
Supplementary material

Anonymous Author(s)

Affiliation

Address

email

1 Definitions

First we recall our geometric definitions of neural network and layer.

Definition 1 (Neural Network). A neural network is a sequence of \mathcal{C}^1 maps Λ_i between manifolds of the form:

$$M_0 \xrightarrow{\Lambda_1} M_1 \xrightarrow{\Lambda_2} M_2 \xrightarrow{\Lambda_4} \dots \xrightarrow{\Lambda_{n-1}} M_{n-1} \xrightarrow{\Lambda_n} M_n \quad (1)$$

We call M_0 the input manifold and M_n the output manifold. All the other manifolds of the sequence are called representation manifolds. The maps Λ_i are the layers of the neural network. We denote with $\mathcal{N}_{(i)} = \Lambda_n \circ \dots \circ \Lambda_i : M_i \rightarrow M_n$ the mapping from the i -th representation layer to the output layer.

Definition 2 (Smooth layer). A map $\Lambda_i : M_{i-1} \rightarrow M_i$ is called a smooth layer if it is the restriction to M_{i-1} of a function $\bar{\Lambda}^{(i)}(x) : \mathbb{R}^{d_{i-1}} \rightarrow \mathbb{R}^{d_i}$ of the form

$$\bar{\Lambda}_\alpha^{(i)}(x) = F_\alpha^{(i)} \left(\sum_\beta A_{\alpha\beta}^{(i)} x_\beta + b_\alpha^{(i)} \right) \quad (2)$$

for $i = 1, \dots, n$, $x \in \mathbb{R}^{d_i}$, $b^{(i)} \in \mathbb{R}^{d_i}$ and $A^{(i)} \in \mathbb{R}^{d_i \times d_{i-1}}$, with $F^{(i)} : \mathbb{R}^{d_i} \rightarrow \mathbb{R}^{d_i}$ a diffeomorphism.

We also need some standard definitions in differential geometry [4].

Definition 3 (Submersion). Let $f : M \rightarrow N$ be a smooth map between manifolds. Then f is a submersion if, in any chart, the Jacobian J_f has rank $\dim(N)$.

Definition 4 (Embedding). Let $f : M \rightarrow N$ be a smooth map between manifolds. f is an embedding if its differential is everywhere injective and if it is an homeomorphism with its image. In other words, f is a diffeomorphism with its image.

Definition 5 (Distribution). A distribution \mathcal{D} of dimension k over a m -dimensional manifold M is a collection of k smooth vector fields v_1, \dots, v_k such that $(v_1)_p, \dots, (v_k)_p$ form a basis of a vector subspace of dimension k in $T_p M$ for every $p \in M$.

Definition 6 (Integrable distribution). A distribution \mathcal{D} of dimension k is an integrable distribution if there exist a manifold M of dimension $m \geq k$ such that the collection of k smooth vector fields v_1, \dots, v_k are generating a vector space of dimension k over $T_p M$ for all $p \in M$.

Definition 7 (Trivial fiber bundle). A trivial fiber bundle is a structure (E, B, π, F) , where E, B and F are topological spaces with $E = B \times F$ and the map $\pi : E \rightarrow B$ is the projection of $B \times F$ on B . The space F is called typical fiber. In the case F is a vector space, then (E, B, π, F) is called a trivial vector bundle.

Definition 8 (Vertical and horizontal spaces). Let (E, M, π, F) be a vector bundle over a manifold M . Then the vertical space $\mathcal{V}_p E$ at $p \in E$ is the vector space $\mathcal{V}_p E = \text{Ker}(d_p \pi) \subset T_p E$. The horizontal space $\mathcal{H}_p E$ is a choice of a subspace of $T_p E$ such that $T_p E = \mathcal{V}_p E \oplus \mathcal{H}_p E$. The spaces $\mathcal{V} E := \sqcup_{p \in E} \mathcal{V}_p E$ and $\mathcal{H} E := \sqcup_{p \in E} \mathcal{H}_p E$ are two bundles called vertical and horizontal bundles respectively.

2 Hypotheses

In the following lemmas and propositions we always assume the following hypotheses to hold true.

Assumption 1. The manifolds M_i are open and path-connected sets of dimension $\dim M_i = d_i$.

Assumption 2. The sequence of maps (1) satisfies the following properties:

1) If $\dim(M_{i-1}) \leq \dim(M_i)$ the map $\Lambda_i : M_{i-1} \rightarrow M_i$ is a smooth embedding.

