Hue University Journal of Science: Natural Science
Vol. 129, No. 1B, 25–29, 2020

pISSN 1859-1388
eISSN 2615-9678

Nghiên cứu tính ổn định của Mạng nơ-ron tích chập bằng phương pháp chuẩn hoá phổ nhằm tăng hiệu suất nhận dạng của mô hình

Nguyễn Thị Hà Phương[1*], Hoàng Trọng Lợi[1]

[1] Khoa Kỹ thuật và Công nghệ, Đại học Huế
[*]email: nthphuong.huet@hueuni.edu.vn

Tóm tắt. Trong những năm gần đây, học sâu đã đạt được nhiều thành tựu nổi bật trong các lĩnh vực như nhận dạng hình ảnh, thị giác máy tính và xử lý ngôn ngữ tự nhiên. Tuy nhiên, các mô hình mạng nơ-ron tích chập (CNN) thường thể hiện độ nhạy cao với nhiễu và biến dạng nhỏ trong dữ liệu đầu vào, làm giảm độ tin cậy và khả năng khái quát hóa của hệ thống. Nghiên cứu này tập trung vào việc phân tích tính ổn định của mạng CNN dưới tác động của nhiễu Gaussian ngẫu nhiên, đồng thời đề xuất ứng dụng chuẩn hóa phổ (Spectral Normalization – SN) như một phương pháp cải thiện độ ổn định của mô hình. Chúng tôi xây dựng hai kiến trúc CNN và SN-CNN, tiến hành huấn luyện và đánh giá trên hai tập dữ liệu chuẩn MNIST và CIFAR-10, với nhiều mức độ nhiễu khác nhau. Kết quả mô phỏng cho thấy mô hình SN-CNN duy trì hiệu năng tốt hơn, với độ chính xác cao hơn 5–7% so với mô hình cơ sở khi dữ liệu đầu vào bị nhiễu, đồng thời thể hiện chỉ số stability thấp hơn đáng kể. Nghiên cứu góp phần làm rõ mối quan hệ giữa tính ổn định, độ bền vững và khả năng khái quát của mạng học sâu, mở ra hướng tiếp cận mới trong thiết kế các mô hình AI đáng tin cậy hơn trong môi trường thực tế có nhiễu.

Từ khoá: Học sâu, mạng nơ-ron tích chập (CNN), chuẩn hoá phổ, tính ổn định, độ bền vững, nhiễu Gaussian.

1 Giới thiệu

Trong những năm gần đây, học sâu (Deep Learning) đã trở thành nền tảng của nhiều hệ thống trí tuệ nhân tạo hiện đại, đặc biệt trong các bài toán nhận dạng hình ảnh, thị giác máy tính và xử lý ngôn ngữ tự nhiên. Trong đó, mạng nơ-ron tích chập (Convolutional Neural Network – CNN) là một trong những kiến trúc phổ biến nhất nhờ khả năng tự động trích chọn đặc trưng không gian của dữ liệu hình ảnh [1]. Tuy nhiên, các nghiên cứu gần đây cho thấy rằng các mô hình học sâu, mặc dù đạt độ chính xác cao, lại dễ bị suy giảm hiệu năng khi dữ liệu đầu vào chứa nhiễu hoặc biến dạng nhỏ [2][3]. Hiện tượng này làm nổi bật nhu cầu đánh giá và cải thiện tính ổn định (stability) và độ bền vững (robustness) của mạng nơ-ron.

Tính ổn định của mạng nơ-ron tích chập (CNN) đã thu hút sự quan tâm đáng kể từ cộng đồng nghiên cứu. Một số công trình đã chứng minh rằng CNN có mối liên hệ chặt chẽ với phép biến đổi *scattering transform*, vốn được biết đến với đặc tính ổn định trước nhiễu và các biến dạng hình học nhỏ [4]. Bên cạnh đó, các nghiên cứu khác đã đi sâu phân tích mối quan hệ giữa ổn định phổ (*spectral stability*) và khả năng khái quát hóa của mạng CNN, cho thấy việc điều chỉnh chuẩn phổ của trọng số có thể làm giảm độ nhạy của mạng đối với nhiễu [5]. Ngoài ra, các phân tích về trị riêng trong các lớp tích chập cũng chỉ ra rằng những mạng CNN có chuẩn phổ được giới hạn thường thể hiện mức độ ổn định cao hơn và khả năng chống nhiễu tốt hơn [3][9].

