Lời nói đầu

DeepSeek-OCR: Phương pháp nén hình ảnh dựa trên ngữ cảnh

Haoran Wei, Yaofeng Sun, Yukun Li

DeepSeek-AI

Tóm tắt

Chúng tôi giới thiệu DeepSeek-OCR như một nghiên cứu sơ bộ về khả năng nén các đoạn văn dài thông qua phương pháp lập bản đồ 2D quang học. DeepSeek-OCR bao gồm hai thành phần chính: DeepEncoder và DeepSeek3B-MoE-A570M, trong đó DeepEncoder đóng vai trò là công cụ xử lý cốt lõi – được thiết kế để duy trì mức độ hoạt động thấp ngay cả với dữ liệu đầu vào có độ phân giải cao, đồng thời đạt được tỷ lệ nén cao nhằm đảm bảo số lượng các "token hình ảnh" được sử dụng ở mức tối ưu và dễ quản lý. Các thí nghiệm cho thấy khi số lượng "token văn bản" chỉ bằng khoảng 10 lần số lượng "token hình ảnh" (tức tỷ lệ nén dưới 10×), mô hình vẫn có thể đạt độ chính xác trong việc giải mã văn bản lên đến 97%; ngay cả ở tỷ lệ nén 20×, độ chính xác này vẫn khoảng 60%. Điều này mở ra nhiều tiềm năng cho các lĩnh vực nghiên cứu như việc nén các đoạn văn dài có tính lịch sử hoặc cơ chế "quên lãng" dữ liệu trong các mô hình LLM. Ngoài ra, DeepSeek-OCR còn thể hiện giá trị thực tiễn cao: trên nền tảng OmniDocBench, nó vượt trội hơn so với GOT-OCR2.0 (sử dụng 256 "token hình ảnh/trang") khi chỉ cần 100 "token hình ảnh"; đồng thời cũng vượt qua MinerU2.0 (trung bình sử dụng hơn 6000 "token hình ảnh/trang") với ít hơn 800 "token hình ảnh". Trong thực tế sản xuất, DeepSeek-OCR có thể tạo ra lượng dữ liệu huấn luyện lớn (hơn 200.000 trang mỗi ngày) cho các mô hình LLM/VLM, chỉ với một máy tính A100-40G. Mã nguồn và thông tin về cấu hình mô hình có thể được tìm thấy tại địa chỉ http://github.com/deepseek-ai/DeepSeek-OCR.

Hình 1: Hình (a) thể hiện tỷ lệ nén (số lượng các ký hiệu văn bản trong dữ liệu đích chuẩn so với số lượng các ký hiệu hình ảnh được mô hình sử dụng) khi thử nghiệm trên bài kiểm thử Fox [21]; Hình (b) so sánh hiệu suất của DeepSeek-OCR trên nền tảng OmniDocBench [27]. DeepSeek-OCR đạt được mức hiệu suất tiên tiến nhất trong số các mô hình end-to-end, trong khi chỉ sử dụng số lượng ký hiệu hình ảnh tối thiểu.

1. Giới thiệu

Các mô hình ngôn ngữ lớn hiện nay đối mặt với những thách thức lớn về mặt tính toán khi xử lý các nội dung văn bản dài, do mức độ tốn kém tăng theo cấp số nhân theo độ dài chuỗi văn bản. Chúng tôi đề xuất một giải pháp tiềm năng: tận dụng phương thức biểu diễn thông tin dưới dạng hình ảnh như một phương tiện nén hiệu quả cho dữ liệu văn bản. Một hình ảnh duy nhất chứa nội dung văn bản có thể truyền tải được nhiều thông tin hơn so với việc sử dụng văn bản kỹ thuật số tương đương, điều này cho thấy rằng phương pháp nén thông qua hình ảnh có thể mang lại tỷ lệ nén cao hơn đáng kể.

Quan điểm này thúc đẩy chúng ta xem xét lại các mô hình ngôn ngữ hình ảnh (Vision-Language Models – VLMs) từ góc độ tập trung vào các mô hình LLM, tìm hiểu cách các bộ mã hóa hình ảnh có thể nâng cao hiệu quả xử lý thông tin văn bản của các mô hình LLM, thay vì chỉ giúp chúng thực hiện các nhiệm vụ đơn giản như trả lời câu hỏi dạng trắc nghiệm (VQA) – lĩnh vực mà con người giỏi nhất. Các nhiệm vụ OCR, với vai trò là phương thức kết nối giữa hình ảnh và ngôn ngữ, tạo ra môi trường thí nghiệm lý tưởng để kiểm tra mô hình này; chúng thiết lập mối liên hệ tự nhiên giữa các biểu diễn hình ảnh và văn bản, đồng thời cung cấp các chỉ số đánh giá định lượng cần thiết.

