;

$$E[H(Z|W)] = \sum_{m=1}^{M} P(w_m) H(Z|w_m) = \sum_{m=1}^{M} P(w_m) \left[\sum_{l=1}^{L} P(z_l|w_m) \log\left(\frac{1}{P(z_l|w_m)}\right)\right] \tag{3}$$

where the term in brackets is named *spatial partial residual entropies*; it quantifies the contribution of each neighborhood to the residual entropy of Z.

Equation 1 can be rewritten as an expectation of the sum of spatial partial information and spatial global residual entropies:

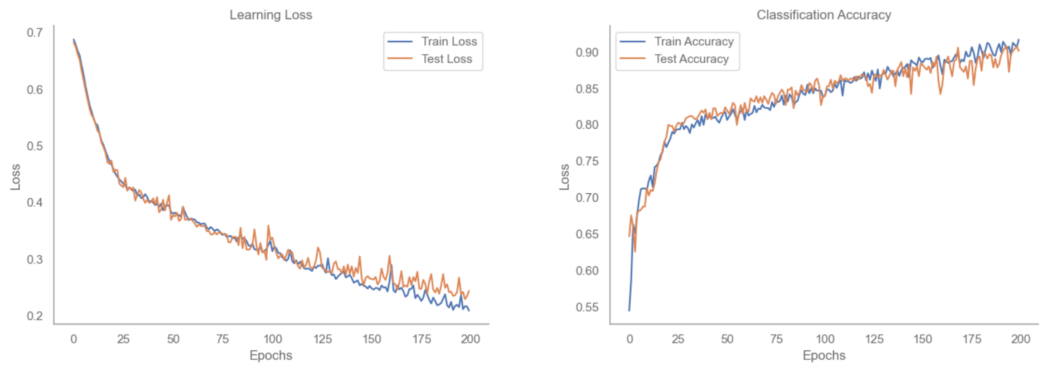

Figure 1: *Left:* Average loss recorded during training and testing phases over 200 epochs. *Right:* Classification accuracy of the neural network over 200 epochs.

$$H(Z) = \sum_{m=1}^{M} P(w_m) \left[D_{\text{KL}}(P(Z|W)||P(Z)) + H(Z|w_m) \right] \tag{4}$$

The formulation above separates the entropy of class co-occurrences between two nearest neighbors within a specific region (space between two concentric hyperspheres) across all neighborhoods. The first term shows the contribution of space; the second term shows the residual entropy after the effect of space is removed. Although other definitions for spatial entropy exist [10], the definition above has the advantage of decoupling the contribution of space and residual entropy both globally and locally (per window).

4 Spatial Entropy and Metalearning

We propose using spatial entropy as a direct measure of task difficulty. In the context of meta-learning, the idea is twofold: spatial entropy can be used as a meta-feature to characterize datasets as a prelude to the construction of a meta-model [1,2]. In addition, spatial entropy can be used as meta-knowledge in transfer learning, to improve learning performance across tasks. Here we simply point to the value of spatial entropy to capture the information contained in the class distribution over the input space.

An example of previous work connecting spatial entropy with supervised learning tools lies in image analysis: rather than relying on histograms alone, spatial entropy brings into the analysis spatial information associated with pixels; this can drastically change the amount of image information. An example is hyperspectral image analysis, where spatial entropy captures the role of space along with multiple hyperspectral bands [11]. A similar study incorporates spatial entropy for the analysis of geographical data [12], specifically on agricultural data. In both studies, the definition of entropy is modified by adding weights to the additive class-entropy terms based on the ratio of the intra- and extra-distance among training examples of similar and different classes respectively; the ratio is computed based on spatial coordinates associated to each example. Different from previous work, we explore the use of spatial entropy in the context of dataset characterization, as a tool for meta-learning.

4.1 Experiments

To demonstrate the behavior of spatial entropy during learning, we designed a set of experiments on artificial datasets that vary in complexity. To to better understand the relationship between learning and the complexity of a task, we compute various spatial metrics that will help us better understand this relationship.