Một hướng tiếp cận nổi bật nhằm tăng tính ổn định là chuẩn hóa phổ (Spectral Normalization-SN), kỹ thuật được giới thiệu bởi [6], giúp giới hạn chuẩn Lipschitz của các lớp mạng, từ đó điều chỉnh biên độ thay đổi đầu ra theo biến động đầu vào.

Phương pháp này đã được chứng minh có hiệu quả trong việc giảm dao động gradient, tăng khả năng khái quát và giảm tác động của nhiễu ngẫu nhiên đối với mô hình học sâu [7][9].

Từ những vấn đề nêu trên, nghiên cứu này hướng tới việc phân tích và đánh giá tính ổn định của mạng CNN dưới tác động của nhiễu Gaussian, đồng thời đề xuất áp dụng chuẩn hóa phổ như một phương pháp điều chỉnh tham số giúp mô hình ổn định và bền vững hơn trước biến động của dữ liệu đầu vào. Kết quả mô phỏng trên hai tập dữ liệu chuẩn MNIST và CIFAR-10 được sử dụng để minh họa hiệu quả của phương pháp, qua đó góp phần cung cấp cơ sở cho các nghiên cứu tiếp theo về mô hình học sâu ổn định và đáng tin cậy hơn trong điều kiện dữ liệu thực tế có nhiễu.

2 Cơ sở lý thuyết và phương pháp đề xuất

2.1 Tính ổn định và độ bền vững trong học sâu

Tính ổn định

Tính ổn định trong học máy nhằm đo lường mức độ mà kết quả học thay đổi khi dữ liệu huấn luyện thay đổi nhỏ [10].

Giả sử ta có một thuật toán học A ánh xạ từ tập dữ liệu huấn luyện $S = \{(x_i, y_i)\}_{i=1}^{n}$ sang một mô hình f_SGọi $S^{(i)}$là tập dữ liệu giống S nhưng thay thế phần tử thứ i bằng một mẫu khác (x_i', y_i').

Thuật toán A được gọi là ổn định theo hàm mất mát nếu tồn tại hằng số $\beta_n \geq 0$ sao cho với mọi $i \in \{1, \dots, n\}$:

$$|\mathcal{L}(f_S, z) - \mathcal{L}(f_{S^{(i)}}, z)| \leq \beta_n, \qquad \forall z = (x, y) \quad (1)$$

trong đó $\mathcal{L}(f, z)$ là hàm mất mát của mô hình f tại mẫu z; β_n càng nhỏ thì thuật toán càng ổn định.

Từ định nghĩa này, chứng minh rằng sai số tổng quát hóa bị chặn bởi β_n.

$$|\mathbb{E}_S[R(f_S)] - R_{emp}(f_S)| \leq \beta_n \quad (2)$$

Trong đó $R(f_S) = \mathbb{E}_{z\sim D}[\mathcal{L}(f_S, z)]$ là rủi ro thực tế, $R_{emp}(f_S) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(f_S, z_i)$ là rủi ro kinh nghiệm. Điều này cho thấy tính ổn định cao giúp mô hình tổng quát hóa tốt hơn.

Trong học sâu, độ ổn định có thể được liên hệ với hàm Lipschitz của mạng nơ-ron:

$$\|f(x_1) - f(x_2)\| \leq L\|x_1 - x_2\| \quad (3)$$

với L là hằng số Lipschitz. Nếu L nhỏ, mô hình thay đổi chậm khi đầu vào thay đổi (tức là có tính ổn định cao). Việc chuẩn hóa trọng số hoặc ràng buộc chuẩn của ma trận giúp giảm L và tăng tính ổn định [3].

Độ bền vững

Độ bền vững phản ánh khả năng mô hình chống chịu trước nhiễu ngẫu nhiên hoặc nhiễu đối kháng trong dữ liệu đầu vào.

Giả sử x là đầu vào và δ là nhiễu (với $\forall\|\delta\| \leq \epsilon$). Mô hình f được xem là ε-robust nếu với mọi x:

$$sign(f(x)) = sign(f(x + \delta)), \quad \forall\|\delta\| \leq \epsilon \quad (4)$$

Điều này nghĩa là nhãn dự đoán không thay đổi khi đầu vào bị nhiễu nhỏ hơn ngưỡng ϵ. Một cách đo độ bền vững phổ biến là tổn thất đối kháng:

$$\mathcal{L}_{adv}(f; x, y) = \max_{\|\delta\|\leq\epsilon} \mathcal{L}(f(x + \delta), y) \quad (5)$$

