TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

HỌC CÓ GIÁM SÁT VỚI DỮ LIỆU NHIỄU BẰNG PHƯƠNG PHÁP LỰA CHỌN DỮ LIỆU

(Supervised learning with noisy data using data selection method)

1 THÔNG TIN CHUNG

Người hướng dẫn:

- TS. Nguyễn Ngọc Thảo (Khoa Công nghệ Thông tin)
- ThS. Trần Trung Kiên

Nhóm Sinh viên thực hiện:

1. Lê Xuân Hoàng (MSSV: 20120089)
2. Lê Xuân Huy (MSSV: 20120494)

Loại đề tài: Nghiên cứu

Thời gian thực hiện: Từ *01/2024* đến *06/2024*

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

Bài toán học có giám sát với dữ liệu nhiễu được phát biểu như sau:

- Cho tập dữ liệu có chứa nhiễu $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ với (x_i, y_i) lần lượt là input và nhãn của mẫu thứ i. Trong khóa luận, chúng em tập trung vào dữ liệu có nhãn nhiễu, tức là chúng ta chỉ quan sát được input và nhãn nhiễu của nó $\{y_i\}_{i=1}^n$ nhưng không quan sát được nhãn thật của nó $\{\widetilde{y}_i\}_{i=1}^n$.
- Yêu cầu: tìm hàm $f : \mathbb{R}^d \to \mathbb{R}$ có thể dự đoán nhãn thật $\widetilde{y} \in \mathbb{R}$ của $x \in \mathbb{R}^d$ mới ở ngoài $\mathcal{D}$.

Bài toán học có giám sát với dữ liệu nhiễu là bài toán thiết yếu trong thời buổi dữ liệu được ví như là "mỏ vàng". Ngày nay, mạng nơ-ron đang ngày càng phát triển đi theo đó là nhu cầu với dữ liệu ngày càng tăng cao. Trên thực tế, dữ liệu thường có rất nhiều nhãn nhiễu, để thu thập được dữ liệu tốt thì tốn rất nhiều thời gian và chi phí. Hơn nữa, mạng nơ-ron có một đặc điểm là có khả năng ghi nhớ bất kỳ nhãn nào (thậm chí ngẫu nhiên) của dữ liệu [1] nên dễ bị quá khớp (overfitting) với nhãn nhiễu. Vì vậy, giải quyết bài toán trên sẽ giúp nâng cao hiệu quả của các mô hình mạng nơ-ron nói riêng và mô hình học máy nói chung, đồng thời giảm các chi phí cho việc thu thập dữ liệu. Hiện nay, đã có nhiều phương pháp để giải quyết bài toán này theo cả hai hướng *model-centric* (tập trung vào xây dựng mô hình) và *data-centric* (tập trung vào xây dựng dữ liệu). Gần đây, hướng *data-centric* có rất nhiều phát triển vượt bậc và được kỳ vọng trong tương lai. Cụ thể với bài toán trên, phương pháp lựa chọn dữ liệu có nhiều nghiên cứu đạt hiệu suất state-of-the-art. Đây là hướng tiếp cận mà chúng em chọn để tìm hiểu.

2.2 Mục tiêu đề tài

- Hiểu rõ hơn về sự ảnh hưởng của dữ liệu nhiễu với các mô hình học máy đồng thời tìm hiểu các nghiên cứu để giải quyết vấn đề này (có những hướng giải quyết nào? Phương pháp nào được đề xuất? Ý tưởng và ưu/nhược điểm các phương pháp đó). Từ đó tìm hiểu sâu hơn về một phương pháp tiềm năng (ngoài ra, phương pháp đó phải khả thi để có thể hoàn thành trong khoảng thời gian thực hiện khóa luận).
- Hiểu rõ lý thuyết và ưu/nhược điểm của phương pháp đã chọn thông qua đọc, cài đặt và thực hiện thí nghiệm trong bài báo (nếu có thời gian thì thực hiện thêm ngoài bài báo).
- Rèn luyện các kỹ năng: tìm kiếm, lên kế hoạch, làm việc nhóm, trình bày, viết bài,...