We train a 3-hidden layer neural network on synthetically generated binary classification dataset $T = \{(X_i, Y_i)\}_{i=1}^{N}$ consisting of $N = 10,000$ data points and $P = 10$ numerical features. To to mimic a real-world problem, we introduce label noise $\epsilon = 0.02$. The neural network is trained using stochastic gradient descent over 200 epochs with a batch size of 128. Loss is computed using the binary cross-entropy loss. Non-linear transformations are achieved by employing the ReLU activation. A sigmoid function is applied to the final layer to produce logits for the binary cross-entropy.

Before training, we reduce the dimensionality of our data from 10 dimensions to 2 learned t-SNE components and compute the initial spatial entropy. Next, at each epoch, we transform our input using the learned representation at the penultimate hidden layer ℓ_3 and compute spatial entropy $H(Z)$ on the new representation $\zeta(X_i)$. To better understand the learning process and the dynamics of spatial entropy, we extract the decomposed spatial metrics such as partial mutual information $P(Z;W)$ and partial residual entropy $H(Z|W)$, proportional spatial mutual information $I_{\text{prop}}(Z;W)$, and relative mutual information $I_{\text{rel}(Z;W)}$ and relative residual entropy $H_{\text{rel}}(Z|W)$. Each of these spatial metrics will help us understand the learning process and the complexity of the task at hand.

Spatial entropy requires a set of distance classes $w_i \in W$ over the data. We define our own range of distance classes to be $w_1 = [0, 1]$, $w_2 =]1, 2]$, $w_3 =]2, 3]$, $w_4 =]3, 4]$, $w_5 =]4, 5]$. In order to make sure our data points fall into these distances, we scale our data appropriately. Distance classes w_1 and w_2 correspond to 4-nearst neighborhood and 8-nearest neighborhood classes [13].

After training our neural network for 200 epochs, we report all the metrics and demonstrate how spatial entropy can assist in better understanding the learning process as well as understand the complexity of the task.

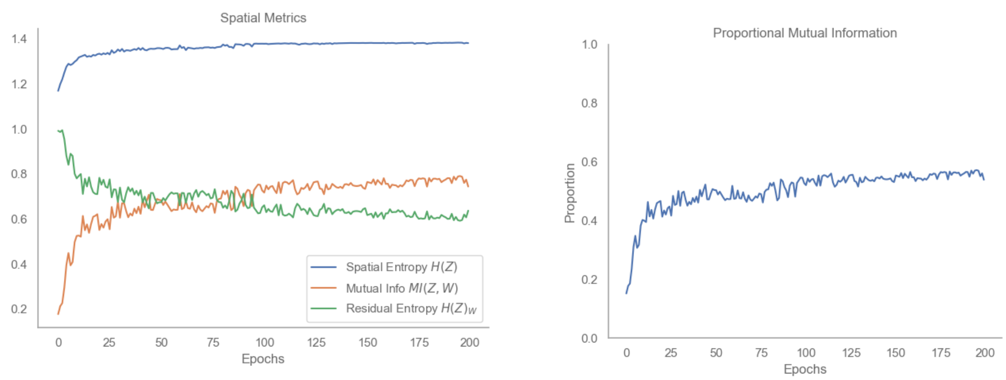

Figure 2: *Left:* Average loss recorded during training and testing phases over 200 epochs. *Right:* Classification accuracy of the neural network over 200 epochs.

4.2 Results

Spatial entropy metrics extracted from the original dataset and the learned representation are shown in Figure 3. The inverse relationship between spatial global residual entropy and spatial global mutual information can be seen on the left plot. Due to the additivity property, summing those quantities (orange and red) together produce the spatial entropy quantity (blue). As the neural network learns a better representation of the classification problem, spatial global mutual information increases while spatial global residual entropy decreases. This trend confirms the role of space in the final learned representation and the complexity of the problem decreases as the separation between the two classes becomes more evident.