Huấn luyện đối kháng [12] tìm mô hình tối ưu bằng:

$$\min_f \mathbb{E}_{(x,y)}\left[\max_{\|\delta\|\leq\epsilon} \mathcal{L}(f(x + \delta), y)\right] \quad (6)$$

Bài toán trên đảm bảo rằng mô hình đạt hiệu năng cao ngay cả trong trường hợp xấu nhất – khi dữ liệu bị nhiễu có chủ đích. Ngoài ra, độ bền vững còn có thể được đo bằng tỷ lệ chính xác sau tấn công đối kháng hoặc giới hạn xác suất sai lệch:

$$Robustness = \mathbb{P}(f = f(x + \delta)), \delta\sim N(0, \sigma^2 I) \quad (7)$$

pISSN 1859-1388
eISSN 2615-9678

Công thức này cho phép ước lượng xác suất mô hình duy trì kết quả đúng khi nhiễu ngẫu nhiên có phương sai σ^2.

2.2 Các mô hình phân tích ổn định đã có

Phân tích giới hạn Lipschitz trong mạng nơ-ron

Một trong những hướng tiếp cận đầu tiên nhằm phân tích tính ổn định của mạng học sâu là thông qua giới hạn Lipschitz. Nếu một mạng nơ-ron f có hằng số Lipschitz L nhỏ, thì đầu ra của mô hình sẽ thay đổi chậm khi đầu vào thay đổi, tức là ổn định hơn. Giới hạn này được định nghĩa như sau:

$$\|f(x_1) - f(x_2)\| \leq L\|x_1 - x_2\| \quad (8)$$

Và khi đầu vào bị nhiễu ϵ:

$$\mathbb{E}_\epsilon[\|f(x+\epsilon) - f(x)\|^2] \leq L^2\mathbb{E}[\|\epsilon\|^2] \quad (9)$$

Đây là định lý cơ bản về độ ổn định trung bình của ánh xạ Lipschitz.

Trong [3] đề xuất Parseval Networks, trong đó các trọng số của mạng được chuẩn hóa sao cho các ma trận trọng số gần trực giao ($W^TW \approx I$). Cách này giúp ràng buộc hằng số Lipschitz của từng tầng và giảm khuếch đại nhiễu khi lan truyền qua mạng. Kết quả cho thấy mô hình đạt độ ổn định cao hơn và ít bị ảnh hưởng bởi nhiễu đối kháng.

Với kỹ thuật mở rộng Spectral Normalization, tức là chia trọng số của mỗi tầng cho giá trị riêng lớn nhất, giúp kiểm soát độ nhạy của mạng mà không ảnh hưởng nhiều đến độ chính xác [6]. Các nghiên cứu này mở đường cho việc phân tích ổn định dưới góc nhìn toán học, dựa trên đặc tính tuyến tính xấp xỉ của mạng.

Phân tích độ bền vững trong CNN và RNN

Đối với các mạng CNN và RNN, phần lớn các nghiên cứu tập trung vào việc đánh giá thực nghiệm độ bền vững của mô hình trước nhiễu hoặc tấn công đối kháng.

Lúc đầu, các nhiễu nhỏ nhưng có hướng có thể khiến mô hình học sâu đưa ra dự đoán sai nghiêm trọng, hiện tượng này được gọi là adversarial vulnerability [2]. Tiếp theo, một số nhà nghiên cứu đã đề xuất Fast Gradient Sign Method (FGSM) để sinh nhiễu đối kháng và phân tích sự thay đổi của hàm mất mát theo hướng gradient [11].

Với CNN, nhiều công trình tập trung vào huấn luyện đối kháng, tức là tối ưu hóa mô hình theo công thức min–max nhằm tăng khả năng kháng nhiễu. Trong khi đó, các nghiên cứu trên RNN (như trong xử lý ngôn ngữ hoặc chuỗi thời gian) cho thấy tính ổn định phụ thuộc mạnh vào hàm kích hoạt và cơ chế lan truyền ngược theo thời gian (BPTT); nhiễu nhỏ ở đầu vào có thể lan truyền và khuếch đại qua nhiều bước thời gian [12][14].

Nhìn chung, các nghiên cứu này chủ yếu mang tính thực nghiệm hoặc bán phân tích, và chưa hình thành khung lý thuyết tổng quát cho độ ổn định thống kê của mạng nơ-ron phi tuyến.

