Supplementary Material: On the combined effect of class imbalance and concept complexity in deep learning

Kushankur Ghosh *Department of Computing Science*
University of Alberta
Edmonton, Canada
kushanku@ualberta.ca Colin Bellinger *Digital Technologies*
National Research Council of Canada
Ottawa Canada
colin.bellinger@nrc-cnrc.gc.ca Roberto Corizzo *Department of Computer Science*
American University
Washington, DC, USA
rcorizzo@american.edu Bartosz Krawczyk *Department of Computer Science*
Virginia Commonwealth University
Richmond, VA, USA
bkrawczyk@vcu.edu Nathalie Japkowicz *Department of Computer Science*
American University
Washington, DC, USA
japkowic@american.edu

Abstract

Structural concept complexity, class overlap, and data scarcity are some of the most important factors influencing the performance of classifiers under class imbalance conditions. When these effects were uncovered in the early 2000s, understandably, the classifiers on which they were demonstrated belonged to the classical rather than Deep Learning categories of approaches. As Deep Learning is gaining ground over classical machine learning and is beginning to be used in critical applied settings, it is important to assess systematically how well they respond to the kind of challenges their classical counterparts have struggled with in the past two decades. The purpose of this paper is to study the behavior of deep learning systems in settings that have previously been deemed challenging to classical machine learning systems to find out whether the depth of the systems is an asset in such settings. The results in both artificial and real-world image datasets (MNIST Fashion, CIFAR-10) show that these settings remain mostly challenging for Deep Learning systems and that deeper architectures seem to help with structural concept complexity but not with overlap challenges in simple artificial domains. Data scarcity is not overcome by deeper layers, either. In the real-world image domains, where overfitting is a greater concern than in the artificial domains, the advantage of deeper architectures is less obvious: while it is observed in certain cases, it is quickly cancelled as models get deeper and perform worse than their shallower counterparts.

I. Introduction

This document provides supplementary figures for the IEEE Big Data 2021 paper entitled: On the combined effect of class imbalance and concept complexity in deep learning.

II. Artificial Domains

A. Backbone

Figure 1 presents the 1-dimensional backbone domain used in our experiments. 125 classification problems were created of this nature. At problem is uni-dimensional with inputs in the [0, 1] range associated with class 1 (+) or 0 (-). The [0, 1] input range is divided into sub-intervals of the same size, each associated with class value 0 or 1. Contiguous intervals have opposite class values. The complexity level, c, can take values from 1 to 5. Depending on its value, different numbers of sub-intervals are created. An example of a backbone model is shown in Figure 1.

B. 5-Dimensional Gaussian

Figure 2 provides a sample of the 5-dimensional Gaussian binary classification domains used in the our experiments. The 10 classification domains of this nature were created and used. The domains were created with increasing levels of overlap from 1 to 10. Overlap 1 (representing the highest level of overlap) the mean for each class is the same, at 0.5. It is then incremented by 1, step-wise, up to nine times to obtain the 9 other distributions.

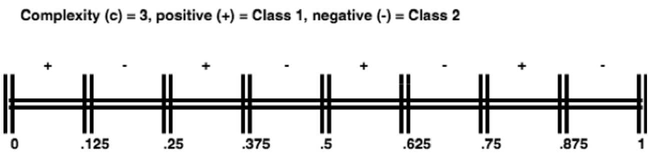
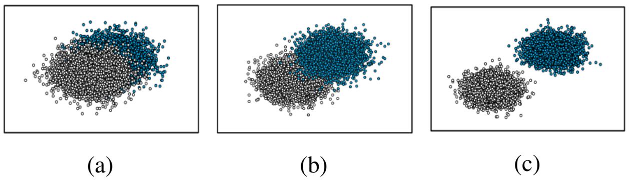

Complexity (c) = 3, positive (+) = Class 1, negative (-) = Class 2
+
-
+
-
+
-
+
-
0
.125
.25
.375
.5
.625
.75
.875
1

Fig. (1) Domain backbone of Complexity 3. In this one-dimensional family of domains, the complexity of the task increases as the number of alternating sub-concepts of each class increases.