2) If $\dim(M_{i-1}) > \dim(M_i)$ the map $\Lambda_i : M_{i-1} \rightarrow M_i$ is a smooth submersion.

Assumption 3. The manifold M_n is equipped with the structure of Riemannian manifold, with metric $g^{(n)}$.

Assumption 4. We assume that the manifolds M_i are diffeomorphic to \mathbb{R}^{d_i} for some $d_1, \dots, d_n \in \mathbb{N}$.

Assumption 5. The matrices of weights in the maps Λ_i , $i = 1, \dots, n$, as per in Definition 2 are of full rank.

3 Proof of the propositions

Proposition 1. Let $\gamma : [0, 1] \rightarrow M_i$ be a piecewise \mathcal{C}^1 curve. Let $k \in \{i, i+1, \dots, n\}$ and consider the curve $\gamma_k = \Lambda_k \circ \dots \circ \Lambda_i \circ \gamma$ on M_k . Then $Pl_i(\gamma) = Pl_k(\gamma_k)$

Proof. It is enough to notice that $\gamma_k : (0, 1) \rightarrow M_k$ is still a piecewise \mathcal{C}^1 curve and that

$$\begin{aligned} Pl_k(\gamma_k) &= \int_0^1 \sqrt{g_{\gamma_k(s)}^{(k)}(\dot{\gamma}_k(s), \dot{\gamma}_k(s))} ds \\ &= \int_0^1 \sqrt{((\Lambda_k \circ \dots \circ \Lambda_i)^* g^{(k)})_{\gamma(s)}(\dot{\gamma}(s), \dot{\gamma}(s))} ds \\ &= Pl_i(\gamma) \end{aligned}$$

where $(\Lambda_k \circ \dots \circ \Lambda_i)^* g^{(k)}$ is the pullback of $g^{(k)}$ via $\Lambda_k \circ \dots \circ \Lambda_i$. \square

Corollary 1. Let $\gamma : [0, 1] \rightarrow M_i$ be a piecewise \mathcal{C}^1 curve. Consider the curve $\Gamma = \mathcal{N}_i \circ \gamma$ on $\mathcal{N}(M_0) \subseteq M_n$. Then $Pl_i(\gamma) = Pl_n(\Gamma)$, with L_n the length of a curve defined using the Riemannian metric $g^{(n)}$.

Proof. The thesis immediately follows from Proposition 1 setting $k = n$. \square

Lemma 1. M_i / \sim_i is an open, path-connected, Hausdorff, second-countable set.

Proof. An elementary property of quotient maps yields that M_i / \sim_i is still a path-connected space and by [3, Corollary 3.17] we also know that π_i is an open map, therefore the quotient set M_i / \sim_i is open. Since pseudometric spaces are completely regular [3, Section 7], we conclude that M_i / \sim_i is Tychonoff and therefore it is in particular T_2 . At last we note that, since π_i is an open quotient, M_i / \sim_i is also second-countable. \square

Proposition 2. If two points $p, q \in M_i$ are in the same class of equivalence, then $\mathcal{N}_i(p) = \mathcal{N}_i(q)$.

Proof. Let $p, q \in M_i$ two points in the same class of equivalence $[p]$. Then, since M_i is path connected by hypothesis, there is a piecewise \mathcal{C}^1 null curve $\gamma : [0, 1] \rightarrow M_0$ connecting q and p , with $Pl_i(\gamma) = 0$. Consider now the curve $\Gamma = \mathcal{N}_i \circ \gamma$ on M_n . By Corollary 1 we conclude that also $Pl_n(\Gamma) = 0$ and being $g^{(n)}$ a Riemannian metric we have that $\mathcal{N}_i(p) = \mathcal{N}_i(q)$. \square

Proposition 3. Let $x, y \in M_i$, then $x \sim_i y$ if and only if $x \sim_{\mathcal{N}_i} y$.