Nghiên cứu gần đây về tính ổn định của Transformer

Trong những năm gần đây, mô hình Transformer đã trở thành kiến trúc nền tảng cho nhiều hệ thống AI hiện đại như BERT, GPT hay ViT. Tuy nhiên, cấu trúc self-attention phi tuyến và cơ chế chuẩn hóa tầng khiến việc phân tích ổn định của Transformer trở nên phức tạp hơn nhiều so với CNN hay RNN [13]. Một trong những công trình đầu tiên về phân tích định lượng tính ổn định của Transformer dưới nhiễu ngẫu nhiên và nhiễu đối kháng được trình bày trong [8]. Tác giả mô hình hóa sự lan truyền nhiễu qua các tầng attention và chứng minh rằng tính ổn định phụ thuộc phi tuyến vào độ lớn của vector truy vấn và trọng số attention. Kết quả thực nghiệm cho thấy một số biến thể của Transformer (như những mô hình sử dụng cơ chế "normalized attention") có khả năng ổn định tốt hơn trước nhiễu Gaussian.

Tuy vậy, các mô hình lý thuyết hiện tại vẫn chưa đưa ra được giới hạn rõ ràng cho độ ổn định thống kê của Transformer, và chưa mô tả được ảnh hưởng của cấu trúc mô hình (số tầng, số đầu

attention, loại hàm kích hoạt) lên mức độ ổn định tổng thể.

Từ các mô hình trên, vẫn còn thiếu một mô hình lý thuyết thống nhất mô tả mối quan hệ giữa nhiễu ngẫu nhiên, cấu trúc mô hình và độ ổn định thống kê. Đặc biệt, việc mô hình hóa nhiễu dưới khung xác suất và đưa ra giới hạn chặn cho Transformer gần như chưa được thực hiện.

2.3 Phương pháp đề xuất

Độ nhạy của mạng nơ-ron đối với nhiễu đầu vào phụ thuộc mạnh vào độ Lipschitz của hàm ánh xạ mà mạng biểu diễn. Nếu một mạng $f_\theta(x)$có hằng số Lipschitz nhỏ, thì với mọi nhiễu nhỏ ϵ, ta có:

$$\|f_\theta(x+\epsilon) - f_\theta(x)\| \le L\|\epsilon\| \quad (10)$$

Trong đó L càng nhỏ thì mô hình càng ổn định trước nhiễu.Vì vậy, một hướng tiếp cận hiệu quả là giới hạn giá trị L bằng cách điều chỉnh chuẩn phổ của các trọng số trong mạng.

Trong mô hình đề xuất, ta có mạng học sâu gồm các tầng tuyến tính W_i và các hàm kích hoạt trơn ϕ_i:

$$f_\theta(x) = W_n\phi_{n-1}(W_{n-1}\phi_{n-2}(\dots\phi_1 W_1 x))) \quad (11)$$

Tổng độ Lipschitz của mạng được chặn trên bởi tích các chuẩn phổ của từng tầng:

$$L(f_\theta) \le \prod_{i=1}^{n}\|W_i\|_2 \quad (12)$$

Trong đó $\|W_i\|_2$ là chuẩn phổ (singular value lớn nhất) của ma trận trọng số W_i.

Để kiểm soát độ ổn định, mô hình đề xuất bổ sung ràng buộc hoặc phạt lên các chuẩn này trong quá trình huấn luyện:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda\sum_{i=1}^{n}(\|W_i\|_2 - 1)^2 \quad (13)$$

với λ là hệ số điều chỉnh mức độ ổn định so với độ chính xác.

Ngoài ra, ta có thể áp dụng chuẩn hóa phổ trực tiếp trong huấn luyện (Spectral Normalization Layer):

$$\widehat{W}_i = \frac{W_i}{\|W_i\|_2} \quad (14)$$

đảm bảo $\|W_i\|_2 \le 1$ cho mọi tầng $\rightarrow L(f_\theta) \le 1$.

Với mô hình có chuẩn hóa phổ, sai lệch đầu ra do nhiễu Gaussian $\epsilon \sim N(0,\sigma^2 I)$ sẽ được chặn bởi công thức 15:

$$\mathbb{E}_\epsilon[\|f_\theta(x+\epsilon) - f_\theta(x)\|^2] \le \sigma^2 \quad (15)$$

độc lập với số tầng, giúp mô hình ổn định trước nhiễu đầu vào, ngay cả khi mạng rất sâu.

Đa số các nghiên cứu trước đó tập trung vào nhiễu đối kháng hoặc ổn định huấn luyện, ít nghiên cứu chuyên sâu về nhiễu ngẫu nhiên trong bối cảnh phân tích lý thuyết tính ổn định toàn cục. Do đó, hướng đề xuất trong bài báo này là phân tích định lượng ảnh hưởng của chuẩn phổ tới độ ổn định trung bình dưới nhiễu Gaussian ngẫu nhiên mở rộng các kết quả trước đó sang miền nhiễu ngẫu nhiên phi đối kháng.