(a) (b) (c)

Fig. (2) Instances of Overlapped Domain: (a) Overlap 3, (b) Overlap 5, and (c) Overlap 9

C. Overlapping Backbone Gaussian

Figure 3 presents the overlapping backbone domain. This domains combines the ideas from the backbone domain and the 5-dimensional Gaussian domain. In these domains, we incorporate overlaps between each subconcept of the backbone by generating Gaussian distributions centered in the middle of each sub-interval. We concentrate on 5 different variation of overlaps (v1, v2, v3, v4, v5) represented in Figure 3 that is arranged in order, starting from low overlap (v1) to high overlap (v5). Each of these distributions are divided into 5 levels of b (as we did previously) by generating $\left(\frac{1250}{\frac{32}{2^b}}\right)$ examples in each minority subconcept and a constant number of 1250 examples in the majority subconcepts. We constructed a testing set by following the balanced testing approach used previously and generating 2000 instances for each class (and, more specifically, 1000 instances in each subconcept).

D. Real-World Domains

Binary domains were selected using a combination of visual inspection of binary T-SNE plots and cross-validation experiments.The aim of the selection was to identify five binary domains with increasing levels of complexity. Binary class pairs from MNIST Fashion and CIFAR-10 were selected for use in the experiments. The class combinations were selected such that the problems because increasingly complex. The selected class combinations are presented in Figure 4 and 5 in the form of T-SNE plots.

III. Methodology

A. The equivalence of macro and weighted averages in the Balanced Setting experiments

To formulate an explanation, considering τ^+, f^+, τ^-, and f^- as true positive, false positive, true negative, and false negative we calculate the sensitivity, S, and specificity, Sp, for class 0 $\left(S^0, Sp^0\right)$ and class 1 $\left(S^1, Sp^1\right)$ as;

$$S^0 = \frac{\tau^-}{\tau^- + f^+} S^1 = \frac{\tau^+}{\tau^+ + f^-} \tag{1}$$

$$Sp^0 = \frac{\tau^+}{\tau^+ + f^-} Sp^1 = \frac{\tau^-}{\tau^- + f^+} \tag{2}$$

Based on the obtained expressions, the Geometric Mean for class 0 $\left(G^0\right)$ and class 1 $\left(G^1\right)$ can be mathematically defined as;

$$G^0 = \sqrt{\frac{\tau^- \cdot \tau^+}{(\tau^- + f^+) \cdot (\tau^+ + f^-)}} \tag{3}$$

$$G^1 = \sqrt{\frac{\tau^+ \cdot \tau^-}{(\tau^+ + f^-) \cdot (\tau^- + f^+)}} \tag{4}$$

Therefore, the macro average Geometric Mean can be written as;

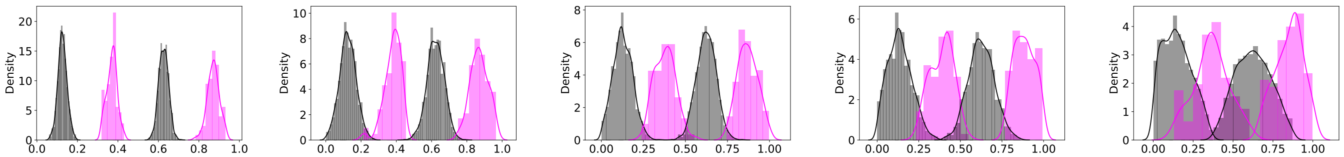


Fig. (3)   Plots of Overlapped Gaussian Distributions on Backbone sorted according to overlap.

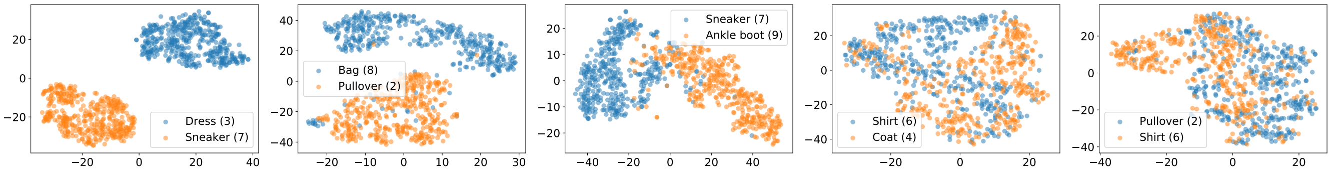


Fig. (4)   T-SNE Plots of Binary MNIST Fashion datasets sorted from the least to the most complex.

$$G^{mac} = \frac{1}{2} \cdot G^0 + \frac{1}{2} \cdot G^1 \tag{5}$$

The weighted average involves a calculation concentrating on the number of examples present in the training set for each class. If we consider the training set as X, then, X_0 represents the total number of instances that belong to class 0 and X_1 is that of class 1, such that, $|X| = X_0 + X_1$. The weighted average can then be denoted as;

$$G^w = \frac{X_0}{|X|} \cdot G^0 + \frac{X_1}{|X|} \cdot G^1 \tag{6}$$

On careful analysis, we find that the expression G^0 and G^1 are similar. Therefore, G^{mac} will be equal to $G^{0,1}$ such that $G^{0,1} = G^0 = G^1$. Similarly, we can also conclude G^w as $G^{0,1}$. Hence, we are getting $G^{mac} = G^w$.