Do đó, chúng tôi giới thiệu DeepSeek-OCR – một hệ thống VLM được thiết kế như một bằng chứng khái niệm ban đầu cho việc nén dữ liệu hình ảnh và văn bản một cách hiệu quả. Công trình của chúng tôi đóng góp ba điểm chính sau:

Trước hết, chúng tôi cung cấp phân tích định lượng toàn diện về tỷ lệ nén các ký hiệu hình ảnh và văn bản. Phương pháp của chúng tôi đạt độ chính xác khi giải mã dữ liệu nén lên đến 96% ở mức độ nén 9–10 lần; độ chính xác khoảng 90% ở mức độ nén 10–12 lần; và khoảng 60% ở mức độ nén 20 lần, dựa trên các bài kiểm thử Fox [21] với nhiều kiểu bố cục tài liệu khác nhau. Độ chính xác thực tế còn cao hơn nữa nếu xét đến sự khác biệt về định dạng giữa dữ liệu nén và dữ liệu chuẩn, như được minh họa trong Hình 1(a). Kết quả này cho thấy các mô hình ngôn ngữ có kích thước lớn có thể học cách giải mã các dữ liệu hình ảnh đã được nén một cách hiệu quả; điều này gợi ý rằng các mô hình LLM lớn hơn cũng có thể dễ dàng đạt được khả năng tương tự thông qua việc thiết kế quá trình huấn luyện ban đầu một cách phù hợp.

Thứ hai, chúng tôi giới thiệu DeepEncoder – một kiến trúc mới mẻ giúp giảm thiểu lượng bộ nhớ cần sử dụng và số lượng các thông tin đầu vào liên quan đến hình ảnh ngay cả với những dữ liệu có độ phân giải cao. Cấu trúc này kết nối các thành phần xử lý thông tin hình ảnh theo thứ tự: bộ xử lý chú ý theo khung cụ thể và bộ xử lý chú ý toàn diện, thông qua một bộ thu nhỏ dữ liệu dạng hình ảnh có cấu trúc lưới tích chập. Thiết kế này đảm bảo rằng bộ xử lý chú ý theo khung cụ thể có thể xử lý một lượng lớn thông tin hình ảnh, trong khi bộ thu nhỏ này giúp giảm bớt số lượng dữ liệu đó trước khi chúng được truyền vào bộ xử lý chú ý toàn diện, từ đó giúp tiết kiệm bộ nhớ và tăng hiệu quả xử lý dữ liệu.

Thứ ba, chúng tôi đã phát triển công cụ DeepSeek-OCR dựa trên các mô hình DeepEncoder và DeepSeek3B-MoE [19, 20]. Như được minh họa trong Hình 1(b), công cụ này đạt được hiệu suất hàng đầu trong số các mô hình end-to-end trên bộ dữ liệu OmniDocBench, trong khi sử dụng số lượng token liên quan đến xử lý hình ảnh tối thiểu. Ngoài ra, chúng tôi còn trang bị cho mô hình này khả năng phân tích các biểu đồ, công thức hóa học, hình dạng hình học đơn giản và hình ảnh tự nhiên, nhằm nâng cao tính ứng dụng thực tiễn của nó. Trong môi trường sản xuất, DeepSeek-OCR có thể tạo ra 33 triệu trang dữ liệu mỗi ngày để phục vụ các mô hình LLM hoặc VLM, bằng cách sử dụng 20 node – mỗi node được trang bị 8 GPU loại A100-40G.

Tóm lại, nghiên cứu này đề xuất một phương pháp tiếp cận mới, sử dụng công nghệ thị giác như một phương tiện nén hiệu quả để xử lý thông tin văn bản trong các mô hình ngôn ngữ lớn. Thông qua công cụ DeepSeek-OCR, chúng tôi đã chứng minh rằng việc kết hợp công nghệ thị giác và văn bản có thể giúp giảm số lượng các ký hiệu được sử

dụng trong quá trình xử lý đáng kể (từ 7 đến 20 lần), từ đó mở ra hướng đi triển vọng để giải quyết những thách thức liên quan đến việc xử lý dữ liệu có bối cảnh dài trong các mô hình ngôn ngữ lớn. Phân tích định lượng của chúng tôi cung cấp những hướng dẫn thực tiễn cho việc tối ưu hóa cách phân bổ các ký hiệu trong các mô hình này, trong khi kiến trúc DeepEncoder được đề xuất đã chứng tỏ tính khả thi thực tế của phương pháp này thông qua các ứng dụng thực tế. Mặc dù nghiên cứu này tập trung vào công nghệ OCR như một ví dụ minh họa ý tưởng, nhưng nó mở ra nhiều khả năng mới để xem xét cách kết hợp hiệu quả giữa công nghệ thị giác và ngôn ngữ, nhằm nâng cao hiệu suất xử lý dữ liệu văn bản quy mô lớn trong các hệ thống ứng dụng thực tế.

Hình 2 | Các bộ mã hóa hình ảnh điển hình được sử dụng trong các hệ thống trí tuệ nhân tạo dựa trên hình ảnh phổ biến hiện nay. Dưới đây là ba loại bộ mã hóa thường được dùng trong các hệ thống mở nguồn này; tuy nhiên, tất cả chúng đều có những hạn chế riêng.