To better understand the impact of the two components onto the entropy we convert partial mutual information and partial residual entropy, to sum up to one. We can identify what component contributed most in entropy. At each distance class w_i, we can see whether the heterogeneity is explained by the role of space (mutual information) or some other sources (residual entropy). Figure 3 highlights the decomposed values at each distance class w_i for the original data and the final learned representation. As evident by the bar chart for the original dataset (left), space plays no role in explaining the heterogeneity in the entropy. Conversely, for smaller distance classes w_1, w_2, and w_3, space contributes almost a majority towards the heterogeneity on the learned representation of the neural network.

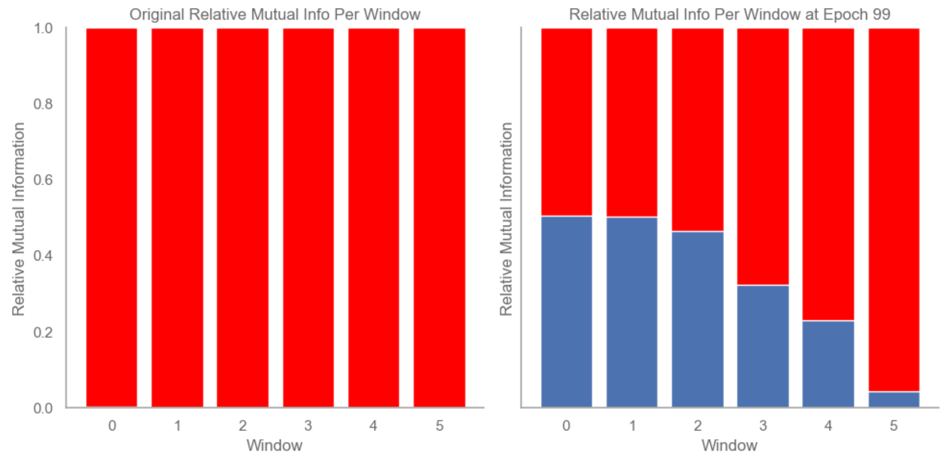

Figure 3: Relative mutual information (blue) and relative residual entropy (red). *Left:* relative metrics per distance class on original dataset. *Right:* relative metrics per distance class on learned representation after 200 epochs. Partial mutual information plays a larger role in the learned representation.

Figure 4 shows the visual representation of our dataset at two stages: original data and final learned representation. The neural network learns a representation that allows for a clean separation of the two classes.

The effect that space has on spatial entropy is highlighted by the proportional mutual information plot in Figure 2. Proportional mutual information is defined as

$$I_{\text{prop}}(Z;W) = \frac{I(Z;W)}{H(Z)} \tag{5}$$

and it states that the contribution of space in the entropy of Z is a proportion of the marginal entropy bounded on $[0,1]$. Datasets with different spatial contributions but with different probability mass p_Z of co-occurrences share the same spatial entropy $H(Z)$ but will have different contributions from the two components.

As our neural network continues learning, the role that space plays in the new representation of spatial entropy increases which translates to a decreased task complexity. Figure 1 highlights the performance of the neural network over 200 epochs. Figure 1 and Figure 2 can be compared side-by-side that further support our findings.

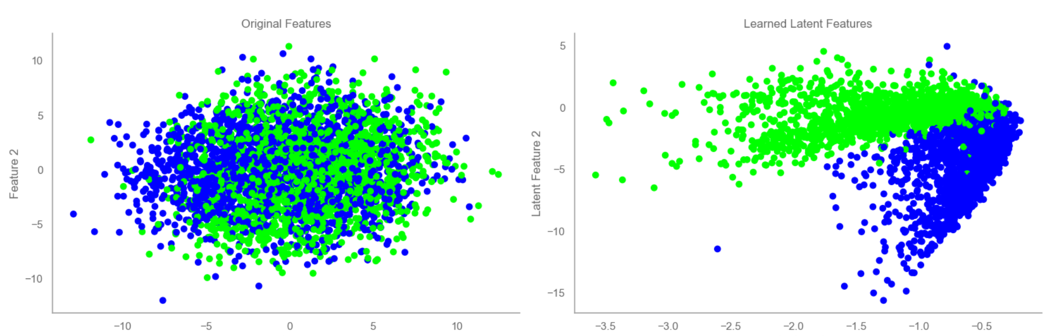

Figure 4: *Left:* t-SNE projection of the original dataset. *Right:* t-SNE projection of the learned representation after 200 epochs. The neural network learns a representation that allows for a simple decision boundary.