IV. Results

A. *Stratified CV Experiments*

In this section, we present the performance of deep (with 5 hidden layers) and shallow (with 1 hidden layer) MLP models that we obtained after applying 10 Fold Stratified Cross Validation on the Backbone and Overlapped Domains. Figure 6 and Figure 7 illustrates the results obtained on the Backbone framework of different sizes whereas, Figure 8 shows the results obtained on the Overlapped Datasets.

B. *Results on Image Domains*

TABLE (I)   Distribution of rank of model performance on MNIST Fashion as a function of depth. These results indicated that on average there is a slight preference for shallower model on the highly imbalanced data.

	Sum of Rank				
Model Depth	Balanced	0.3	0.15	0.05	0.025
1	12.0	15.0	13.0	21.0	13.0
2	9.0	10.0	9.0	17.0	10.0
3	7.0	6.0	14.0	9.0	12.0
4	18.0	16.0	12.0	13.0	17.0
5	22.0	18.0	20.0	9.0	22.0

C. *Model Embeddings*

In order to understand how they are impacted by concept complexity, class overlap and imbalance on the learned models, Figure 9 plots to model embeddings for the CIFAR-10 and MNIST Fashion data. To advance this analysis, we focus on an ‘easy’ classification problem and a ‘hard’ classification problem.

References

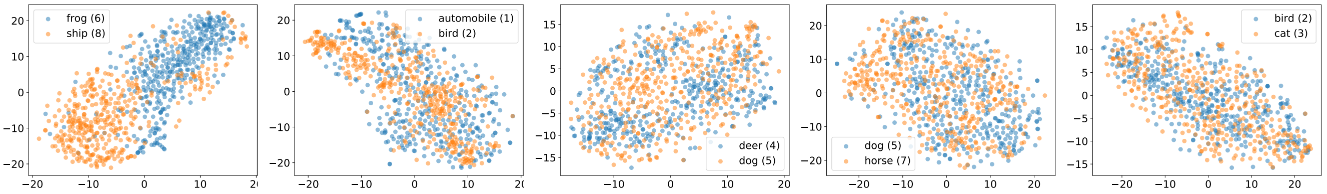
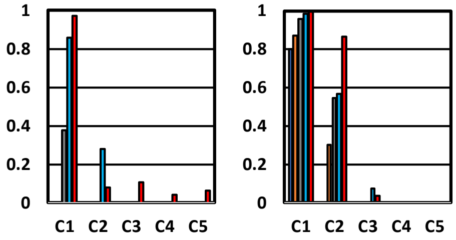


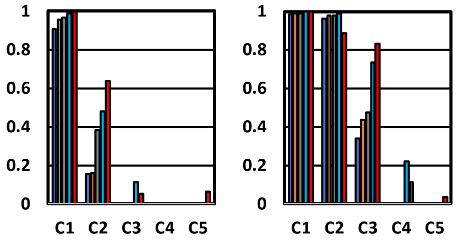
Fig. (5) T-SNE Plots of Binary CIFAR10 datasets sorted from the least to the most complex.



(a) Model 1

(b) Model 5

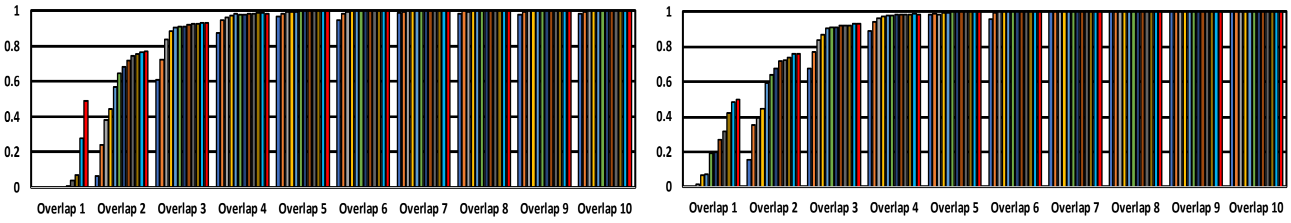
Fig. (6) MLP generated Macro G-Mean Scores by doing 10-fold Stratified Cross Validation for Size 1: (a) Model 1 and (b) Model 5.