2. Các công trình liên quan

2.1. Các bộ mã hóa hình ảnh điển hình trong các mô hình trí tuệ nhân tạo dựa trên hình ảnh

Các hệ thống mã hóa hình ảnh mã nguồn mở hiện nay sử dụng ba loại cấu trúc chính để xử lý dữ liệu hình ảnh, như được minh họa trong Hình 2. Loại đầu tiên là kiến trúc gồm hai thành phần, điển hình là công cụ Vary [36], vốn sử dụng các bộ mã hóa SAM [17] hoạt động song song để tăng số lượng tham số hình ảnh, từ đó nâng cao khả năng xử lý các hình ảnh có độ phân giải cao. Mặc dù phương pháp này mang lại khả năng điều chỉnh các tham số và sử dụng ít bộ nhớ hoạt động, nhưng nó gặp phải nhiều hạn chế: quy trình xử lý hình ảnh trước khi mã hóa phức tạp hơn, và việc sử dụng song song các bộ mã hóa trong quá trình huấn luyện gặp nhiều khó khăn. Loại thứ hai là phương pháp dựa trên việc chia hình ảnh thành các khối nhỏ để xử lý song song, điển hình là InternVL2.0 [8]; phương pháp này giúp giảm lượng bộ nhớ cần thiết khi xử lý hình ảnh độ phân giải cao. Tuy có thể xử lý các hình ảnh có độ phân giải rất cao, nhưng phương pháp này vẫn có những hạn chế do độ phân giải của các bộ mã hóa thường thấp (dưới 512×512), khiến việc xử lý các hình ảnh lớn trở nên phức tạp và tạo ra nhiều tham số hình ảnh. Loại thứ ba là phương pháp mã hóa với độ phân giải linh hoạt, điển hình là Qwen2-VL [35]; phương pháp này áp dụng công nghệ NaViT [10] để xử lý trực tiếp toàn bộ hình ảnh mà không cần chia chúng thành các khối nhỏ. Mặc dù có khả năng thích ứng với nhiều độ phân giải khác nhau, nhưng phương pháp này gặp nhiều khó khăn khi xử lý các hình ảnh lớn do lượng bộ nhớ cần thiết quá lớn, có thể dẫn đến tình trạng quá tải bộ nhớ GPU; đồng thời, quá trình huấn luyện cũng đòi hỏi thời gian dài hơn do số lượng tham số hình ảnh tạo ra quá nhiều. Những yếu tố này sẽ làm chậm quá trình xử lý dữ liệu và tạo ra hiệu suất thấp hơn trong quá trình thực hiện các tác vụ liên quan đến hình ảnh.

2.2. Các mô hình OCR hoạt động theo nguyên lý từ đầu đến cuối

OCR, đặc biệt là các công cụ phân tích văn bản từ hình ảnh, luôn là chủ đề được quan tâm sâu rộng trong lĩnh vực chuyển đổi hình ảnh thành văn bản. Với sự phát triển mạnh

mẽ của các mô hình trí tuệ nhân tạo, ngày càng nhiều mô hình OCR hoạt động theo nguyên lý "từ đầu đến cuối" đã xuất hiện, giúp thay đổi đáng kể cấu trúc truyền thống của các hệ thống OCR (vốn yêu cầu sử dụng các mô hình riêng biệt để phát hiện và nhận diện thông tin). Mô hình Nougat [6] là một ví dụ điển hình; nó đã áp dụng cấu trúc này để phân tích văn bản các bài báo khoa học trên arXiv, chứng minh tiềm năng của các mô hình này trong việc xử lý các nhiệm vụ phức tạp liên quan đến nhận diện hình ảnh. Mô hình GOT-OCR2.0 [38] mở rộng phạm vi ứng dụng của OCR2.0 sang nhiều loại tác vụ phân tích hình ảnh khác và thiết kế một mô hình OCR với sự cân bằng giữa hiệu suất và tính hiệu quả, từ đó làm nổi bật thêm tiềm năng của các nghiên cứu về OCR theo nguyên lý "từ đầu đến cuối". Ngoài ra, các mô hình thị giác tổng quát như series Qwen-VL [35], series InternVL [8] và nhiều phiên bản cải tiến của chúng cũng liên tục được nâng cấp để tăng cường khả năng phân tích văn bản từ hình ảnh. Tuy nhiên, vẫn còn một câu hỏi quan trọng mà các mô hình hiện tại chưa giải đáp được: Đối với một văn bản gồm 1000 từ, ít nhất cần bao nhiêu thông tin thị giác là đủ để giải mã nó? Câu hỏi này có ý nghĩa quan trọng đối với các nghiên cứu liên quan đến nguyên tắc "Một hình ảnh giá trị hơn ngàn lời nói".

Hình 3 | Cấu trúc của hệ thống DeepSeek-OCR. DeepSeek-OCR bao gồm một bộ mã hóa sâu (DeepEncoder) và một bộ giải mã sâu DeepSeek-3B-MoE. Bộ mã hóa sâu là thành phần cốt lõi của hệ thống này, bao gồm ba thành phần chính: công cụ SAM [17] dùng để thu thập thông tin dựa trên cơ chế chú ý theo khung cửa sổ; công cụ CLIP [29] dùng để xử lý kiến thức thông qua cơ chế chú ý toàn diện; và một bộ nén dữ liệu gồm 16 thành phần, có vai trò kết nối hai thành phần trên lại với nhau.