Từ công thức trong Lipschitz công thức 9 và chuẩn hoá phổ ta thay thế L bằng $\prod_i\|\boldsymbol{W_i}\|_\mathbf{2}$ kết hợp lại ta có công thức ổn định của mô hình đề xuất như sau:

$$Stability_{SN}(f_\theta) = \mathbb{E}_{x,\epsilon}[\|f_\theta(x+\epsilon) - f_\theta(x)\|^2] \le \prod_i\|W_i\|_2^2\mathbb{E}[\|\epsilon\|^2] \quad (16)$$

Nếu mỗi $\|W_i\|_2 \le 1$ thì giới hạn trên chỉ còn $\mathbb{E}[\|\epsilon\|^2] = d\sigma^2$, không phụ thuộc vào độ sâu mạng. Điều này chứng minh rằng chuẩn hoá phổ giúp duy trì độ ổn định tuyến tính theo nhiễu, bất kể độ sâu kiến trúc.

3 Kết quả thực nghiệm

3.1 Dữ liệu và cấu trúc mô hình huấn luyện

Dữ liệu chương trình

Để đánh giá hiệu quả của phương pháp chuẩn hóa phổ trong việc tăng cường tính ổn định của mô hình học sâu, chúng tôi tiến hành mô phỏng trên tập dữ liệu MNIST và CIFAR-10.

pISSN 1859-1388
eISSN 2615-9678

Bảng 1. Tập dữ liệu MNIST và CIFAR-10

Đặc điểm	MNIST	CIFAR-10
Loại dữ liệu	Ảnh chữ số viết tay (0–9)	Ảnh vật thể thực (10 loại)
Số lớp	10 (0 đến 9)	10 (máy bay, ô tô, chim, mèo, nai, chó, ếch, ngựa, tàu, xe tải)
Kích thước ảnh	28×28 pixel, grayscale (1 kênh)	32×32 pixel, RGB (3 kênh)
Số lượng ảnh	60.000 train + 10.000 test	50.000 train + 10.000 test

Hình. 1. Một số hình ảnh trong hai tập dữ liệu CIFAR-10 và MNIST.

Cấu trúc CNN

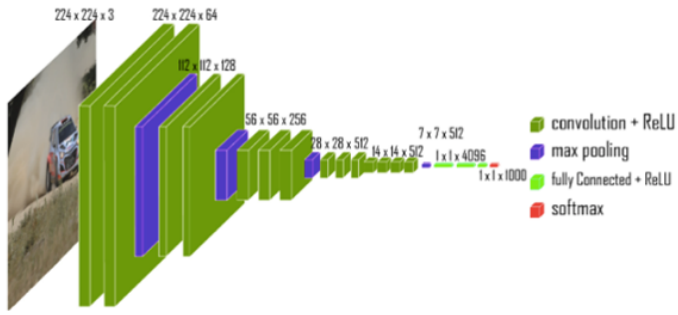
Trong bài báo này, tôi sử dụng cấu trúc CNN (Convolutional Neural Network) được sử dụng để huấn luyện và kiểm định độ ổn định.

Mô hình chứa tổng cộng 16 tầng chứa trọng số, bao gồm: 13 lớp tích chập, 3 lớp kết nối đầy đủ và sử dụng hàm kích hoạt ReLU và softmax ở đầu. Mô hình thực hiện được mô tả trong Hình : 1) Lớp Convolution, trích xuất đặc trưng bằng các lớp tích chập nhỏ sử dụng bộ lọc 3x3, stride=1 để bộ lọc di chuyển từng bước 1, giữ nguyên độ phân giải. Sử dụng hàm kích hoạt ReLU để giảm tính phi tuyến; 2) Giảm chiều bằng lớp MaxPooling 2×2, stride = 2 giúp giảm kích thước đầu vào và giữ lại thông tin quan trọng; 3) Lớp Fully Connected (FC): Ảnh được làm phẳng thành một vector 1 chiều. Ba lớp FC gồm 2 lớp 4096 nơ-ron và 1 lớp Softmax để đưa ra kết quả dự đoán với 1000 nhãn [15].

Hai mô hình CNN được sử dụng huấn luyện song song:

- Mô hình thường (Baseline CNN): sử dụng trọng số gốc, không áp dụng bất kỳ kỹ thuật điều chuẩn nào.
- Mô hình chuẩn hoá phổ (SN-CNN): áp dụng chuẩn hoá phổ cho tất cả các lớp tích chập và fully-connected theo phương pháp của [6].

