|  |
| --- |
| BAN CƠ YẾU CHÍNH PHỦ  **HỌC VIỆN KỸ THUẬT MẬT MÃ**  ¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯  Logo HvKTMM |
| BÁO CÁO KẾT THÚC HỌC PHẦN  Ngành: An toàn thông tin  ĐỀ TÀI  Survey về tấn công dữ liệu nhạy cảm trong machine learning |
| Môn học : Machine Learning  Lớp : AT13CLC01  Giảng viên hướng dẫn : TS.Trần Anh Tú  Sinh viên thực hiện : Bùi Văn Công  Hoàng Đăng Luân |
| Hà Nội, 2020 |

# **Nhận Xét Của Giảng Viên**

# **Mục Lục**

[Survey về tấn công dữ liệu nhạy cảm trong machine learning 1](#_Toc43266357)

[**Nhận Xét Của Giảng Viên** 2](#_Toc43266358)

[**Mục Lục** 3](#_Toc43266359)

[Survey về tấn công dữ liệu nhạy cảm trong machine learning 4](#_Toc43266360)

[I. Tổng quan về Machine Learning và Deep Learning. 4](#_Toc43266361)

[1. Machine Learning. 4](#_Toc43266362)

[2. Deep Learning . 5](#_Toc43266363)

[II. Dữ liệu nhạy cảm trong Machine Learning 8](#_Toc43266364)

[1. Xác định dữ liệu nhạy cảm 8](#_Toc43266365)

[2. Dữ liệu nhạy cảm trong các cột 8](#_Toc43266366)

[3. Dữ liệu nhạy cảm trong bộ dữ liệu dựa trên văn bản phi cấu trúc 8](#_Toc43266367)

[4. Dữ liệu nhạy cảm trong dữ liệu phi cấu trúc dạng tự do 8](#_Toc43266368)

[5. Dữ liệu nhạy cảm trong sự kết hợp của các lĩnh vực 9](#_Toc43266369)

[6. Dữ liệu nhạy cảm trong nội dung phi cấu trúc 10](#_Toc43266370)

[III. Các tấn công trong Machine Learning và Deep Learning lên dữ liệu. 10](#_Toc43266371)

[1. Phân loại tấn công gây mất an toàn. 10](#_Toc43266372)

[a) Causative Attack. 10](#_Toc43266373)

[b) Exploratory Attack 11](#_Toc43266374)

[c) Targeted Attack 11](#_Toc43266375)

[d) Indiscriminate Attack 11](#_Toc43266376)

[2. Các kiểu tấn công tấn công bất lợi . 11](#_Toc43266377)

[a. Tấn công suy luận thành viên 12](#_Toc43266378)

[b. Tấn công đảo ngược mô hình(Model Inversion Attacks) 13](#_Toc43266379)

[c. Công cụ chia sẻ bí mật: đánh giá và kiểm tra ghi nhớ ngoài ý muốn trong các neural networks 14](#_Toc43266380)

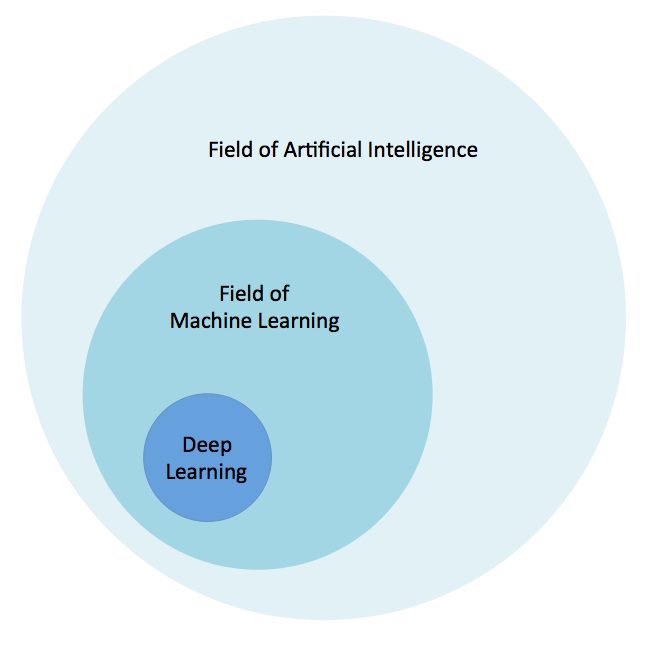
Survey về tấn công dữ liệu nhạy cảm trong machine learning

