Hipótesis de cambio de régimen: fundamentos para dinámicas desacopladas en el entrenamiento de redes neuronales

Cristian Pérez-Corral[1], Alberto Fernández-Hernández[1], Jose I. Mestre[2], Manuel F. Dolz[2], Jose Duato[3], Enrique S. Quintana-Ortí[1]

Universitat Politècnica de València, Valencia
{cpercor, a.fernandez}@upv.es, quintana@disca.upv.es
Universitat Jaume I, Castellón
{jmiravet, dolzm}@uji.es
Openchip & Software Technologies, Barcelona
jose.duato@openchip.com

Resumen Este artículo es un resumen del trabajo publicado en la *International Joint Conference on Neural Networks* [1] titulado *Regime Change Hypothesis: Foundastions for Decoupled Dynamics in Neural Network Training*.
A pesar del éxito empírico de las Deep Neural Networks (DNNs), su dinámica de entrenamiento sigue siendo difícil de caracterizar. En redes basadas en ReLU, los patrones de activación determinan las regiones a trozos lineales en las que la red se comporta de forma afín. Esta geometría motiva el estudio de si el entrenamiento presenta dos escalas temporales: una fase inicial con cambios relevantes en los patrones de activación y una fase posterior donde se refinan los parámetros sujetos a dichos patrones. En este trabajo, se propone un resultado teórico que justifica este comportamiento bajo ciertos supuestos, y se validan empíricamente, iteración a iteración, los cambios en los parámetros y en los patrones de activación en Multilayer Perceptrons (MLPs), Convolutional Neural Networks (CNNs) y modelos basados en Transformers. Los resultados muestran que, en general, los patrones de activación se estabilizan antes que los parámetros, lo que respalda la hipótesis planteada y motiva el desarrollo de otras estrategias de optimización.

Palabras clave: Patrones de activación · redes neuronales ReLU · dinámica de aprendizaje.

1. Introducción

El Deep Learning (DL) se ha consolidado como el paradigma dominante en numerosos dominios, entre ellos la Computer Vision (CV) y el Natural Language Processing (NLP) [2]. En redes con funciones de activación a trozos, y en particular con ReLU, la expresividad puede entenderse a partir de la partición del espacio de entrada en regiones definidas por patrones de activación, dentro de cada una de las cuales la red implementa una transformación afín [3].

La evolución de estos patrones durante el entrenamiento aporta información relevante sobre la dinámica de aprendizaje, y hay trabajos donde estudian tanto la evolución de estadísticas de activación [4] como la separación entre *conocimiento estructural*, asociado a las máscaras de activación, y *conocimiento cuantitativo*, asociado a los parámetros [5]. Esto motiva la hipótesis central de este trabajo: el entrenamiento podría admitir una interpretación en dos fases. Para apoyar esta hipótesis, se aporta soporte teórico y empírico a esta idea. En particular, se da la formulación un resultado de estabilidad local de patrones de activación en MLPs con ReLU, se propone una instrumentación práctica para comparar la evolución de parámetros y activaciones durante el entrenamiento, y se muestra evidencia empírica en MLPs, CNNs y modelos basados en Transformers de que, en general, los patrones de activación se estabilizan antes que los parámetros. A esta idea se la ha denominado **hipótesis de cambio de régimen**.

2. Fundamentos teóricos

Definición 1. *Un **Multilayer Perceptron (MLP)** con activaciones ReLU y L capas ocultas es una aplicación $f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_{L+1}}$ dada por*

$$z_l(x) = W_{l-1}h_{l-1}(x) + b_{l-1}, \qquad h_l(x) = \sigma(z_l(x)),$$

para $l = 1, \ldots, L$, con $h_0(x) = x$ y $\sigma(t) = \max(t, 0)$ aplicada componente a componente. La capa de salida es afín $f(x) = W_L h_L(x) + b_L$.

Definición 2. *Sea $f(\cdot; w)$ un MLP como el anterior. Para $x \in \mathbb{R}^{n_0}$ fijo, definimos el **patrón de activación** como $\mathrm{A}(x; w) = \big(\mathrm{A}_1(x; w), \ldots, \mathrm{A}_L(x; w)\big)$, donde*

$$\mathrm{A}_l(x; w) = \mathbf{1}\{z_l(x; w) > 0\} \in \{0, 1\}^{n_l}.$$

Es decir, cada componente indica si la neurona correspondiente está activa o inactiva para la muestra x.

Este patrón determina la región poliédrica del espacio de entrada en la que se encuentra la muestra. En una misma región, la red actúa como una transformación afín [3]. El siguiente resultado formaliza que, en general, estos patrones son localmente estables frente a pequeñas perturbaciones de los parámetros.

Proposición 1 (Estabilidad local de patrones de activación). *Sea $f(\cdot; w)$ un MLP con ReLU y $w \in \mathbb{R}^p$. Entonces existe un conjunto $B \subset \mathbb{R}^p$ de medida nula tal que, para todo $w_0 \in \mathbb{R}^p \setminus B$, existe un conjunto $Z \subset \mathbb{R}^{n_0}$ también de medida nula con la siguiente propiedad: para todo $x \in \mathbb{R}^{n_0} \setminus Z$ existe $\varepsilon > 0$ tal que, si $\|w - w_0\| < \varepsilon$, el patrón de activación de x coincide bajo w_0 y bajo w.*

La proposición (demostrada en [1, pág 3., Proposición 1]) establece que, para casi toda configuración de parámetros y casi toda muestra, perturbaciones suficientemente pequeñas no cambian el patrón de activación. En consecuencia, la

Cuadro 1: Comparación entre la convergencia de parámetros y de patrones de activación en un subconjunto representativo de experimentos.