Cả hai mô hình được huấn luyện trong 50 epoch với cùng siêu tham số: tốc độ học 0.001, batch size 64, hàm mất mát cross-entropy.



Hình. 2. Cấu trúc CNN (Nguồn: https://fptshop.com.vn/tin-tuc/danh-gia/convolutional-neural-network-173457)

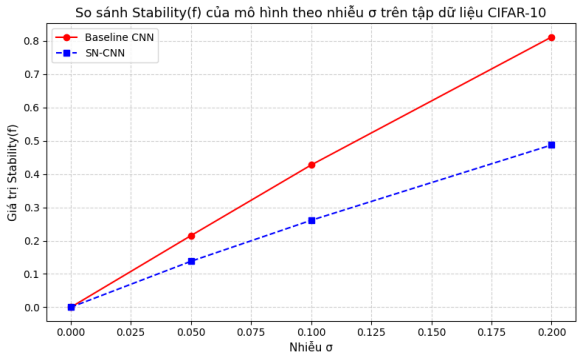
Để mô phỏng nhiễu ngẫu nhiên, dữ liệu đầu vào được cộng nhiễu Gaussian $N(0,\sigma^2)$ với $\sigma \in \{0.0, 0.05, 0.1, 0.2\}$.

Từ công thức 16 ta thực hiện tính độ ổn định. Tuy nhiên, trong thực nghiệm, độ ổn định không cần tính bình phương mà định lượng bằng chuẩn sai biệt trung bình giữa đầu ra của mạng khi có và không có nhiễu:

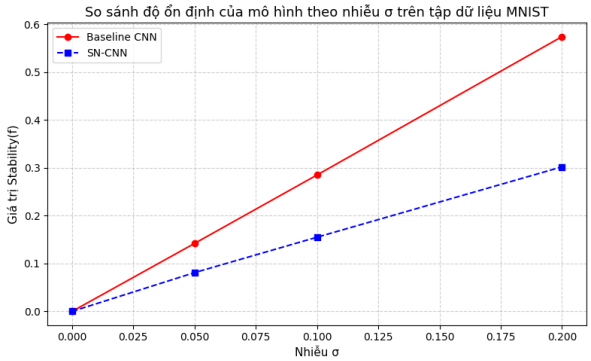
$$Stability_{SN}(f) = \mathbb{E}_{x,\epsilon}[\|f(x+\epsilon) - f(x)\|] \quad (17)$$

Trong đó $f(x)$ là đầu ra của mạng, và ϵ là nhiễu Gaussian độc lập. Giá trị Stability(f) càng nhỏ thể hiện mô hình càng ổn định.

3.2 Kết quả của mô hình đề xuất



a)



b)

Hình. 3. Độ ổn định của mô hình Baseline CNN và SN-CNN trên các tập dữ liệu: a) CIFAR-10, b) MNIST.

Hình 3 thể hiện kết quả chuẩn hoá phổ giúp giảm đáng kể độ nhạy của mô hình đối với nhiễu đầu vào. Khi mức nhiễu tăng, đường cong Stability(f) của mô hình SN-CNN tăng chậm hơn so với mô hình thông thường, chứng tỏ khả năng ổn định cao hơn.

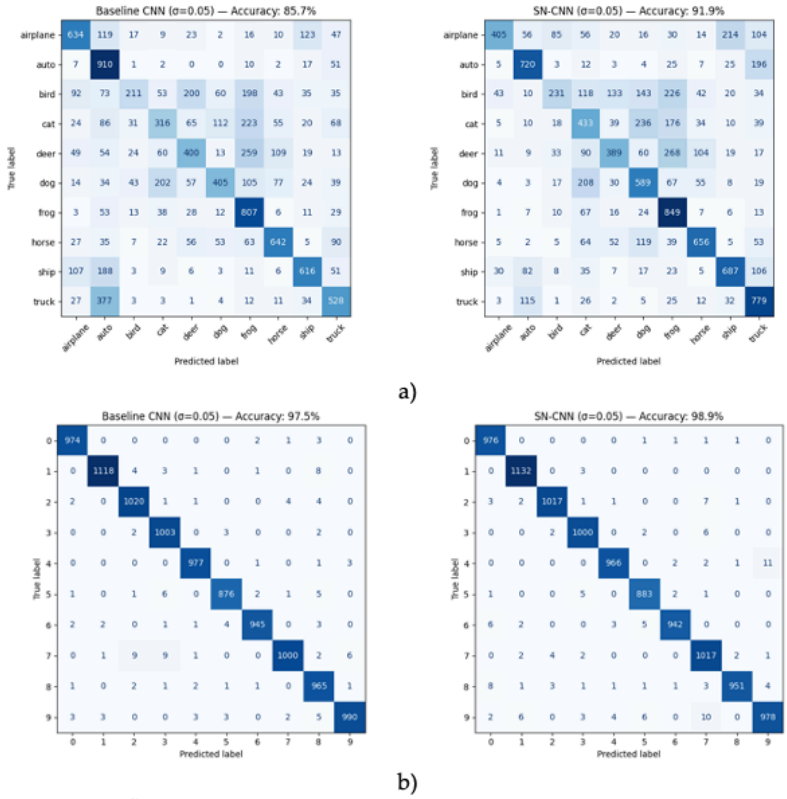
Mức cải thiện trung bình ≈ 40–50% về chỉ số Stability(f) đối với tập dữ liệu MNIST, ≈ 30-40% đối với tập dữ liệu CIFAR-10.