1. Tổng quan về Machine Learning và Deep Learning.
   * + 1. Machine Learning.

Machine learning là một lĩnh vực con của Trí tuệ nhân tạo(Artificial Intelligence) sử dụng các thuật toán cho phép máy tính có thể học từ dữ liệu để thực hiện các công việc thay vì được lập trình một cách rõ ràng.nó cho phép máy tính nội hóa các khái niệm tìm thấy trong dữ liệu để hình thành dự đoán cho các tình huống mới. Để đạt được mức độ chính xác đáng tin cậy, các mô hình yêu cầu các bộ dữ liệu lớn để 'học hỏi' từ đó. Để bảo vệ quyền riêng tư cá nhân trong bối cảnh dữ liệu lớn, các kỹ thuật ẩn danh khác nhau đã được sử dụng theo quy ước. Ba liên quan nhất là k-anonymity, L-diversity, và T-closeness. Trong k-anonymity, các cột cụ thể của QI(Quasi-Identifiers) (ví dụ: tên, tôn giáo) được xóa hoặc thay đổi (ví dụ: thay thế một độ tuổi cụ thể bằng một khoảng tuổi), do đó, trong tập dữ liệu giờ đây sẽ luôn có ít nhất 2 hàng với chính xác các thuộc tính tương tự (sau đó sẽ là một nhóm ẩn danh 2). L-diversity và T-closeness là phần mở rộng của khái niệm này. Những sửa đổi này được áp dụng trước khi dữ liệu được chia sẻ; đây được gọi là xuất bản dữ liệu bảo mật. Tuy nhiên, với sự phát triển của AI, hình thức bảo vệ này có thể không còn đủ nữa.

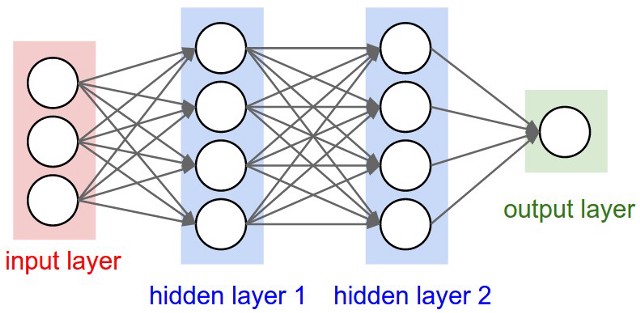
Những thách thức bảo mật dữ liệu cụ thể cho AI? Nhiều thách thức về quyền riêng tư được trích dẫn trong bối cảnh dữ liệu lớn cũng liên quan đến AI, ví dụ: có thể xác định lại thông tin cá nhân bằng cách sử dụng bộ dữ liệu lớn, chỉ sử dụng lượng thông tin cá nhân tối thiểu hoặc thiếu minh bạch về cách sử dụng dữ liệu của người tiêu dùng . Sự trỗi dậy của AI khác biệt về mặt định lượng đối với lượng dữ liệu khổng lồ, ví dụ, Baidu sử dụng hàng thập kỷ dữ liệu âm thanh để đào tạo thuật toán nhận dạng giọng nói của họ. Mặt khác, có tính đa chiều cao của dữ liệu được xem xét bởi các mô hình. Các kỹ thuật thống kê thông thường sẽ tính đến một số lượng hạn chế các biến được lựa chọn cẩn thận. Do các kỹ thuật chính quy mới và giảm chi phí tính toán (cũng do dịch vụ đám mây do Google hoặc AWS cung cấp), không gian tính năng có thể đã được mở rộng đáng kể, do đó các mô hình ML hiện có thể xem xét hàng ngàn biến để đưa ra dự đoán. Một ví dụ đáng chú ý là việc Google sử dụng mạng lưới thần kinh để giảm 40% hóa đơn làm mát của họ. Việc khai thác dữ liệu phi cấu trúc thông qua các kỹ thuật Deep Learning và khả năng kết hợp tất cả dữ liệu này trong một mô hình đã dẫn đến sự phong phú thông tin không lường trước được. Với các thuật toán có thể tạo ra các kết luận từ các bộ dữ liệu lớn và phức tạp như vậy, ba vấn đề khái niệm mới phát sinh.