Modelo	Datos	BVS	Ep.	AUC_W	AUC_P	ρ
MLP	Breast Cancer	0.97	18	0.32	0.26	**1.23**×
EfficientNetB0	CIFAR-100	0.51	159	0.58	0.18	**3.2**×
ViT-B16	CIFAR-100	0.61	50	0.87	0.08	**11**×

mayor parte del régimen de la red es lineal. Este resultado motiva la **hipótesis de cambio de régimen**: durante el entrenamiento, los patrones de activación tienden a estabilizarse antes que los parámetros, de modo que las fases tardías pueden interpretarse como un refinamiento paramétrico dentro de una colección casi fija de regiones afines.

3. Experimentos

En esta sección se evalúa si la hipótesis es correcta empíricamente y los patrones de activación se estabilizan antes que los parámetros. Para ello se ha seleccionado un subconjunto representativo de arquitecturas: un MLP, una CNN y un Transformer, para comprobar que la hipótesis no es dependiente de la arquitectura elegida. Para cuantificar la convergencia de parámetros medimos la media de la distancia ℓ_1: sea $w_i \in \mathbb{R}^p$ el vector de parámetros entrenables tras procesar el lote i-ésimo,

$$\Delta w_i = \frac{1}{p}\|w_i - w_{i-1}\|_1. \tag{1}$$

En la misma línea, para cuantificar la convergencia de patrones de activación, fijamos un subconjunto S de validación y medimos la fracción media de bits que cambian entre dos iteraciones consecutivas:

$$\Delta a_i = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{|S|\,n_l}\sum_{x\in S}\left\|\mathrm{A}_l(x;w_i) - \mathrm{A}_l(x;w_{i-1})\right\|_1. \tag{2}$$

Ambas trayectorias se suavizan mediante media móvil y se normalizan de forma robusta para hacer comparables sus tendencias dentro de cada ejecución. Para obtener un resumen global, calculamos el promedio temporal de cada curva normalizada, denotado por AUC_W para parámetros y AUC_P para patrones. Definimos además el factor de aceleración relativo como $\rho = \mathrm{AUC}_W/\mathrm{AUC}_P$, de modo que valores $\rho > 1$ indican una estabilización más rápida de los patrones de activación.

La Tabla 1 resume los resultados. En las tres arquitecturas consideradas se observa $\rho > 1$, lo que indica que los patrones de activación presentan menor cambio acumulado que los parámetros. Este efecto es moderado en el MLP, más

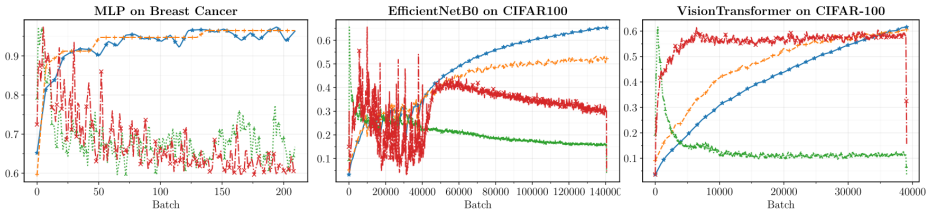


Figura 1: Curvas de entrenamiento y validación, además de las curvas de convergencia de los parámetros de la red y los patrones de activación.

acusado en la CNN y especialmente notable en el Transformer, donde la separación entre ambas dinámicas resulta muy marcada. En conjunto, estos resultados apoyan la hipótesis de cambio de régimen y sugieren que la estabilización temprana de las activaciones no es exclusiva de una familia concreta de modelos.

La Figura 1 ilustra visualmente este comportamiento: las curvas asociadas al cambio de activación descienden antes que las correspondientes a los parámetros, lo que sugiere que una parte importante de las fases tardías del entrenamiento transcurre dentro de regímenes de activación relativamente estables.

4. Conclusiones

Los resultados teóricos y empíricos presentados apoyan la hipótesis de cambio de régimen, al mostrar que los patrones de activación tienden a estabilizarse antes que los parámetros durante el entrenamiento, abriendo la puerta a nuevas estrategias de optimización más eficientes.

Referencias

1. C. Pérez-Corral, A. Fernández-Hernández, J. I. Mestre, M. F. Dolz, J. Duato, and E. S. Quintana-Ortí, "Regime change hypothesis: Foundations for decoupled dynamics in neural network training," 2026.
2. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
3. M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, "On the expressive power of deep neural networks," in *International Conference on Machine Learning (ICML)*, 2017.
4. D. Hartmann, D. Franzen, and S. Brodehl, "Studying the evolution of neural activation patterns during training of feed-forward relu networks," *Frontiers in Artificial Intelligence*, 2021.
5. J. Duato, J. I. Mestre, M. F. Dolz, E. S. Quintana-Orti, and J. Cano, "Decoupling structural and quantitative knowledge in relu-based deep neural networks," in *Proceedings of the 5th Workshop on Machine Learning and Systems*, ser. EuroMLSys '25, 2025.