Proof. If $x \sim_i y$, then there is a piecewise \mathcal{C}^1 null curve γ with $\gamma(0) = x$ and $\gamma(1) = y$ and we have that $Pl_i(\gamma) = Pl_n(\mathcal{N}_i \circ \gamma) = 0$. Since $g^{(n)}$ is a non-degenerate Riemannian metric, $Pl_n(\mathcal{N}_i \circ \gamma) = 0$ entails that the tangent vector to $\mathcal{N}_i \circ \gamma(s)$ is the zero vector for every $s \in (0, 1)$ and therefore $\mathcal{N}_i \circ \gamma$ is the constant curve $\mathcal{N}_i \circ \gamma(s) = \mathcal{N}_i(x)$. This proves $x \sim_i y \Rightarrow x \sim_{\mathcal{N}_i} y$. Let us now assume

70 that $x \sim_{\mathcal{N}_i} y$. By definition we know that there is a piecewise \mathcal{C}^1 curve $\gamma : [0, 1] \rightarrow M_i$ such that
71 $\gamma(0) = x, \gamma(1) = y$ and $\mathcal{N}_i \circ \gamma(s) = \mathcal{N}_i(x) \forall s \in [0, 1]$. It remains to prove that γ is a null curve.
72 This follows from the fact that, being $\mathcal{N}_i \circ \gamma$ a constant curve, then $Pl_i(\gamma) = l(\mathcal{N} \circ \gamma) = 0$. \square

73 **Corollary 2.** *Under the hypothesis of Proposition 3, one has that $M_i / \sim_i = M_i / \sim_{\mathcal{N}_{i+1}}$. Moreover,*
74 *if two points $p, q \in M_i$ are connected by a \mathcal{C}^1 curve $\gamma : [0, 1] \rightarrow M_i$ satisfying $\mathcal{N}_i(p) = \mathcal{N}_i \circ \gamma(s)$*
75 *for every $s \in [0, 1]$, then they lie in the same class of equivalence.*

76 *Proof.* The statement follows immediately from Propositions 2 and 3 making use of the definitions
77 of the quotients \sim_i and $\sim_{\mathcal{N}_{i+1}}$. \square

78 **Theorem 1** (Godement's criterion, [2, 1]). *Let X be a smooth manifold and $R \subset X \times X$ be an*
79 *equivalence relation. The quotient X/R is a smooth manifold if and only if*

80 1) *R is a submanifold of $X \times X$*

81 2) *The projection map on the second component $pr_2 : R \subset X \times X \rightarrow X$ is a submersion.*

82 Now we can prove that M_i / \sim_i is a smooth manifold.

83 **Proposition 4.** $\frac{M_i}{\sim_i}$ *is a smooth manifold of dimension $\dim(\mathcal{N}(M_0))$.*

84 *Proof.* We prove that the quotient M_i / \sim_i is a smooth manifold using Godement's criterion (Theo-
85 rem 1). The graph \mathcal{G}_{i+1} of $\sim_{\mathcal{N}_i}$ is the union of $C_p \times C_p$, with C_p a connected component of $\mathcal{N}_i^{-1}(p)$,
86 with $p \in \mathcal{N}_i(M_{i-1}) \subseteq M_n$ and therefore \mathcal{G}_{i+1} is a submanifold of $M_i \times M_i$. Furthermore, the
87 restriction of the projection pr_2 to R is the restriction of the identity map to C_p for some $p \in M_i$,
88 which is a diffeomorphism with its image and therefore a submersion. The statement about the
89 dimension follows from the proof of 2) \Rightarrow 1) of Theorem 1, see [2, Lemma 9.4], taking in account
90 that $T_p \mathcal{N}_i = \dim(\text{Ker}(g^{(i)}))$ is constant. \square

91 This proposition, along with [2, Lemma 9.4 and Lemma 9.9], yields that the classes of equivalence
92 $[p]$ are the leaves of a simple foliation of M_i and that π_i is a smooth submersion.

93 **Proposition 5.** $\pi_i : M_i \rightarrow M_i / \sim_i$ *is a smooth fiber bundle, with $\text{Ker}(d\pi_i) = \mathcal{V}M_i$, which is*
94 *therefore an integrable distribution. Every class of equivalence $[p]$ is a path-connected submanifold*
95 *of M_i and coincide with the fiber of the bundle over p .*

96 *Proof.* The first part of the statement follows applying Proposition 4 together with [2, Lemma 9.4 and
97 Lemma 9.9]. The second part of the statement is then a consequence of the definitions of equivalence
98 class and vertical bundle. \square