* + - 1. Deep Learning .

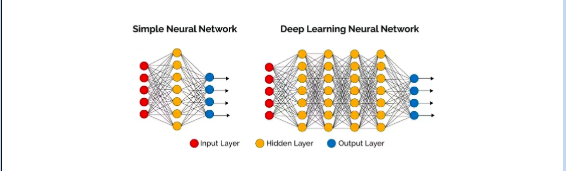


Mối quan hệ giữa DLvà ML, mạng neuron, và trí tuệ nhân tạo.

* **Deep learning** là: một tập hợp con của **Machine Learning**, có khả năng khác biệt ở một số khía cạnh quan trọng so với **Machine Learning** nông truyền thống, cho phép máy tính giải quyết một loạt các vấn đề phức tạp không thể giải quyết được …
* Một ví dụ về một nhiệm vụ Machine Learning đơn giản:
  + Có thể dự đoán doanh số bán kem sẽ thay đổi như thế nào dựa trên nhiệt độ ngoài trời.
  + Việc đưa ra dự đoán chỉ sử dụng một vài tính năng dữ liệu theo cách này là tương đối đơn giản và có thể được thực hiện bằng cách sử dụng một kỹ thuật Machine Learning gọi là hồi quy tuyến tính với độ dốc giảm dần.
* Deep learning được cấu tạo và có kiến trúc như sau:
* **Deep neural networks**
  + **Deep belief networks,**
  + **Recurrent neural networks**
  + **Convolutional neural networks**
* **Deep neural networks**
  + - * là hệ thống cấu trúc thần kinh phức tạp gồm nhiều đơn vị **neural network** mà trong đó, ngoài các lớp nguồn vào (**input**), nguồn ra (**output**) thì có hơn một lớp ẩn (**hidden layer**)
* Là một artificial neural network (ANN) có nhiều lớp giữa các lớp đầu vào đầu ra
* DNN tìm thấy thao tác toán học chính xác để biến đầu vào thành đầu ra, cho dù đó là mối quan hệ tuyến tínhhoặc một mối quan hệ phi tuyến tính.
* Ví dụ: một DNN được đào tạo để nhận ra các giống chó sẽ đi qua hình ảnh đã cho và tính xác suất để con chó trong hình ảnh là một giống chó nhất định.

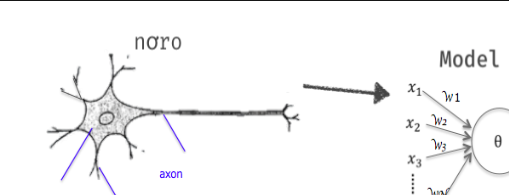


* + - Layer input: nhận các input đầu vào data.
    - Hidden layers: thực hiện các tính toán toán học dựa trên outputs
    - Layer output: trả về dữ liệu data output.
* Deep Neural Network được xây dựng với mục đích mô phỏng hoạt động não bộ phức tạp của con người và được áp dụng vào nhiều lĩnh vực khác nhau, mang lại thành công và những hiệu quả đáng kinh ngạc cho con người.
* Có khả năng máy tự đào sâu học hỏi nhờ vào công nghệ Deep Neural Network để từ đó xử lý và giải quyết những dữ liệu phi định danh hay phi cấu trúc.



Sơ đồ mô phỏng đơn vị Neural Network & Deep Learning Neural Network.

* **Neural Network** là mô hình toán học mô phỏng và biểu hiện cho một số chức năng của nơron thần kinh trong não bộ con người (Xem hình bên dưới).



**Neural là mô hình toán học mô phỏng nơron thần kinh con người.**

* **Deep belief networks (** **DBN)**
* Là một **generative graphical model** hoặc thay vào đó là một lớp **Deep neural network**  và bao gồm nhiều lớp **latent variables** và với các kết nối giữa các lớp nhưng không phải giữa các đơn vị trong mỗi lớp.
* **Recurrent neural networks :**
  + Là một class của artifical neural networks nơi mà nó có thể kết nói giữa các nút tạo thành một biểu đồ dọc theo một chuỗi thời gian.
  + Điều này cho phép nó thể hiện nhiều các hành vi tạp thời khác nhau.
  + Xuất phát từ **feedforward neural networks**
  + **RNN** có thể sử dụng trạng thái bên trong (memory) để xử lý các chuổi string có độ dài thay đổi theo đầu vào, giúp cho chúng ta có thể áp dụng cho các tác vụ như không nhận dạng , nhận dạng chữ viết tay đã được kết nối , hay nhận dạng giọng nói …
* **Convolutional neural networks:**
* Là một class của **deep neural networks**  thường được sử dụng để phân tích hình ảnh trực quan.
* Và còn được gọi là mạng neural nhân tạo