5 Conclusions

In this paper, we study the use of spatial entropy as a novel meta-feature to evaluate the complexity of a task. We closely follow how the spatial characteristics of a classification problem change throughout the learning process of a neural network. Decomposing the spatial entropy into its components allows us to better understand the role of space as a source of heterogeneity on entropy. Our preliminary experiments demonstrate that heterogeneity from other sources other than space is highly prevalent in farther distance classes w_i while mutual information is highest in closer distance classes. This highlights the role of space in entropy. We conclude that as the neural network learns a better representation of the input, space contributes in entropy.

5.1 Future Work

Spatial entropy provides a novel insight into measuring task complexity and understanding how learning and task complexity interact throughout the process. Future work could see exploring spatial entropy over image data as it can offer interesting insights on pixel density and learning features of a convolutional neural network. It can prove to be one of the new groups of meta-features specifically for image data. Another extension would see spatial entropy measured during learning of a multi-class task where overlap of classes is prominent. It can help understand how a neural network decomposes a difficult multi-class problem and what role spatial entropy plays. Perhaps maximizing proportional mutual information as an objective function can offer new insights and improved performance on certain tasks.

References

[1] Brazdil, P. & Giraud-Carrier, C. & Soares, C. & Vilalta R. (2008). *Metalearning: Applications to Data Mining.* Springer Verlag. ISBN: 978-3-540-73262-4

[2] Vilalta R., Giraud-Carrier C., Brazdil P. (2005) Meta-Learning: Concepts and Techniques. *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers.* Oded Maimon and Lior Rokach, Editors. Springer Publishers.

[3] Rendell, L. & Cho, H. (1990). Empirical learning as a function of concept character. Machine Learning **5**, 267-–298. https://doi.org/10.1007/BF00117106

[4] Tran, A. & Nguyen, C. V. & Hassner, T. (2019). Transferability and Hardness of Supervised Classification Tasks. *Conference Information*, pp. xxx–yyy. Editorial.

[5] Ho, T. K. & Basu, M. (2002). Complexity Measures of Supervised Classification Problems. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(3).

[6] Song, L. & Vempala, S.S. & Wilmes, J. & Xie, B. (2017). On the Complexity of Learning Neural Networks. *Proceedings of Neural Information Processing Systems NeurIPS-17.*

[7] Cesa-Bianchi, N. & Mansour, Y. & Shamir, O. (2015). On the Complexity of Learning with Kernels. *Proceedings of The 28th Conference on Learning Theory, Proceedings of Machine Learning Research* 40:297-325.

[8] Altieri, L. & Cocchi, D. & Roli, G. (2018). A new approach to spatial entropy measures. *Environ Ecol Stat* **25**:95–110. https://doi.org/10.1007/s10651-017-0383-1.

[9] Altieri, L. & Cocchi, D. & Roli, G. (2019). Advances in spatial entropy measures. *Stoch Environ Res Risk Assess* **33**:1223–1240. https://doi.org/10.1007/s00477-019-01686-y

[10] Batty, M. & Morphet, R. & Masucci, P. et al. (2014). Entropy, complexity, and spatial information. *J Geogr Syst* **16**:363–385. https://doi.org/10.1007/s10109-014-0202-2

[11] Wang, B. & Wang, X. & Chen, Z. (2012). Spatial Entropy Based Mutual Information in Hyperspectral Band Selection for Supervised Learning. *International Journal of Numerical Analysis and Modeling* **9**(2):181-192.

[12] Li, X. & Claramunt, C. (2006). A Spatial Entropy-Based Decision Tree for Classification of Geographical Information. *Transactions in GIS.* **10**:451-467.

[13] Cressie, N. A. (1993). Statistics for spatial data (No. 519.5 C74).