99 4 ChatGPT prompts

100 In order to generate a small dataset of 100 sentences for hate speech detection, we prompt ChatGPT
101 (3.5 version) with several requests. The first one is: *can you generate 100 sentences with [CLS] and*
102 *[MASK] tokens as input for BERT? Do not enumerate them while printing them, I want to do a copy*
103 *paste directly in a txt file.*

104 Then: *can you do the same but generating sentences for training a BERT model for hate speech*
105 *detection?*

106 And finally: *can you do the same but without the [MASK] token and with some of them being offensive*
107 *(37%), other hate speech (39%) and others neutral (24%)? Still add the [CLS] token.*

108 This yields the dataset we used for exploration of BERT input embedding space.

5 Using interpretation outputs as alternative prompts (cont.)

ViT experiments were conducted using 1000 iteration outcomes from both SiMEC and SiMExp, applied to a subset of the MNIST dataset containing 14 images of the digit “4”. BERT for MLM experiments involved 1000 iterations from both SiMEC and SiMExp, applied to a subset of 8 sentences from the “fill in the mask” dataset (see Section 4 for more details).

For ViT experiments, we first extract the original predicted class $i^* = \arg \max_i y_i$, which represents the output whose equivalence class we aim to explore. Then, we run the interpretation algorithm for each $p_0 \dots p_K$ to obtain K interpretations in the form of images. Finally, we classify the new images, obtaining the corresponding predictions $Y = \mathbf{y}^{(0)} \dots \mathbf{y}^{(K)}$, where each $\mathbf{y}^{(k)} \in \mathbb{R}^N$, N being the number of prediction classes (e.g., $N = 10$ in MNIST). We visualize the prediction trend for the i^* th value in every $\mathbf{y}^{(0)} \dots \mathbf{y}^{(K)}$ categorizing the images into two subsets: those that lead to a change in prediction $Y_c = \{\mathbf{y}^{(k)} \in Y \mid \arg \max_i y_i^{(k)} \neq i^*\}$ and those that don’t $Y_s = \{\mathbf{y}_i \in Y \mid \arg \max_i y_i^{(k)} = i^*\}$.

Considering BERT for MLM experiments, we proceed as illustrated in the main text.

Figure 1 illustrates a scenario similar to the one described in the main text. Specifically, it shows SiMExp’s tendency to identify modifications that lower the prediction value for the original equivalence class, i^* .

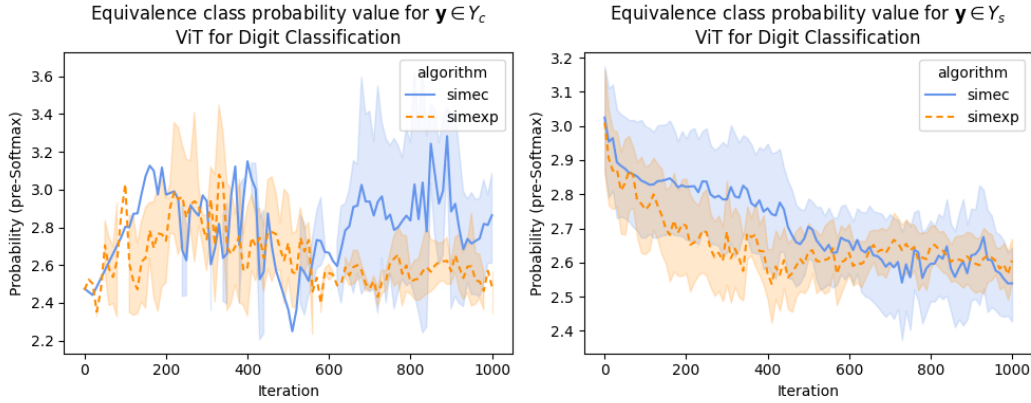


Figure 1: Results of ViT experiments involving 1000 iterations from both SiMEC and SiMExp, applied to subsets of MNIST dataset, each containing 14 sentences. The left plot shows prediction values for i^* for each $\mathbf{y} \in Y_c$, whereas the right plot depicts prediction values for $\mathbf{y} \in Y_s$. In general, both plots show a similar trend where SiMExp generally identifies alternative prompts that lower the prediction value for i^* .

Results for experiments using BERT for MLM are depicted in Figure 2. The plot on the left side exhibits the expected behavior. However, when selecting alternative tokens that lead to the same equivalence class, SiMExp seems unable to lower the prediction value of i^* . While our primary concern is validating the expected behavior when the prediction class changes (left plot), a deeper investigation into using interpretation outputs as prompts will be crucial in future work.