1. Dữ liệu nhạy cảm trong Machine Learning
   * + 1. Xác định dữ liệu nhạy cảm

Dữ liệu nhạy cảm có thể tồn tại trong môi trường của bạn trong một số tình huống. Các phần sau đây bao gồm năm trường hợp phổ biến nhất và các phương pháp hiện tại bạn có thể sử dụng để xác định dữ liệu nhạy cảm trong mỗi tình huống.

* + - 1. Dữ liệu nhạy cảm trong các cột

Dữ liệu nhạy cảm có thể được giới hạn trong các cột cụ thể trong bộ dữ liệu có cấu trúc. Ví dụ: bạn có thể có một tập hợp các cột chứa tên, họ và địa chỉ gửi thư của người dùng. Trong trường hợp này, bạn xác định cột nào có dữ liệu nhạy cảm, quyết định cách bảo mật chúng và ghi lại các quyết định này.

* + - 1. Dữ liệu nhạy cảm trong bộ dữ liệu dựa trên văn bản phi cấu trúc

Dữ liệu nhạy cảm có thể là một phần của bộ dữ liệu dựa trên văn bản không có cấu trúc và thường có thể được phát hiện bằng các mẫu đã biết. Ví dụ: số thẻ tín dụng trong bảng điểm trò chuyện có thể được phát hiện một cách đáng tin cậy bằng cách sử dụng mẫu biểu thức chính quy thông thường cho số thẻ tín dụng. Các lỗi phát hiện biểu thức thông thường, dẫn đến phân loại sai, có thể được giảm thiểu bằng cách sử dụng các công cụ phức tạp hơn như [API ngăn chặn mất dữ liệu đám mây](https://cloud.google.com/dlp) (DLP).

* + - 1. Dữ liệu nhạy cảm trong dữ liệu phi cấu trúc dạng tự do

Dữ liệu nhạy cảm có thể tồn tại ở dạng dữ liệu phi cấu trúc tự do như báo cáo văn bản, bản ghi âm thanh, hình ảnh hoặc biên lai được quét. Các bộ dữ liệu này làm cho việc xác định dữ liệu nhạy cảm của bạn khó khăn hơn đáng kể, nhưng có nhiều công cụ có sẵn để giúp bạn.

Đối với các tài liệu văn bản miễn phí, bạn có thể sử dụng hệ thống xử lý ngôn ngữ tự nhiên như [API ngôn ngữ tự nhiên trên đám mây](https://cloud.google.com/natural-language) để xác định các thực thể, địa chỉ email và dữ liệu nhạy cảm khác.

Đối với bản ghi âm, bạn có thể sử dụng dịch vụ chuyển lời nói thành văn bản như [API](https://cloud.google.com/speech) của [Cloud Speech](https://cloud.google.com/speech) và sau đó áp dụng bộ xử lý ngôn ngữ tự nhiên.

Đối với hình ảnh, bạn có thể sử dụng dịch vụ phát hiện văn bản như [API](https://cloud.google.com/vision) của [Cloud Vision](https://cloud.google.com/vision) để tạo văn bản thô từ hình ảnh và cách ly vị trí của văn bản đó trong hình ảnh. API có thể cung cấp tọa độ cho vị trí của một số mục được nhắm mục tiêu trong hình ảnh và bạn có thể sử dụng thông tin này để che giấu tất cả các khuôn mặt từ hình ảnh của dòng máy tính tiền trước khi đào tạo mô hình học máy để ước tính thời gian chờ đợi trung bình của khách hàng.

Đối với video, bạn có thể phân tích từng video thành các khung hình riêng lẻ và coi chúng là tệp hình ảnh hoặc bạn có thể sử dụng công cụ xử lý video như [API thông minh](https://cloud.google.com/video-intelligence) trên nền tảng đám mây cùng với [API](https://cloud.google.com/video-intelligence) của Cloud Speech để xử lý âm thanh.