2.3 Phạm vi của đề tài

Đề tài chỉ tập trung vào dữ liệu hình ảnh có chứa nhãn nhiễu và chỉ xét các bài toán phân loại. Tập dữ liệu được sử dụng trong đề tài là CIFAR-10 và CIFAR-100. Về mô hình, đề tài chỉ tập trung vào mô hình mạng nơ-ron. Về mục tiêu, đề tài chỉ tìm hiểu và cài đặt lại phương pháp của một bài báo uy tín; ngoài ra, có thể có thêm các thí nghiệm ngoài bài báo nhằm thấy rõ hơn về ưu/nhược điểm của phương pháp. Lý do chúng em giới hạn đề tài như vậy là vì:

- Để đảm bảo về mặt thời gian thực hiện. Chúng em thấy rằng việc hiểu rõ phương pháp (và các kiến thức nền tảng bên dưới) và cài đặt lại phương pháp là tốn khá nhiều thời gian.
- Việc hiểu rõ phương pháp (và các kiến thức nền tảng bên dưới) là cơ sở để có các cải tiến trong tương lai cũng như vận dụng cho bài toán khác.

Tất nhiên, nếu có thời gian thì chúng em cũng có thể sẽ có những đề xuất hoặc

cải tiến. Tuy nhiên, chúng em xác định rằng đây không phải mục đích chính.

2.4 Cách tiếp cận dự kiến

Dưới đây là các nghiên cứu về các phương pháp giải quyết bài toán huấn luyện mạng nơ-ron với dữ liệu có nhãn nhiễu mà chúng em tìm hiểu đến thời điểm hiện tại, cũng như phương pháp mà chúng em chọn để tìm hiểu sâu. Bài toán đã được giải theo nhiều hướng khác nhau, sau đây là các hướng đi phổ biến:

- Ước lượng ma trận chuyển đổi nhiễu:
  - Ý tưởng: ước lượng ma trận T là tỷ lệ của một nhãn bị chuyển đổi thành nhãn khác do yếu tố nhiễu, sau đó thay đổi hàm lỗi huấn luyện theo ma trận này.
  - Nghiên cứu tiêu biểu: F-correction (Patrini et al., 2017) [2] là một trong những phương pháp đạt hiệu suất rất tốt.
  - Nhược điểm: việc ước lượng ma trận chuyển đổi nhiễu rất khó chính xác, đặc biệt khi số lượng lớp cần phân loại càng nhiều.
- Thiết kế hàm lỗi huấn luyện:
  - Ý tưởng: cải tiến lại hàm lỗi huấn luyện sao cho mô hình có thể hoạt động tốt với dữ liệu nhiễu.
  - Nghiên cứu tiêu biểu: $\mathcal{L}_{DMI}$ (Xu et al., 2019) [3] là hàm lỗi huấn luyện hiệu quả cho bất kỳ mạng nơ-ron phân loại nào mà không cần bất kỳ thông tin bổ sung nào.
  - Nhược điểm: chỉ hoạt động hiệu quả trong các trường hợp đơn giản, khi dữ liệu ít phức tạp hoặc số lượng lớp cần phân loại ít.
- Sử dụng kỹ thuật regularization:
  - Ý tưởng: thêm các ràng buộc, các điều kiện vào quá trình huấn luyện để ngăn chặn mô hình học theo các nhãn nhiễu.

  - Nghiên cứu tiêu biểu: Weight decay (Krogh, A. and Hertz, J.A., 1992) [4] là một kỹ thuật không còn xa lạ, được sử dụng phổ biến để giảm thiểu overfitting.
  - Nhược điểm: khó đạt được hiệu suất tốt khi dữ liệu có tỉ lệ nhãn nhiễu cao.

- Lựa chọn dữ liệu:
  - Ý tưởng: phương pháp lựa chọn dữ liệu có thể được thực hiện bằng hai cách là chọn mẫu hoặc điều chỉnh trọng số mẫu. Chọn mẫu nghĩa là chọn ra một tập hợp con từ tập dữ liệu sao cho tập con đó được xem là "sạch" nhất. Trong khi đó, điều chỉnh trọng số mẫu là đánh trọng số thấp cho những mẫu được xem là nhiễu và đánh trọng số cao cho những mẫu tốt.
  - Nghiên cứu tiêu biểu: Decoupling (Malach, ShalevShwartz, 2017) [5], MentorNet (Jiang et al., 2018) [6], Co-teaching (Han et al., 2018) [7], INCV (Chen et al., 2019) [8], CRUST (Mirzasoleiman et al., 2020) [9].
  - Hướng nghiên cứu này là hướng nghiên cứu hứa hẹn nhất với nhiều bài báo đạt hiệu suất state-of-the-art. Trong các nghiên cứu trên, phương pháp CRUST có hiệu suất tốt nhất.