6 Example of feature importance in image classification

As, to the best of our knowledge, we are not aware of any explainable version of the MNIST dataset, we show the results of the same methods applied to ViT in an example, reported in Figure 3. This example is in line with the results obtained on the sample of 100 images from the MNIST dataset and shows that our method gives higher importance to pixels that are contained in the actual shape of the digit or that would form another digit if white (in the case of 6, these are the upper right ones, which would transform 6 into 8). Though on different scales, the number of pixels assigned high importance by our approach is larger than in the AR approach and lower with respect to the Relevancy approach.

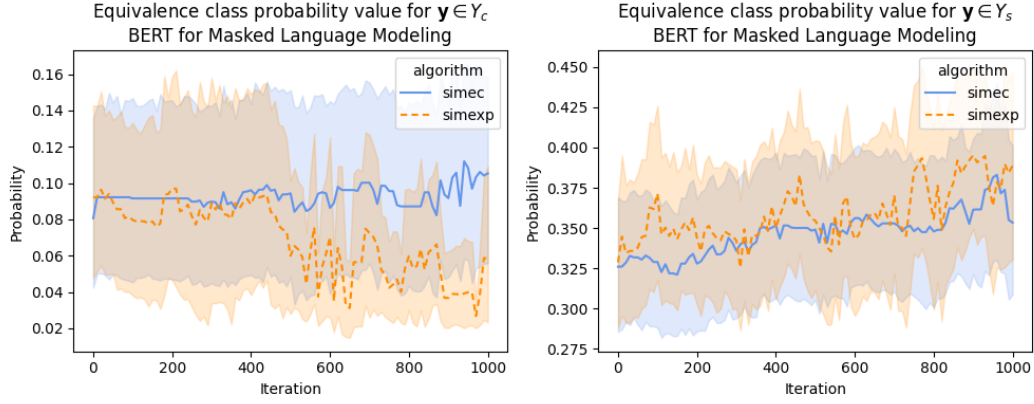


Figure 2: Results of BERT experiments involving 1000 iterations from both SiMEC and SiMExp, applied to subsets of a “fill in the mask” dataset, each containing 8 sentences. The left plot shows prediction values for i^* for each $y \in Y_c$, whereas the right plot depicts prediction values for $y \in Y_s$. Here, the expected behavior can be noted on the left side, where SiMExp tries to predict tokens which in turn lower the probability of the equivalence class prediction i^* . However, the same cannot be said for the plot on the right, where SiMExp is unable to find tokens leading to lower probabilities for i^* .

139 This contributes to give a precise (contrary to AR) and noiseless (contrary to the Relevancy method)
 140 explanation.

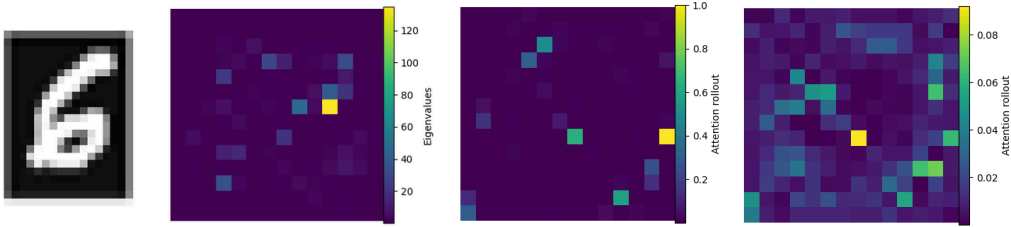


Figure 3: Example of feature importance with our method (left), AR (center) and the Relevancy method (right) on a sample image from the MNIST dataset.

References

- [1] N. Bourbaki. *Variétés différentielles et analytiques: Fascicule de résultats*. Springer-Verlag Berlin Heidelberg, 1967.
- [2] R. Fernadnes. Lectures on differential geometry, 2024. To be published by World Scientific, <https://www.math.tecnico.ulisboa.pt/~mabreu/GD/RLF-notes.pdf>.
- [3] T. Pirttimäki. A survey of Kolmogorov quotients, 2019. arXiv:1905.01157 [math.GN].
- [4] L. W. Tu. *An Introduction to Manifolds*. Springer-Verlag New York, 2 edition, 2011. doi: 10.1007/978-1-4419-7400-6.