Những kỹ thuật này vẫn phải được xem xét và phê duyệt tư vấn pháp lý của riêng bạn và tùy thuộc vào mức độ hệ thống của bạn có thể xử lý văn bản miễn phí, sao chép âm thanh, hiểu hình ảnh và phân đoạn video để xác định dữ liệu nhạy cảm tiềm năng. Các API Google được liệt kê ở trên, cũng như Cloud DLP, là những công cụ mạnh mẽ mà bạn có thể kết hợp vào đường dẫn tiền xử lý của mình. Tuy nhiên, các phương pháp tự động này là không hoàn hảo và bạn sẽ muốn xem xét việc duy trì chính sách quản trị để đối phó với bất kỳ thông tin nhạy cảm nào còn sót lại sau khi chà.

* + - 1. Dữ liệu nhạy cảm trong sự kết hợp của các lĩnh vực

Dữ liệu nhạy cảm có thể tồn tại dưới dạng kết hợp của các trường hoặc biểu hiện từ một xu hướng trong trường được bảo vệ theo thời gian. Ví dụ, một thông lệ tiêu chuẩn để giảm khả năng xác định người dùng là làm mờ hai chữ số mã zip cuối cùng, giảm mã zip từ năm xuống còn ba ("zip3"). Tuy nhiên, sự kết hợp của zip3 liên quan đến công việc và zip3 được liên kết với địa chỉ nhà có thể đủ để xác định người dùng có các kết hợp công việc tại nhà bất thường. Tương tự, xu hướng địa chỉ nhà zip3 theo thời gian có thể đủ để xác định một cá nhân đã di chuyển nhiều lần.

Xác định xem một tập dữ liệu có thực sự được bảo vệ khi đối mặt với một cuộc tấn công phân tích tần số hay không đòi hỏi phải có chuyên môn thống kê. Bất kỳ kịch bản nào phụ thuộc vào các chuyên gia của con người đều đưa ra các thách thức về khả năng mở rộng và có thể yêu cầu nghịch lý cùng một kỹ sư dữ liệu kiểm tra dữ liệu để kiểm tra dữ liệu thô cho các vấn đề tiềm ẩn. Lý tưởng nhất, bạn sẽ tạo ra các cách tự động để xác định và định lượng rủi ro này, một nhiệm vụ nằm ngoài phạm vi của bài viết này.

Bất kể, bạn nên làm việc với các kỹ sư dữ liệu và tư vấn pháp lý của bạn để đánh giá rủi ro của bạn trong các tình huống này.

* + - 1. Dữ liệu nhạy cảm trong nội dung phi cấu trúc

Dữ liệu nhạy cảm đôi khi tồn tại trong nội dung phi cấu trúc do thông tin theo ngữ cảnh được nhúng. Ví dụ: bảng điểm trò chuyện có thể bao gồm cụm từ "Tôi đã gọi từ ngày hôm qua từ văn phòng của tôi. Tôi phải đến sảnh tầng mười tám bởi Cafe Deluxe Espresso, vì tầng bốn có khả năng tiếp nhận di động kém."

Dựa trên bối cảnh và phạm vi của dữ liệu đào tạo và lời khuyên của cố vấn pháp lý của bạn, bạn có thể muốn lọc các khía cạnh của nội dung này. Do tính chất không có cấu trúc và tập hợp lớn các cụm từ phức tạp có thể cho phép suy luận tương tự, đây là một kịch bản khó giải quyết với các công cụ lập trình, nhưng đáng để xem xét quản trị chặt chẽ hơn về quyền truy cập vào toàn bộ dữ liệu phi cấu trúc.

Để phát triển mô hình, thường có hiệu quả khi lấy một mẫu của dữ liệu này đã được một người đáng tin cậy xem xét và xem xét và làm cho nó có sẵn để phát triển mô hình. Sau đó, bạn có thể sử dụng các hạn chế bảo mật và tự động hóa phần mềm để xử lý bộ dữ liệu đầy đủ thông qua quy trình đào tạo mô hình sản xuất.

1. Các tấn công trong Machine Learning và Deep Learning lên dữ liệu.
   * + 1. Phân loại tấn công gây mất an toàn.
   1. Causative Attack.