3. Phương pháp tiếp cận

3.1. Kiến trúc

Như được minh họa trong Hình 3, DeepSeek-OCR sử dụng một kiến trúc VLM tích hợp từ đầu đến cuối, bao gồm cả bộ mã hóa và bộ giải mã. Bộ mã hóa (tên là DeepEncoder) có nhiệm vụ trích xuất các đặc điểm hình ảnh, chuyển đổi chúng thành dạng các "token" và sau đó nén những thông tin hình ảnh này lại. Bộ giải mã được sử dụng để tạo ra kết quả mong muốn dựa trên các token hình ảnh và các thông tin hướng dẫn được cung cấp. DeepEncoder chứa khoảng 380 triệu tham số, được cấu tạo chủ yếu từ mô-đun SAM-base có 80 triệu tham số [17] và mô-đun CLIP-large có 300 triệu tham số [29], được kết nối liên tiếp với nhau. Bộ giải mã sử dụng kiến trúc MoE có 3 tỷ tham số hoạt động [19, 20], với tổng số 570 triệu tham số được sử dụng trong quá trình xử lý dữ liệu. Trong các phần tiếp theo, chúng ta sẽ tìm hiểu sâu hơn về các thành phần cấu tạo của mô hình này, quy trình xử lý dữ liệu và các kỹ thuật huấn luyện mô hình.

3.2. DeepEncoder

Để nghiên cứu khả năng áp dụng phương pháp nén dữ liệu hình ảnh dựa trên ngữ cảnh, chúng ta cần một bộ mã hóa hình ảnh có những đặc điểm sau: 1. Có khả năng xử lý các độ phân giải cao; 2. Tiêu thụ ít tài nguyên khi xử lý độ phân giải cao; 3. Sử dụng ít "token" hình ảnh; 4. Hỗ trợ các đầu vào có độ phân giải khác nhau; 5. Có số lượng tham số vừa phải. Tuy nhiên, như đã mô tả ở Phần 2.1, các bộ mã hóa mã nguồn mở

hiện có chưa thể đáp ứng đầy đủ tất cả các yêu cầu này. Do đó, chúng tôi đã tự thiết kế một bộ mã hóa hình ảnh mới, có tên là DeepEncoder.

3.2.1. Kiến trúc của DeepEncoder

DeepEncoder chủ yếu bao gồm hai thành phần: một thành phần trích xuất đặc điểm từ dữ liệu hình ảnh, được điều khiển bởi thuật toán chú ý theo cửa sổ (window attention); và một thành phần trích xuất đặc điểm kiến thức hình ảnh, sử dụng thuật toán chú ý toàn cục (global attention). Để tận dụng những lợi ích từ các nghiên cứu trước đây về việc huấn luyện sơ bộ, chúng tôi lựa chọn SAM-base (kích thước khối dữ liệu 16) và CLIP-large làm cấu trúc chính cho hai thành phần này. Đối với CLIP, chúng tôi loại bỏ lớp nhúng dữ liệu khối đầu tiên, vì đầu vào của nó không còn là hình ảnh nữa mà là các mã token thu được từ quá trình xử lý trước đó. Giữa hai thành phần này, chúng tôi áp dụng một mô-đun hội tụ 2 lớp để thực hiện thao tác giảm độ phân giải các mã token hình ảnh xuống còn 1/16. Mỗi lớp hội tụ có kích thước kernel là 3, bước chạy là 2, độ đệm là 1; số lượng kênh tín hiệu tăng từ 256 lên 1024. Nếu chúng ta đưa vào một hình ảnh có kích thước 1024×1024, DeepEncoder sẽ chia nó thành ***LATEX INLINE_1*** mã token khối. Vì nửa đầu của bộ mã hóa này chủ yếu sử dụng thuật toán chú ý theo cửa sổ và chỉ tiêu tốn 80MB bộ nhớ, nên mức độ sử dụng bộ nhớ vẫn được coi là hợp lý. Trước khi đi vào giai đoạn chú ý toàn cục, 4096 mã token này sẽ được xử lý qua bộ phận nén, làm giảm số lượng chúng xuống còn ***LATEX INLINE_1*** mã token; điều này giúp kiểm soát được lượng bộ nhớ cần thiết cho quá trình xử lý.

Hình 4 | Để kiểm tra hiệu suất của mô hình ở các tỷ lệ nén khác nhau (đòi hỏi số lượng các "token thị giác" khác nhau) và nâng cao tính ứng dụng thực tế của DeepSeek-OCR, chúng tôi đã cấu hình nó với nhiều chế độ độ phân giải khác nhau.

Bảng 1: Khả năng hỗ trợ nhiều độ phân giải của DeepEncoder. Vì mục đích nghiên cứu lẫn ứng dụng, chúng tôi đã thiết kế DeepEncoder với nhiều độ phân giải khác nhau và các chế độ điều chỉnh độ phân giải một cách linh hoạt.

Chế độ	Độ phân giải gốc	Độ phân giải động
	Tiny	Small
Độ phân giải	512	640
Token	64	100
Quy trình xử lý	Thay đổi kích thước	Thay đổi kích thước

3.2.2. Hỗ trợ nhiều độ phân giải khác nhau

Giả sử chúng ta có một hình ảnh chứa 1000 ký tự quang học, và chúng ta muốn kiểm tra cần bao nhiêu "token thị giác" để giải mã hình ảnh đó. Điều này đòi hỏi mô hình phải hỗ trợ việc sử dụng một số lượng khác nhau các "token thị giác". Nói cách khác, bộ mã hóa sâu (DeepEncoder) cần phải hỗ trợ nhiều độ phân giải khác nhau.