(a) Model 1

(b) Model 5

Fig. (7) MLP generated Macro G-Mean Scores by doing 10-fold Stratified Cross Validation for Size 5: (a) Model 1 and (b) Model 5.



(a)

(b)

Fig. (8) MLP generated Macro G-Mean Scores by doing 10-Fold Stratified Cross Validation on the Overlapped Datasets: (a) Model 1 and (b) Model 5.

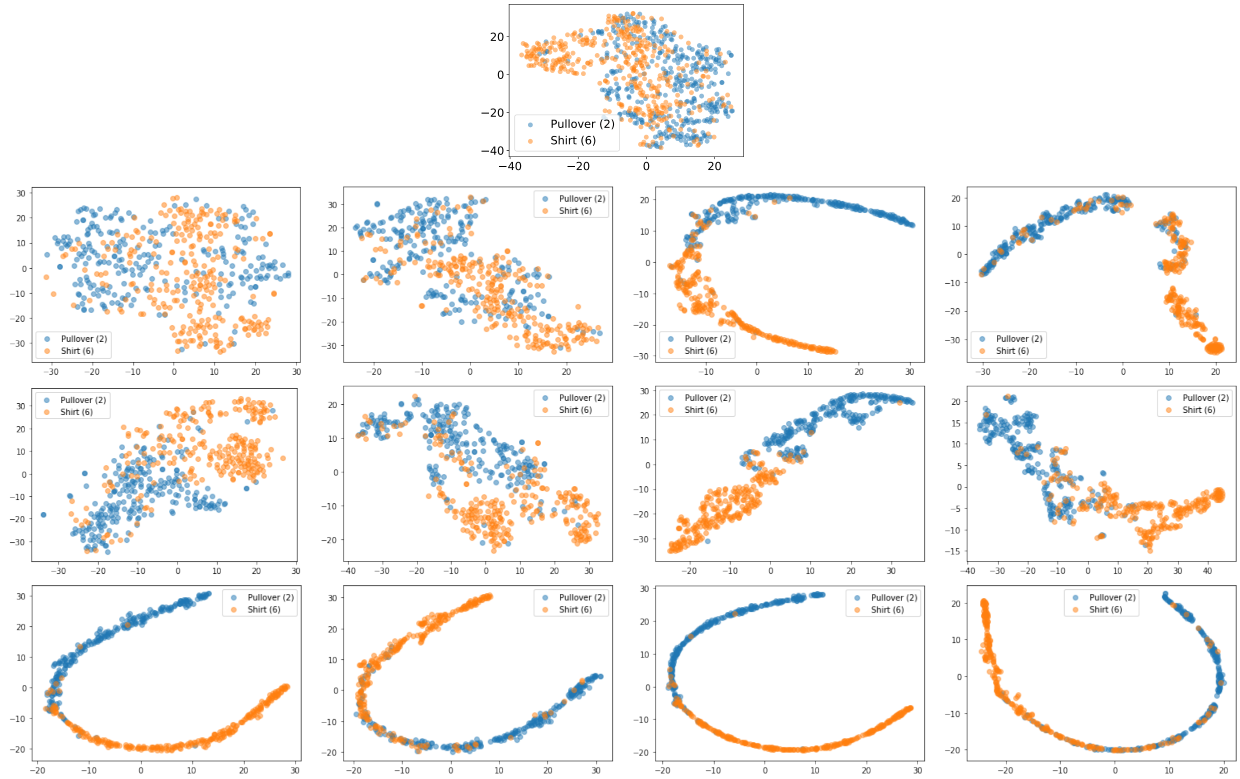


Fig. (9) These plots illustrate the representation learned on the MNIST Fashion for classes 2 versus 6. The plot on the top row shows the T-SNE embedding of the original data. This reveals that the classification problem is complex due to the presence of class overlap and sub-concepts. In the subsequent rows, the first and second columns show the T-SNE plots for the representations learned after the last CNN block on the balanced (first column) and imbalanced (second column) data. The third and fourth columns show the T-SNE plots for the representations learned after the last fully connected layers for balanced (third column) and imbalanced (fourth column). Row-wise, the plot represent the results of CNNs with 1, 3, and 5 CNN blocks. These plots illustrate that adding more CNN blocks on the balanced data improves separability of the classes. However, whilst 3 blocks in the imbalanced case appears to slightly improve separability, 5 blocks exacerbates the class overlap.