Trong cuộc tấn công nguyên nhân được sử dụng để làm giảm hiệu suất và độ tin cậy của các quy trình đào tạo, thuật toán học máy đã cung cấp dữ liệu đào tạo không chính xác sau khi sửa đổi trong nhãn của các mẫu không nằm trong giới hạn quyết định. Nhiều nhà nghiên cứu đã thực hiện các cuộc tấn công nguyên nhân vào các hình ảnh và tiết lộ rằng nó làm giảm rõ rệt hiệu suất của giai đoạn đào tạo.

Điều này có nghĩa là các đối thủ có khả năng thay đổi đầu vào của dữ liệu đào tạo, điều này trở thành nguyên nhân của sự thay đổi các tham số của các mô hình học tập trong quá trình tái chế, dẫn đến giảm đáng kể việc trình bày các công việc trong các nhiệm vụ phân loại thành công

* 1. Exploratory Attack

Các cuộc tấn công thăm dò về cơ bản không ảnh hưởng đến một tập dữ liệu đào tạo. Mục tiêu chính của các cuộc tấn công khám phá là để có được kiến thức liên quan đến thuật toán học tập càng nhiều càng tốt về hệ thống cơ bản. Tấn công xâm lược mô hình, trích xuất mô hình và suy luận thành viên là những ví dụ về các cuộc tấn công khám phá.

* + Các mỗi đe dọa về tính bảo mật trong Deep Learning có thể chia thành 3 phần.
    - **Integrity Attack.** Cuộc tấn công toàn vẹn
    - **Availability Attack.** Cuộc tấn công sẵn có trái ngược với cuộc tấn công toàn vẹn.
    - **Privacy Violation Attack** . Trong cuộc tấn công vi phạm quyền riêng tư.
  1. Targeted Attack

Nó rất nguy hiểm và nó trực tiếp làm giảm hiệu suất của bộ phân loại trong một mẫu cụ thể hoặc một bộ của một trong các mẫu**.**

* 1. Indiscriminate Attack

Một cuộc tấn công Indiscriminate là tiểu loại của cuộc tấn công ngộ độc. Mục tiêu chính của kẻ tấn công là tăng lỗi phân loại chung. Hơn nữa, cuộc tấn công bừa bãi luôn chọn một giá trị ngẫu nhiên từ mẫu đào tạo. Nó ngẫu nhiên thất bại phân loại.

* + - 1. Các kiểu tấn công tấn công bất lợi .

Mặc dù Deep Learning trở nên thành công để thu hút sự chú ý của ngành công nghiệp, những thách thức về bảo mật và quyền riêng tư của nó, thật không may, nó không thể có được sự chú ý đầy đủ như nó cần phải có.

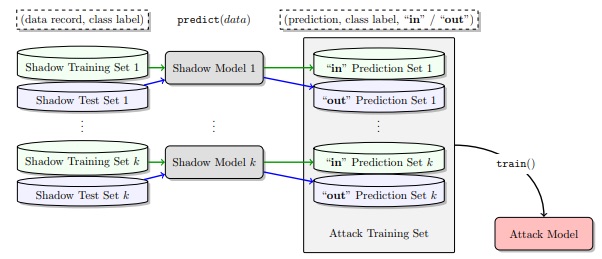
Trong quá trình nghiên cứu, nhiều loại tấn công nhắm vào các ứng dụng DL và chứa các cuộc tấn công DoS, tấn công trốn tránh và tấn công chấm dứt hữu cơ được tiết lộ. Mặc dù tất cả các cuộc tấn công này đều khác nhau về bản chất và về các mục tiêu tấn công của chúng, các nguồn tấn công của kẻ tấn công trong các ứng dụng Deep Learning về cơ bản là từ ba góc độ sau**.**

Các cuộc tấn công bảo mật chống lại các hệ thống học máy , chẳng hạn như các cuộc **tấn công suy luận thành viên** và các **cuộc tấn công đảo ngược mô hình**, có thể làm lộ thông tin cá nhân hoặc nhạy cảm. Một số cuộc tấn công không yêu cầu truy cập trực tiếp vào mô hình nhưng có thể được sử dụng so với API mô hình. Các mô hình được cá nhân hóa, như văn bản dự đoán, có thể phơi bày thông tin có độ nhạy cao

* + - * 1. Tấn công suy luận thành viên

Điều gì sẽ xảy ra nếu Attacker có thể xác định liệu dữ liệu của bạn có được sử dụng để huấn luyện một mô hình với độ chính xác 70-90% hay không, ngay cả khi Attacker không có thông tin đầy đủ? Với sự ra đời của việc học máy tinh vi hơn và cũng được cá nhân hóa hoặc được nhóm lại trở nên phổ biến, các cuộc tấn công này có thể phơi bày thông tin nhạy cảm về các cá nhân.