Chúng tôi đáp ứng các yêu cầu nêu trên thông qua việc áp dụng phương thức ghép nối động các thông tin mã hóa vị trí, đồng thời thiết kế nhiều chế độ xử lý khác nhau để

thực hiện việc huấn luyện mô hình đồng thời, nhằm giúp một mô hình DeepSeek-OCR duy nhất có thể hỗ trợ nhiều độ phân giải khác nhau. Như được minh họa trong Hình 4, công cụ DeepEncoder chủ yếu hỗ trợ hai chế độ đầu vào chính: độ phân giải gốc và độ phân giải động; mỗi chế độ này lại bao gồm nhiều tiểu chế độ con khác nhau.

Độ phân giải gốc hỗ trợ bốn chế độ con: Tiny, Small, Base và Large, với các độ phân giải và số lượng token tương ứng là 512×512 (64), 640×640 (100), 1024×1024 (256) và 1280×1280 (400). Vì các chế độ Tiny và Small có độ phân giải tương đối thấp, để tránh lãng phí token hình ảnh, các hình ảnh sẽ được xử lý bằng cách điều chỉnh kích thước trực tiếp. Đối với các chế độ Base và Large, để bảo toàn tỷ lệ khung hình ban đầu của hình ảnh, chúng sẽ được điền thêm ký tự để đạt đến kích thước mong muốn. Sau khi điền thêm ký tự, số lượng token hình ảnh hợp lệ sẽ ít hơn so với số lượng thực tế; công thức tính toán là:

$$N_{\text{valid}} = \left\lceil N_{\text{actual}} \times \left[1 - \left(\frac{\max(w,h) - \min(w,h)}{\max(w,h)} \right) \right] \right\rceil$$

Trong đó, ***LATEXInline_1*** và ***LATEXInline_1*** lần lượt đại diện cho chiều rộng và chiều cao của hình ảnh đầu vào ban đầu.

Độ phân giải động có thể được cấu thành từ hai độ phân giải nguyên bản khác nhau. Ví dụ, chế độ Gundam bao gồm ***LATEXInline_4*** ô hình ảnh (các góc nhìn cục bộ) và một góc nhìn tổng thể gồm ***LATEXInline_4*** ô hình ảnh. Phương pháp phân chia hình ảnh thành các ô này tuân theo tiêu chuẩn InternVL2.0 [8]. Việc hỗ trợ độ phân giải động chủ yếu nhằm phục vụ các ứng dụng cụ thể, đặc biệt là đối với các tập tin hình ảnh có độ phân giải siêu cao (chẳng hạn như ảnh báo chí). Việc phân chia hình ảnh thành các ô hình ảnh nhỏ là một biện pháp giúp giảm lượng bộ nhớ cần thiết để xử lý dữ liệu hình ảnh. Đáng chú ý là do độ phân giải nguyên bản của chúng tôi khá lớn, nên khi sử dụng độ phân giải động, hình ảnh không bị phân mảnh quá nhiều (số lượng các ô hình ảnh luôn nằm trong khoảng từ 2 đến 9). Số lượng các ô hình ảnh được DeepEncoder đưa ra khi sử dụng chế độ Gundam là ***LATEXInline_4***, trong đó ***LATEXInline_4*** chính là số lượng các ô hình ảnh đó. Đối với những hình ảnh có chiều rộng và chiều cao nhỏ hơn 640 pixel, giá trị của ***LATEXInline_4*** sẽ được đặt thành 0; trong trường hợp này, chế độ Gundam sẽ tự động chuyển sang chế độ Base.

Chế độ Gundam được huấn luyện cùng với bốn chế độ độ phân giải khác nhau, nhằm mục đích tạo ra một mô hình có thể hỗ trợ nhiều độ phân giải khác nhau. Lưu ý rằng chế độ Gundam-master (gồm 1024×1024 các góc nhìn cục bộ và 1280×1280 góc nhìn tổng thể) được hình thành thông qua việc tiếp tục huấn luyện trên mô hình DeepSeek-OCR đã được đào tạo trước đó. Việc sử dụng chế độ này chủ yếu nhằm đảm bảo sự cân bằng trong quá trình xử lý dữ liệu; bởi vì độ phân giải của chế độ Gundam-master quá lớn, nếu huấn luyện nó cùng với các chế độ khác sẽ làm chậm tốc độ tổng thể của quá trình huấn luyện.

3.3. Bộ giải mã MoE

Bộ giải mã của chúng tôi sử dụng mô hình DeepSeekMoE [19, 20], cụ thể là phiên bản DeepSeek-3B-MoE. Trong quá trình thực hiện các phép tính dự đoán, mô hình này kích hoạt 6 trong tổng số 64 mô-đun chuyên biệt và 2 mô-đun chung, với khoảng 570 triệu tham số được sử dụng trong quá trình tính toán. Mô hình DeepSeekMoE 3B rất phù hợp cho các nghiên cứu về các hệ thống trí tuệ nhân tạo tập trung vào từng lĩnh vực cụ thể (chẳng hạn như công nghệ nhận dạng ký tự quang học – OCR), vì nó vừa giữ được khả năng biểu đạt mạnh mẽ của các mô hình có dung lượng lớn, vừa sở hữu hiệu suất tính toán cao như các mô hình nhỏ hơn.