Với những gì đã trình bày ở trên, trong khóa luận, chúng em sẽ tập trung tìm hiểu và cài đặt phương pháp CRUST được trình bày trong bài báo. Phương pháp này đạt hiệu suất rất cao, điều quan trọng hơn là phương pháp này đảm bảo về mặt lý thuyết. Phương pháp như vậy sẽ an toàn hơn nhiều trong các hệ thống yêu cầu độ an toàn cao như máy bay, ô tô tự lái và thiết bị y tế. Việc tìm hiểu và cài đặt phương pháp trên không chỉ giúp chúng em hiểu rõ hơn về phương pháp lựa chọn dữ liệu mà còn hiểu các cơ sở lý thuyết để một mô hình hoạt động tốt với dữ liệu nhiễu.

2.5 Kết quả dự kiến của đề tài

- Cài đặt lại được từ đầu phương pháp được đề xuất trong bài báo [9].
- Có được các kết quả thí nghiệm để cho thấy phương pháp tự cài đặt ra được các kết quả như trong bài báo.
- Có được các kết quả thí nghiệm để thấy rõ về ưu/nhược điểm của phương pháp.
- Nếu có thời gian thì có thể thực hiện những thí nghiệm khác ngoài bài báo và tìm hướng cải tiến.

2.6 Kế hoạch thực hiện

STT	**Công việc**	**Thời gian bắt đầu**	**Thời gian kết thúc**	**Phân công**
1	Tìm những hướng nghiên cứu phù hợp	01/01/2024	15/01/2024	Huy, Hoàng
2	Xác định hướng nghiên cứu sẽ chọn	16/01/2024	23/01/2024	Huy, Hoàng
3	Tìm những paper xoay quanh hướng nghiên cứu đã chọn	24/01/2024	10/02/2024	Huy, Hoàng
4	Xác định paper chính sẽ chọn	11/01/2024	18/02/2024	Huy, Hoàng
5	Lên kế hoạch những việc cần làm	19/02/2024	15/03/2024	Huy, Hoàng
6	Mô tả bài toán, mục tiêu bài toán	21/02/2024	28/02/2024	Huy
7	Tìm hiểu các bài báo và các phương pháp liên quan đã được đề xuất	21/02/2024	06/03/2024	Huy, Hoàng
8	Xác định phương pháp mà nhóm chọn và hướng tiếp cận	28/02/2024	06/03/2024	Hoàng
9	Viết kế hoạch thực hiện	06/03/2024	12/03/2024	Huy, Hoàng
10	Viết đề cương	12/03/2024	15/03/2024	Huy, Hoàng
11	Đọc hiểu sâu paper chính	15/03/2024	15/05/2024	Huy, Hoàng
12	Đọc các paper liên quan	15/03/2024	15/05/2024	Huy, Hoàng
13	Đọc hiểu code	21/03/2024	10/04/2024	Huy, Hoàng
14	Viết code cài đặt	11/04/2024	15/05/2024	Huy, Hoàng
15	Thí nghiệm chương trình	01/05/2024	15/05/2024	Huy, Hoàng
16	Thực hiện những thí nghiệm khác ngoài bài báo và tìm hướng cải tiến	15/05/2024	01/06/2024	Huy, Hoàng
17	Viết cuốn và slides	15/05/2024	15/06/2024	Huy, Hoàng

Bảng 1: Bảng kế hoạch

Tài liệu

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *Communications of the ACM*, vol. 64, 2016.

[2] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017.

[3] Y. Xu, P. Cao, Y. Kong, and Y. Wang, "L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise," in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[4] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Proc, NeurIPS*, pp. 950–957, 1992.

[5] E. Malach and S. Shalev-Shwartz, "Decoupling "when to update" from "how to update"," *Advances in Neural Information Processing Systems*, 2017.

[6] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *ICML*, 2018.

[7] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *NeurIPS*, pp. 8535–8545, 2018.

[8] P. Chen, B. B. Liao, G. Chen, and S. Zhang, "Understanding and utilizing deep neural networks trained with noisy labels," in *International Conference on Machine Learning*, pp. 1062–1070, 2019.

[9] B. Mirzasoleiman, K. Cao, and J. Leskovec, "Coresets for robust training of neural networks against noisy labels," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

XÁC NHẬN
CỦA NGƯỜI HƯỚNG DẪN
(Ký và ghi rõ họ tên)

TP. Hồ Chí Minh, ngày 05 tháng 04 năm 2024
NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)

Trần Trung Kiên

Lê Xuân Huy

Nguyễn Ngọc Thảo

Lê Xuân Hoàng