**INTRO**

Trong kỷ nguyên số, internet và mạng xã hội đã trở thành mảnh đất màu mỡ cho tin giả lan rộng, gây ảnh hưởng sâu sắc đến nhận thức cộng đồng.

Sự tiến bộ của công nghệ, đặc biệt là trí tuệ nhân tạo và các công cụ kỹ thuật số, càng khiến tin giả trở nên phức tạp, tinh vi và khó nhận diện. Điều này đặt ra một thách thức lớn cho cả người tiêu dùng thông tin và các cơ quan truyền thông trong cuộc chiến chống lại thông tin sai lệch.

Chính vì vậy dự án của chúng tôi đặt mục tiêu xây dựng một nền tảng kỹ thuật dựa trên những nghiên cứu khoa học hiện đại, hướng tới phát triển một mô hình có khả năng nhận diện và phân biệt tin tức thật - giả. Qua đó, chúng tôi hy vọng góp phần tạo nên một không gian truyền thông đáng tin cậy hơn, giảm thiểu những tác động tiêu cực của tin giả đối với xã hội.

**Phạm vi nghiên cứu**

Phạm vi nghiên cứu của dự án tập trung vào việc giải quyết một trong những thách thức lớn của thời đại số: vấn đề tin giả. Mục tiêu của chúng tôi là phát triển một hệ thống phân loại tin tức giả dựa trên mô hình mạng nơ-ron tích chập (CNN). Dưới đây là các bước chính trong nghiên cứu:

*1. Sử dụng mô hình CNN:*

- Chúng tôi chọn mô hình CNN vì khả năng mạnh mẽ trong việc nhận diện các đặc trưng của văn bản. Các lớp tích chập và pooling trong CNN giúp nhận biết các mẫu thông tin quan trọng, hỗ trợ phân loại tin tức hiệu quả hơn.

*2. Xây dựng và huấn luyện mô hình:*

- Chúng tôi sẽ xây dựng và huấn luyện mô hình CNN trên tập dữ liệu ISOT (Integrated Student Outcomes Toolkit). Quá trình này nhằm giúp mô hình hiểu và phân loại tin tức trên nhiều loại dữ liệu khác nhau, tăng cường độ chính xác trong việc phát hiện tin giả.

*3. Xử lý dữ liệu văn bản:*

- Do tin tức là dữ liệu văn bản, mô hình cần được trang bị các kỹ thuật xử lý ngôn ngữ tự nhiên để hiểu rõ ý nghĩa và phân biệt giữa tin thật và tin giả.

*4. Đánh giá hiệu suất:*

- Hiệu quả của mô hình sẽ được đo lường qua các chỉ số như Độ chính xác (Accuracy), Độ chính xác truy hồi (Precision), Tỉ lệ phát hiện (Recall) và F1-score. Những kết quả này giúp đánh giá khả năng của mô hình trong việc phân loại tin tức và ứng dụng trong thực tế.

*Vd cho Precision:*Trong bài toán phân loại tin giả, nếu mô hình dự đoán một bài viết là **giả**:

* Nếu bài viết thực sự là giả (dự đoán đúng) → đó là một **True Positive**.
* Nếu bài viết thực ra là thật (dự đoán sai) → đó là một **False Positive**.

*Vd cho recall* : Trong bài toán phân loại tin giả, nếu có 100 bài viết giả mà mô hình chỉ phát hiện đúng 80 bài (bỏ sót 20 bài), thì:

* **True Positives** là 80 (dự đoán đúng các bài giả).
* **False Negatives** là 20 (dự đoán sai, bỏ sót các bài giả).

**Mục tiêu‑phương pháp‑dữ liệu**

*Mục tiêu*

Mục tiêu chính của nghiên cứu là phát triển và đánh giá một hệ thống có khả năng xác định tính chính xác của các bài báo thông qua việc đánh giá độ tin cậy của thông tin trong chúng. Hệ thống sẽ áp dụng một cách linh hoạt các phương pháp học máy kết hợp với đánh giá nội dung để đưa ra những nhận định chính xác.

*Phương pháp*

Chúng tôi sẽ sử dụng mô hình CNN (Mạng Nơ-ron Tích chập) để phân loại các bài báo thành hai loại: tin tức thật và tin tức giả, từ đó xác định độ tin cậy của nội dung.

*Dữ liệu*

Bộ dữ liệu đầu vào bao gồm các bài báo thật và giả, được thu thập từ liên kết: [chèn link ở đây].

**Giới thiệu về CNN**

CNN là một loại mạng nơ-ron nhận dữ liệu đầu vào dưới dạng vector hoặc ma trận số. Nó thực hiện các phép toán như cộng và nhân ma trận để phát hiện các đặc trưng nổi bật trong dữ liệu.

Cấu trúc của CNN bao gồm nhiều lớp tích chập, sử dụng các bộ lọc (filters) để nhận diện những đặc điểm quan trọng, đặc biệt hiệu quả với dữ liệu hình ảnh hoặc chuỗi.