Bộ giải mã tái tạo lại biểu diễn văn bản gốc từ các mã ẩn dấu đã được nén của thuật toán DeepEncoder như sau:

$$f_{\text{dec}} : \mathbb{R}^{n \times d_{\text{latent}}} \rightarrow \mathbb{R}^{N \times d_{\text{text}}}; \quad \hat{\mathbf{X}} = f_{\text{dec}}(\mathbf{Z}) \quad \text{where } n \leq N$$

Trong đó, các chuỗi ký tự ***LATEXInline_2*** là những thông tin ẩn được nén lại từ mô hình DeepEncoder, còn các chuỗi ký tự ***LATEXInline_2*** sau đó là biểu diễn văn bản được tái tạo lại. Hàm ***LATEXInline_2*** đại diện cho một mối quan hệ phi tuyến tính có thể được các mô hình ngôn ngữ được thiết kế gọn gàng học được một cách hiệu quả thông qua phương thức huấn luyện kiểu OCR. Có thể suy đoán rằng các mô hình LLM, nhờ vào các quá trình tối ưu hóa huấn luyện chuyên biệt, sẽ thể hiện khả năng tích hợp những công nghệ này một cách tự nhiên hơn.

3.4. Động cơ xử lý dữ liệu

Chúng tôi đã xây dựng những bộ dữ liệu huấn luyện phức tạp và đa dạng cho hệ thống DeepSeek-OCR, bao gồm:
- **Dữ liệu OCR 1.0**: Chủ yếu bao gồm các nhiệm vụ OCR truyền thống như nhận diện chữ trong hình ảnh hoặc tài liệu;
- **Dữ liệu OCR 2.0**: Tập trung vào việc phân tích các hình ảnh nhân tạo phức tạp, chẳng hạn như biểu đồ thông dụng, công thức hóa học hay dữ liệu liên quan đến hình học phẳng;
- **Dữ liệu thị giác tổng quát**: Nhằm trang bị cho DeepSeek-OCR khả năng hiểu biết về hình ảnh một cách tổng quát, đồng thời giữ nguyên giao diện thị giác cơ bản của hệ thống.

Dữ liệu OCR 1.0

Dữ liệu tài liệu là ưu tiên hàng đầu đối với DeepSeek-OCR. Chúng tôi đã thu thập được 30 triệu trang tài liệu dạng PDF đa dạng ngôn ngữ từ Internet; trong đó, tiếng Trung và tiếng Anh chiếm khoảng 25 triệu trang, còn các ngôn ngữ khác chiếm 5 triệu trang. Đối với những dữ liệu này, chúng tôi đã tạo ra hai loại bộ dữ liệu chuẩn để huấn luyện mô hình: ghi chú sơ lược và ghi chú chi tiết.

Ghi chú sơ lược được trích xuất trực tiếp từ toàn bộ tập dữ liệu bằng công cụ ***LATEXInline_1***, nhằm giúp mô hình nhận diện văn bản trong hình ảnh, đặc biệt là đối với các

ngôn ngữ thiểu số. Ghi chú chi tiết bao gồm 2 triệu trang tài liệu tiếng Trung và tiếng Anh; chúng được đánh dấu bằng các mô hình bố cục tiên tiến (như PP-DocLayout [33]) và các mô hình OCR (như MinuerU [34] và GOT-OCR2.0 [38]) để tạo ra dữ liệu phục vụ việc nhận diện văn bản. Đối với các ngôn ngữ thiểu số, trong phần nhận diện, chúng tôi nhận thấy rằng các mô hình bố cục có khả năng tổng quát hóa tốt. Trong phần xử lý văn bản, chúng tôi sử dụng công cụ ***LATEXInline_1*** để tạo ra các đoạn dữ liệu nhỏ để huấn luyện mô hình GOT-OCR2.0; sau đó, dùng mô hình đã được huấn luyện này để đánh dấu các đoạn dữ liệu đó sau khi xử lý bố cục, từ đó tạo ra 600.000 mẫu dữ liệu. Trong quá trình huấn luyện DeepSeek-OCR, các ghi chú sơ lược và chi tiết được sử dụng với những cách khác nhau.

Bộ dữ liệu chuẩn dùng cho phần ghi chú chi tiết có thể được xem ở Hình 5. Chúng tôi cũng thu thập thêm 3 triệu bản dữ liệu loại Word, từ đó tạo ra các cặp dữ liệu hình ảnh-văn bản chất lượng cao mà không cần thông tin về bố cục, bằng cách trích xuất nội dung trực tiếp từ các tài liệu này. Những dữ liệu này đặc biệt hữu ích đối với các công thức toán học và các bảng biểu định dạng HTML. Ngoài ra, chúng tôi cũng sử dụng một số dữ liệu mã nguồn mở [28, 37] như tài liệu bổ sung.

Đối với công nghệ nhận dạng ký tự trong hình ảnh các cảnh vật tự nhiên, mô hình của chúng tôi hỗ trợ chủ yếu tiếng Trung và tiếng Anh. Các nguồn dữ liệu hình ảnh được lấy từ LAION [31] và Wukong [13], và đã được đánh dấu bằng công cụ PaddleOCR [9]; mỗi nguồn chứa 10 triệu mẫu dữ liệu cho cả tiếng Trung và tiếng Anh. Tương tự như công nghệ nhận dạng ký tự trong tài liệu, công nghệ này cũng cho phép người dùng điều khiển việc có xuất ra các khung xác định vị trí ký tự hay không thông qua các lệnh đặc biệt.