Hình 4a) thể hiện ma trận nhầm lẫn của hai mô hình CNN khi thực nghiệm trên tập dữ liệu CIFAR-10 có nhiễu mức σ = 0.05. Mô hình Baseline CNN cho thấy mức độ nhầm lẫn cao giữa các lớp có đặc trưng thị giác tương tự, đặc biệt là giữa "cat–dog" và "automobile–truck". Các phần tử ngoài đường chéo chính xuất hiện với giá trị lớn, phản ánh khả năng phân tách đặc trưng kém khi dữ liệu bị nhiễu. Ngược lại, mô hình SN-CNN với chuẩn hóa phổ thể hiện ma trận nhầm lẫn tập trung hơn dọc theo đường chéo chính, cho thấy khả năng nhận dạng ổn định hơn và giảm rõ rệt các sai lệch giữa các lớp. Kết quả này minh chứng rằng chuẩn hóa phổ giúp kiểm soát độ lớn của ma trận trọng số, làm tăng tính ổn định và độ bền vững của mô hình trước biến thiên nhiễu, đồng thời cải thiện độ chính xác từ 85.7% lên 91.9%.

Hình 4b) mô tả kết quả phân loại trên tập dữ liệu MNIST giữa hai mô hình Baseline CNN và SN-CNN. Cả hai mô hình đều đạt hiệu suất cao do đặc trưng đơn giản của bộ dữ liệu, tuy nhiên SN-CNN vẫn thể hiện độ chính xác vượt trội hơn (97.5% so với 98.9%). Đường chéo trong ma trận của SN-CNN thể hiện rõ ràng hơn, cho thấy tỉ lệ nhận dạng đúng cao hơn ở một số chữ số dễ nhầm như "3", "5" và "8". Mức cải thiện tuy nhỏ nhưng ổn định, phản ánh khả năng học khái quát tốt hơn của mô hình khi được kiểm soát phổ của trọng số. Kết quả này khẳng định hiệu quả của chuẩn hoá phổ không chỉ trong điều kiện dữ liệu nhiễu mà còn ở các bài toán có dữ liệu tiêu chuẩn, giúp mô hình đạt được độ ổn định hội tụ cao và giảm hiện tượng overfitting.

Các lỗi nhận dạng chủ yếu xảy ra giữa những lớp có đặc trưng hình ảnh tương tự nhau, chẳng hạn như "cat" (mèo) thường bị nhầm với "dog" (chó), "deer" (nai) nhầm với "horse" (ngựa), "truck"

pISSN 1859-1388
eISSN 2615-9678



Hình. 4. Ma trận nhầm lẫn của hai mô hình Baseline CNN và SN-CNN tại $\sigma = 0.05$ trên các tập dữ liệu: a) CIFAR-10, b) MNIST.

nhầm với "automobile", và "bird" nhầm với "airplane", "4" bị nhầm thành "9", "5" nhầm với "6", "7" nhầm với "1" trong một số trường hợp nhiễu cao.