Kiểu tấn công này được gọi là Tấn công suy luận thành viên (MIA), và nó được tạo ra bởi Giáo sư Reza Shokri



Đầu tiên, dữ liệu huấn luyện đầy đủ phải được thu thập từ chính mô hình thông qua các truy vấn tuần tự của các đầu vào có thể hoặc được thu thập từ các bộ dữ liệu công khai hoặc riêng tư mà kẻ tấn công có quyền truy cập. Sau đó, kẻ tấn công xây dựng một số mô hình bóng - mô phỏng theo mô hình (nghĩa là lấy đầu vào và đầu ra tương tự của mô hình đích). Các mô hình bóng này phải được điều chỉnh cho độ chính xác cao và thu hồi trên các mẫu dữ liệu huấn luyện được thu thập. Lưu ý: cuộc tấn công nhằm mục đích phân chia thử nghiệm và đào tạo khác nhau cho mỗi mô hình bóng, do đó bạn phải có đủ dữ liệu để thực hiện bước này.

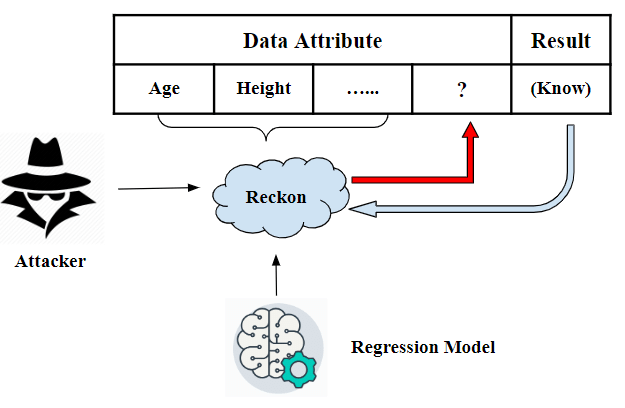
Cuối cùng, trong một kiến trúc có phần giống với một mạng lưới nghịch cảnh generative adversarial network (GAN), Kẻ tấn công có thể huấn luyện một kẻ phân biệt đối xử tìm hiểu sự khác biệt về đầu ra giữa dữ liệu huấn luyện đã thấy và dữ liệu kiểm tra hoặc xác thực chưa thấy dựa trên đầu vào và đầu ra cho nhiều mô hình bóng. Bộ phân biệt đối xử này sau đó được sử dụng để đánh giá API mục tiêu và xác định xem một điểm dữ liệu nằm trong hoặc ngoài mô hình đích. Do đó, các cuộc tấn công này có thể được chạy mà không có quyền truy cập đầy đủ vào mô hình hoặc trong môi trường được gọi là môi trường Hộp đen.

Những dữ liệu nào các cuộc tấn công có thể phơi bày? Đối với AI trong chăm sóc sức khỏe, nó có thể phơi bày thứ gì đó nhạy cảm như tình trạng HIV. AI trong điều trị sức khỏe thường phân tách các cá nhân dựa trên bệnh của họ. Nếu kẻ tấn công có thể truy vấn API của một mô hình khuyến nghị điều trị cho bệnh nhân HIV, việc tiết lộ rằng họ là một phần của dữ liệu huấn luyện cũng cho chúng tôi biết tình trạng HIV dương tính của họ. Khi việc sử dụng AI phát triển trong ngành chăm sóc sức khỏe, chúng ta cần phải nhận thức được các lỗ hổng bảo mật và quyền riêng tư trong các mô hình chúng ta đào tạo và sử dụng.

* 1. Tấn công đảo ngược mô hình(Model Inversion Attacks)

Điều gì sẽ xảy ra nếu Attacker có thể học được khuôn mặt của bạn chỉ bằng cách biết tên của bạn? Bạn có thể nghĩ rằng Attacker đang sử dụng Google Image Search, nhưng thực tế, thông tin này được tiết lộ chỉ bằng cách truy cập vào một mô hình nhận dạng khuôn mặt đã được đào tạo để nhận diện khuôn mặt của bạn.