Dữ liệu OCR 2.0

Theo tiêu chuẩn GOT-OCR2.0 [38], chúng tôi coi dữ liệu liên quan đến biểu đồ, công thức hóa học và hình học phẳng là dữ liệu OCR 2.0. Đối với dữ liệu biểu đồ, dựa trên công cụ OneChart [7], chúng tôi sử dụng các thư viện pyecharts và matplotlib để hiển thị 10 triệu hình ảnh biểu đồ, bao gồm các loại biểu đồ đường thẳng, thanh, tròn và kết hợp phổ biến. Quá trình phân tích dữ liệu biểu đồ được xem là việc chuyển đổi hình ảnh thành bảng dữ liệu HTML, như minh họa trong Hình 6(a). Đối với công thức hóa học, chúng tôi sử dụng định dạng SMILES từ PubChem làm nguồn dữ liệu và sử dụng RDKit để hiển thị chúng dưới dạng hình ảnh, tạo ra tổng cộng 5 triệu cặp hình ảnh-không gian văn bản. Đối với hình ảnh hình học phẳng, chúng tôi áp dụng phương pháp Slow Perception [39] để tạo ra các dữ liệu mới; cụ thể, chúng tôi sử dụng kích thước "perception-ruler" là 4 để mô hình hóa từng đoạn thẳng. Để tăng đa dạng của dữ liệu được hiển thị, chúng tôi áp dụng phương pháp tăng cường dữ liệu bằng cách dịch chuyển các hình ảnh đó theo các hướng khác nhau trong hệ tọa độ gốc, sao cho các điểm tương ứng trên hình ảnh gốc vẫn nằm tại cùng một vị trí. Dựa trên các phương pháp này, chúng tôi đã thu thập được tổng cộng 1 triệu bộ dữ liệu hình học phẳng, như minh họa trong Hình 6(b).

3.4.3. Dữ liệu thị giác tổng quát

DeepEncoder có thể hưởng lợi từ những kết quả đạt được trong quá trình huấn luyện sơ bộ của CLIP, đồng thời cũng sở hữu đủ số lượng tham số để tiếp nhận và xử lý các kiến thức thị giác tổng quát. Do đó, chúng tôi cũng đã chuẩn bị một số dữ liệu phù hợp cho mô hình DeepSeek-OCR. Theo hướng dẫn của DeepSeek-VL2 [40], chúng tôi đã tạo ra các dữ liệu liên quan đến các nhiệm vụ như ghi chú văn bản, phát hiện đối tượng và xác định mối quan hệ giữa các đối tượng trong hình ảnh. Lưu ý rằng DeepSeek-OCR không phải là một mô hình VLM tổng quát, và phần dữ liệu này chỉ chiếm khoảng 20% tổng lượng dữ liệu được sử dụng. Chúng tôi giới thiệu loại dữ liệu này chủ yếu nhằm bảo đảm tính tổng quát của công cụ thị giác này, giúp các nhà nghiên cứu quan tâm đến mô hình của chúng tôi và các nhiệm vụ thị giác có thể tiếp tục phát triển công việc của mình một cách thuận lợi trong tương lai.

3.4.4. Dữ liệu chỉ chứa văn bản

Để đảm bảo khả năng xử lý ngôn ngữ của mô hình, chúng tôi đã sử dụng 10% dữ liệu huấn luyện ban đầu chỉ chứa văn bản, trong đó tất cả các dữ liệu đều được xử lý sao cho có độ dài bằng 8192 ký tự – đây cũng chính là độ dài chuỗi dữ liệu được sử dụng trong DeepSeek-OCR. Nói chung, khi huấn luyện DeepSeek-OCR, dữ liệu OCR chiếm 70%, dữ liệu thị giác thông thường chiếm 20%, và dữ liệu chỉ chứa văn bản chiếm 10%.

Hình 5 | Hiển thị các ghi chú chi tiết từ công nghệ OCR 1.0. Chúng tôi trình bày dữ liệu đính thực theo định dạng xen kẽ giữa văn bản và tọa độ; mỗi đoạn văn bản đều đi kèm với tọa độ và nhãn của nó trong hình ảnh gốc. Tất cả các tọa độ đều được quy đổi sang khoảng từ 0 đến 1000.

Hình 6: Đối với các biểu đồ, chúng tôi không sử dụng định dạng từ điển [7] của OneChart, mà thay vào đó sử dụng định dạng bảng HTML cho các tiêu đề biểu đồ; cách này giúp tiết kiệm được một lượng lớn token. Đối với các đối tượng hình học phẳng, chúng tôi chuyển đổi dữ liệu thực tế thành định dạng từ điển, trong đó các khóa bao gồm các thông tin như đoạn thẳng, tọa độ đầu mút, loại đoạn thẳng, v.v., nhằm tăng tính dễ đọc. Mỗi đoạn thẳng đều được mã hóa theo phương pháp Slow Perception [39].