* CNN không chỉ giới hạn ở việc xử lý hình ảnh mà còn có thể được áp dụng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và chuỗi thời gian. Trong các ứng dụng NLP, thường sử dụng CNN một chiều (1D) để nhận diện các chuỗi hoặc phát hiện các mẫu trong dữ liệu văn bản.
* Mô hình CNN cho phép điều chỉnh kích thước và số lượng bộ lọc, giúp tối ưu hóa việc phát hiện các đặc trưng tùy thuộc vào loại dữ liệu.

Hình 1 mô phỏng quy trình tích chập của CNN:

* **Input Features**: Dữ liệu đầu vào được trình bày dưới dạng chuỗi hoặc mảng giá trị.
* **Kernel Size**: Kích thước bộ lọc được sử dụng để quét qua dữ liệu đầu vào.
* **Patch from Input Features**: Mỗi lần bộ lọc quét qua một phần nhỏ của dữ liệu đầu vào (patch) và thực hiện tích vô hướng (dot product) giữa bộ lọc và phần dữ liệu này.
* **Output Features**: Sau khi thực hiện phép tích chập, các giá trị kết quả tạo thành bản đồ đặc trưng (feature map), chứa các đặc điểm đã được trích xuất từ dữ liệu đầu vào.

**Nguyên lý hoạt động của CNN**

Mạng CNN bao gồm nhiều lớp tích chập chồng lên nhau, mỗi lớp sử dụng các hạt nhân (kernels) hay bộ lọc (filters) để phát hiện các đặc trưng quan trọng.

Các lớp này áp dụng các hàm kích hoạt phi tuyến như ReLU (Rectified Linear Unit) để giúp mạng học được các đặc trưng phức tạp hơn.

Thông tin từ các lớp trước được truyền đến các lớp sau, cho phép CNN xây dựng các trừu tượng thông tin từ mức độ thấp đến cao.

* **Lớp Tích chập (Convolutions)**: Tại đây, các hạt nhân trượt qua các vùng nhỏ của hình ảnh đầu vào để tạo ra các đặc trưng.
* **Lớp Gộp cực đại (Max Pooling)**: Lớp này giảm kích thước không gian của bản đồ đặc trưng, đồng thời giữ lại những đặc trưng quan trọng nhất.
* **Lớp Liên kết hoàn toàn (Fully Connected)**: Cuối cùng, các đặc trưng sẽ được làm phẳng (flatten) và chuyển qua các lớp liên kết hoàn toàn để đưa ra dự đoán.

Hình 2 mô tả quy trình hoạt động của CNN từ tích chập đến gộp cực đại và lớp liên kết hoàn toàn. Các vùng màu đen và trắng đại diện cho các giá trị sau khi đi qua các bộ lọc và các lớp khác nhau.

Mô hình CNN có khả năng tự động học và phân tích các đặc trưng của dữ liệu nhờ vào việc sử dụng các lớp lọc đa dạng, cho phép phát hiện thông tin từ đơn giản đến phức tạp.

### Hình ảnh 2 mô tả chi tiết quy trình xử lý của mạng nơ-ron tích chập (CNN):

**Dữ liệu đầu vào (Input)**:

* 1. Dữ liệu đầu vào thường là một hình ảnh, được biểu diễn dưới dạng lưới ô vuông đen trắng. Mỗi ô vuông đại diện cho một pixel, với các mức độ màu sắc (đen, trắng, xám) tương ứng với các giá trị cường độ khác nhau.

**Lớp Tích chập (Convolutional Layer)**:

* 1. Lớp tích chập đầu tiên sử dụng các bộ lọc (kernels) để quét qua toàn bộ hình ảnh đầu vào.
  2. Mỗi bộ lọc tập trung vào việc phát hiện các đặc trưng khác nhau, chẳng hạn như cạnh, góc, hoặc các chi tiết cụ thể của hình ảnh. Kết quả của quá trình tích chập tạo ra một tập hợp các "bản đồ đặc trưng" (feature maps) phản ánh các đặc trưng khác nhau.
  3. Trong hình, ta thấy hai bộ lọc với kích thước khác nhau: (16, 2, 2) và (32, 5, 2), đại diện cho số lượng bộ lọc, kích thước của bộ lọc, và bước dịch chuyển (stride).

**Lớp Gộp cực đại (Max Pooling Layer)**:

* 1. Sau lớp tích chập là lớp gộp cực đại, có chức năng giảm kích thước của bản đồ đặc trưng nhưng vẫn bảo toàn các thông tin quan trọng. Lớp này thường lấy giá trị lớn nhất trong từng vùng con, giúp giảm số lượng thông tin mà vẫn giữ lại các đặc trưng chính.
  2. Hình ảnh cho thấy bản đồ đặc trưng sau lớp pooling có kích thước nhỏ hơn so với lớp tích chập, góp phần làm giảm độ phức tạp của mạng và tiết kiệm tài nguyên tính toán.

**Làm phẳng (Flatten)**:

* 1. Sau khi đi qua các lớp tích chập và pooling, các bản đồ đặc trưng sẽ được làm phẳng thành một vector một chiều, để có thể đưa vào các lớp liên kết hoàn toàn.
  2. Bước này chuyển đổi các đặc trưng không gian thành một chuỗi các giá trị mà mạng có thể xử lý trong các lớp liên kết hoàn toàn phía sau.