Bảng 2. Kết quả nhận dạng của các mô hình trên tập dữ liệu CIFAR-10

Mức nhiễu σ	Baseline CNN	SN-CNN
0.00	97.2%	97.1%
0.05	85.7%	91.9%
0.10	70.5%	84.5%
0.20	45.1%	69.8%

Bảng 3. Kết quả nhận dạng của các mô hình trên tập dữ liệu MNIST

Mức nhiễu σ	Baseline CNN	SN-CNN
0.00	99.25%	99.21%
0.05	97.48%	98.86%
0.10	93.12%	97.20%
0.20	78.45%	91.85%

Khi $\sigma = 0$ (không nhiễu), cả hai mô hình đều hoạt động gần tối ưu. Khi σ tăng, độ chính xác của Baseline CNN giảm nhanh, trong khi SN-CNN giảm chậm hơn và cải thiện tính ổn định và độ bền vững.

Phản ánh hiệu quả của chuẩn hoá phổ trong việc hạn chế biến động đầu ra khi dữ liệu bị nhiễu.

Nhìn chung, việc áp dụng chuẩn hóa phổ đã cho thấy khả năng cải thiện đáng kể tính ổn định của CNN mà không cần thay đổi kiến trúc mạng hay tăng độ phức tạp tính toán đáng kể. Phương pháp này giúp mô hình trở nên ổn định hơn trước biến thiên dữ liệu, duy trì được hiệu năng cao trên nhiều loại tập dữ liệu khác nhau, từ ảnh đơn giản (MNIST) đến ảnh màu phức tạp (CIFAR-10). Những kết quả này mở ra tiềm năng áp dụng chuẩn hoá phổ trong các hệ thống học sâu khác, chẳng hạn như GAN, RNN hay Transformer, nơi yêu cầu về ổn định mô hình đóng vai trò quan trọng trong huấn luyện và triển khai thực tế.

4 Kết luận

Trong bài báo này, chúng tôi đã tập trung phân tích và đánh giá tính ổn định của mạng nơ-ron tích chập khi chịu tác động của nhiễu Gaussian ngẫu nhiên, đồng thời đề xuất ứng dụng chuẩn hóa phổ như một phương pháp cải thiện độ ổn định và độ bền vững của mô hình học sâu.

Kết quả mô phỏng trên hai tập dữ liệu chuẩn MNIST và CIFAR-10 cho thấy mô hình SN-CNN thể hiện khả năng duy trì độ chính xác tốt hơn đáng kể so với mô hình CNN thông thường, đặc biệt trong điều kiện dữ liệu bị nhiễu, cụ thể độ chính xác cao hơn 5–7%. Chỉ số stability(f) của SN-CNN nhỏ hơn, chứng tỏ mức dao động đầu ra trước biến động đầu vào được kiểm soát tốt hơn. Điều này cho thấy việc giới hạn chuẩn Lipschitz thông qua chuẩn hóa phổ giúp mô hình giảm nhạy cảm với nhiễu, ổn định hơn trong huấn luyện và tổng quát hóa tốt hơn. Nghiên cứu góp phần củng cố mối liên hệ giữa tính ổn định thống kê, chuẩn Lipschitz và hiệu năng mô hình học sâu.

Trong các nghiên cứu tiếp theo, nhóm tác giả dự kiến mở rộng mô hình sang các dạng nhiễu đối kháng và biến dạng hình học, đồng thời đề xuất các ràng buộc phổ thích nghi nhằm tối ưu hóa giữa tính ổn định và khả năng biểu diễn của mô hình.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. (2015); 521(7553): 436–444.
2. Szegedy C, et al. Intriguing properties of neural networks. ICLR 2014.
3. Cissé M, Bojanowski P, Grave E, Dauphin Y, Usunier N. Parseval networks: Improving robustness to adversarial examples. ICML 2017.
4. MallatS. Group invariant scattering. Communications on Pure and Applied Mathematics. 2012; 65(10): 1331–1398.
5. Bietti A, Mairal J. On the inductive bias of neural tangent kernels. Advances in Neural Information Processing Systems (NeurIPS). 2019.
6. Miyato T, Kataoka T, Koyama M, Yoshida Y. Spectral normalization for generative adversarial networks. ICLR 2018.
7. Yoshida Y, Miyato T. Spectral Norm Regularization for Improving the Generalizability of Deep Learning. 2017; arXiv:1705.10941.
8. Qin C, Chen X, Zhang C. Understanding perturbation stability of Transformers. ICML 2023 Workshop on Reliable ML. 2023.
9. Sedghi H, Gupta V, Long P. M. The singular values of convolutional layers. ICLR. 2019.
10. Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. Journal of Machine Learning Research, 2, 499–526.
11. Goodfellow I. J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. ICLR 2015.
12. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. ICLR 2018.
13. Vaswani A, et al. Attention is all you need. NeurIPS 2017.
14. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. ICML 2013.
15. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2016; pp. 770-778.