Kiểu tấn công này được gọi là tấn công đảo ngược mô hình và nó được tạo ra bởi Matt Fredrikson và các nhà nghiên cứu.



Attack model of model inversion attack

Tương tự như cuộc tấn công suy luận thành viên, cuộc tấn công này chỉ cần truy cập API vào mô hình học máy và có thể được chạy dưới dạng một loạt các truy vấn tiến bộ. Để bắt đầu, kẻ tấn công có thể sử dụng hình ảnh cơ sở - nếu họ biết bất cứ điều gì về mục tiêu của họ (chẳng hạn như tuổi tác, giới tính hoặc chủng tộc), họ có thể cố gắng chọn một hình ảnh gần hơn. Sau đó, kẻ tấn công chạy một loạt các truy vấn thực hiện tấn công đảo ngược - thay đổi pixel để tăng độ chính xác hoặc độ tin cậy của hệ thống máy học. Tại một số điểm, độ tin cậy cao đạt được, tạo ra một hình ảnh tương tự như hình ảnh được hiển thị ở trên - mặc dù không hoàn hảo, nhưng khá rõ ràng về cách người đó nhìn.

Làm thế nào hợp lý là cuộc tấn công này? Bây giờ chúng ta thấy việc sử dụng thường xuyên các hệ thống nhận dạng khuôn mặt, bao gồm cả việc sử dụng cơ sở dữ liệu liên bang và nhập cư Hoa Kỳ được chia sẻ cho những việc như đăng ký hàng không, có nghĩa là việc truy cập và sử dụng các hệ thống này cũng đang tăng lên. Với mỗi nhà thầu hoặc nhà thầu phụ có quyền truy cập vào hệ thống hoặc API, sẽ có một cuộc tấn công tiềm năng khác ở đường chân trời. Cùng với vi phạm quyền riêng tư đi kèm với cuộc tấn công này, quyền truy cập vào hộ chiếu, thị thực hoặc các hình ảnh quan trọng và chính thức khác có thể là mối đe dọa an ninh tiềm ẩn.

* 1. Công cụ chia sẻ bí mật: đánh giá và kiểm tra ghi nhớ ngoài ý muốn trong các neural networks

Điều gì sẽ xảy ra nếu Attacker có thể thu thập số thẻ tín dụng của bạn chỉ với quyền truy cập vào một mô hình dựa trên văn bản được cá nhân hóa hoặc bán cá nhân hóa? Bạn có thể nghĩ về điều này giống như một bàn phím dự đoán, điều gì xảy ra nếu bạn gõ “My phone number is…”?

Kiểu tấn công này đã được điều tra bởi Nicholas Carlini và một số nhà nghiên cứu học thuật và não bộ khác của Google trong một bài báo có tên The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets. Không giống như hai cuộc tấn công trước, cuộc tấn công này đòi hỏi quyền truy cập hộp trắng vào mô hình - vì vậy kẻ tấn công cần phải có được mô hình được đào tạo hoặc chính nhị phân. Sau đó, họ tiến hành bằng cách truy vấn mô hình văn bản để tối đa hóa khả năng nhân vật tiếp theo khám phá các mẫu đã biết (như số ID, số thẻ tín dụng hoặc số điện thoại).

Với việc tăng cường sử dụng học sâu trong văn bản, đặc biệt là AI đàm thoại trong dịch vụ khách hàng, khả năng ghi nhớ và trích xuất bí mật là một rủi ro bảo mật thực sự. Về cơ bản, nó cho phép một cuộc tấn công lừa đảo chống lại bất kỳ người dùng nào đã từng chia sẻ thông tin cá nhân với người máy.

Tiết lộ bí mật (Disclosure of secrets) là mối quan tâm đặc biệt trong các mô hình mạng thần kinh phân loại hoặc dự đoán chuỗi văn bản ngôn ngữ tự nhiên, ngay cả khi văn bản dữ liệu đào tạo nhạy cảm hoặc riêng tư là rất hiếm, người ta nên cho rằng các mô hình được đào tạo tốt đã chú ý đến các chi tiết chính xác của nó. Những người sử dụng các mô hình như vậy có thể phát hiện ra tình cờ, do tình cờ hoặc do mục đích mà việc nhập các tiền tố văn bản nhất định khiến các mô hình xuất ra sự tiết lộ văn bản một cách đáng ngạc nhiên.

1. Kết luận