**Các lớp Liên kết hoàn toàn (Fully-Connected Layers)**:

* 1. Sau khi làm phẳng, các đặc trưng sẽ được đưa vào các lớp liên kết hoàn toàn. Đây là những lớp cuối cùng, hoạt động tương tự như một mạng nơ-ron thông thường để tổng hợp các đặc trưng và đưa ra dự đoán cuối cùng.
  2. Trong hình, các ô vuông nhỏ dần về phía bên phải đại diện cho các lớp liên kết hoàn toàn, dẫn đến một ô đầu ra, biểu thị cho kết quả dự đoán của mô hình.

*Áp dụng mô hình CNN để kiểm tra xác thực các bài báo là tin thật hay là bài báo tin*

*giả , chúng ta sẽ xử lý chữ , mỗi từ hoặc một chuỗi từ có thể xem như một “ ảnh ”*

*một chiều(1D) , với chiều dài là độ dài của câu và chiều rộng là số chiều của biểu*

*diễn từ*

**Lớp Embedding**

Chức năng: Ánh xạ từng từ trong tập dữ liệu thành một vector số thực.

Cách hoạt động: Mỗi từ được đại diện bằng một vector với kích thước cố định (ví dụ: 100 hoặc 300 chiều). Các vector này được học trong quá trình huấn luyện, giúp mô hình hiểu rõ mối quan hệ ngữ nghĩa giữa các từ. Việc sử dụng lớp embedding giúp giảm thiểu độ phức tạp của dữ liệu đầu vào và cung cấp một cách biểu diễn phù hợp cho các từ.

**Lớp Convolutional**

- Chức năng: Quét qua chuỗi từ để trích xuất các đặc trưng.

- Cách hoạt động: Các bộ lọc (kernels) di chuyển qua các từ (hoặc nhóm từ) trong chuỗi, thực hiện các phép toán tích chập để tìm kiếm các mẫu (patterns) đặc trưng trong văn bản. Mỗi bộ lọc có thể nhận diện các đặc trưng khác nhau như cụm từ thông dụng, ngữ cảnh, hoặc cấu trúc câu.

**Lớp Max Pooling**

- Chức năng: Giảm kích thước của đầu ra và tăng tính trừu tượng.

- Cách hoạt động: Sau khi lớp convolutional, lớp max pooling sẽ lấy giá trị lớn nhất từ các vùng nhất định trong đầu ra. Điều này không chỉ giúp giảm số lượng thông tin mà mô hình phải xử lý, mà còn giúp làm nổi bật các đặc trưng quan trọng hơn, tăng cường khả năng tổng quát của mô hình.

**Lớp Flatten**

- Chức năng: Chuyển đổi đầu ra từ các bản đồ đặc trưng (feature maps) thành một vector một chiều (1D).

- Cách hoạt động: Các đặc trưng đã được trích xuất qua các lớp trước đó được làm phẳng thành một vector. Điều này cho phép các lớp sau (fully connected) nhận đầu vào dưới dạng một vector duy nhất, tạo điều kiện cho việc xử lý và phân tích tiếp theo.

**Lớp Fully Connected (Dense)**

- Chức năng: Học các mối quan hệ giữa các đặc trưng.

- Cách hoạt động: Lớp này kết nối tất cả các nơ-ron từ lớp trước với các nơ-ron trong lớp này. Các nơ-ron trong lớp fully connected sẽ tính toán và học các mối quan hệ phức tạp giữa các đặc trưng đã được trích xuất, nhằm tối ưu hóa khả năng dự đoán của mô hình.

**Lớp Dropout**

- Chức năng: Giảm nguy cơ quá mức học (overfitting) và cải thiện khả năng tổng quát hóa.

- Cách hoạt động: Trong quá trình huấn luyện, một tỷ lệ phần trăm của các nơ-ron sẽ bị "bỏ qua" (không tham gia vào quá trình tính toán). Điều này giúp ngăn chặn mô hình trở nên quá phù hợp với dữ liệu huấn luyện, đảm bảo rằng mô hình có khả năng hoạt động tốt với dữ liệu chưa thấy (test data).

**Lớp Output**

- Chức năng: Xuất ra xác suất cho mỗi lớp (giả mạo hoặc không giả mạo).

- Cách hoạt động: Lớp này thường sử dụng hàm kích hoạt softmax (hoặc sigmoid cho bài toán nhị phân) để chuyển đổi đầu ra của mô hình thành xác suất. Kết quả này cho phép mô hình đưa ra quyết định cuối cùng về việc xác thực bài báo, xác định xem nó có phải là tin thật hay tin giả.

**Tóm lại**

Mô hình CNN với các lớp như embedding, convolutional, pooling, flatten, fully connected, dropout và output, tạo ra một quy trình hiệu quả để xử lý và phân tích văn bản. Quy trình này không chỉ giúp trích xuất đặc trưng mà còn tối ưu hóa khả năng dự đoán của mô hình, từ đó tăng cường độ chính xác trong việc xác định tính xác thực của các bài báo.