3.5. Quy trình đào tạo dữ liệu

Quy trình đào tạo của chúng tôi khá đơn giản và bao gồm hai giai đoạn chính: a) Đào tạo riêng bộ công cụ DeepEncoder; b) Đào tạo hệ thống DeepSeek-OCR. Lưu ý rằng chế độ Gundam-master được tạo ra bằng cách tiếp tục đào tạo mô hình DeepSeek-OCR đã được huấn luyện sẵn với dữ liệu mẫu số lượng 6 triệu bản. Vì quy trình đào tạo này giống hệt với các chế độ khác, chúng tôi sẽ không trình bày chi tiết thêm ở đây.

3.5.1. Đào tạo mô hình DeepEncoder

Theo phương pháp Vary [36], chúng tôi sử dụng một mô hình ngôn ngữ gọn nhẹ [15] và cơ sở hạ tầng dự đoán các ký tự tiếp theo để huấn luyện thuật toán DeepEncoder. Trong giai đoạn này, chúng tôi sử dụng toàn bộ dữ liệu từ các phiên bản OCR 1.0 và 2.0 đã nêu trước đó, cùng với 100 triệu dữ liệu thông thường được lấy mẫu từ tập dữ

liệu LAION [31]. Tất cả các dữ liệu này đều được huấn luyện trong 2 chu kỳ, với kích thước mỗi lần huấn luyện là 1280 dữ liệu; thuật toán tối ưu hóa AdamW [23] cùng lịch trình điều chỉnh tốc độ học máy loại cosine annealing [22] được sử dụng, với tốc độ học máy được thiết lập là ***LATEXInline_0***. Độ dài chuỗi dữ liệu được sử dụng cho việc huấn luyện là 4096.

3.5.2. Đào tạo DeepSeek-OCR

Sau khi DeepEncoder được triển khai xong, chúng tôi sử dụng dữ liệu được đề cập ở Phần 3.4 để huấn luyện mô hình DeepSeek-OCR. Toàn bộ quá trình huấn luyện được thực hiện trên nền tảng HAI-LLM [14]. Mô hình này áp dụng phương thức song song hóa dữ liệu và được chia thành 4 phần: hai phần thuộc về DeepEncoder và hai phần thuộc về bộ giải mã. Đối với DeepEncoder, chúng tôi coi các thành phần SAM và bộ nén dữ liệu là các công cụ xử lý dữ liệu hình ảnh; chúng được đặt trong phần PP0 và các tham số của chúng được "đóng băng" (không thay đổi trong quá trình huấn luyện). Phần CLIP được xem là lớp đầu vào dùng để tạo ra các vector đại diện cho dữ liệu hình ảnh, và nó được đặt trong phần PP1 với các tham số có thể thay đổi được trong quá trình huấn luyện. Đối với phần mô hình ngôn ngữ, vì DeepSeek3B-MoE gồm 12 lớp, chúng tôi chia 6 lớp cho mỗi trong hai phần PP2 và PP3. Chúng tôi sử dụng 20 node, mỗi node được trang bị 8 GPU loại A100-40G, để thực hiện quá trình huấn luyện. Tỷ lệ song song hóa dữ liệu là 40, và kích thước mỗi lô dữ liệu được xử lý là 640. Chúng tôi sử dụng thuật toán tối ưu hóa AdamW, với lịch trình huấn luyện dựa trên các bước cụ thể, và tỷ lệ học tập ban đầu được đặt ở mức ***LATEXInline_0***. Đối với dữ liệu chỉ chứa văn bản, tốc độ huấn luyện là 90 tỷ token mỗi ngày; còn đối với dữ liệu đa phương thức, tốc độ huấn luyện là 70 tỷ token mỗi ngày.

Bảng 2 | Chúng tôi đã kiểm tra tỷ lệ nén văn bản giữa hình ảnh và văn bản khi sử dụng công cụ DeepSeek-OCR, bằng cách sử dụng tất cả các tài liệu tiếng Anh có độ dài từ 600 đến 1300 ký tự trong bộ dữ liệu đánh giá của Fox [21]. Các ký tự văn bản đại diện cho số lượng ký tự sau khi văn bản gốc được phân tích bằng công cụ tokenizer của DeepSeek-OCR. Các ký tự hình ảnh, với giá trị lần lượt là 64 hoặc 100, đại diện cho số lượng ký tự hình ảnh được tạo ra sau khi hình ảnh đầu vào được điều chỉnh kích thước thành 512×512 hoặc 640×640 bởi công cụ DeepEncoder.

Các từ khóa văn bản	Số lượng ký hiệu hình ảnh = 64	Số lượng ký hiệu hình ảnh = 100
	Phương thức nén chính xác	10,5 lần
600–700	96,5%	10,5 lần
700–800	93,8%	11,8 lần
800–900	83,8%	13,2 lần
900–1000	85,9%	15,1 lần
1000–1100	79,3%	16,5 lần
1100–1200	76,4%	17,7 lần
1200–1300	59,1%	19,7 lần

4. Đánh giá

4.1 Nghiên cứu về việc nén dữ liệu hình ảnh kết hợp văn bản

Chúng tôi chọn các bài kiểm thử Fox [21] để đánh giá khả năng nén và giải nén của công cụ DeepSeek-OCR đối với các tài liệu chứa nhiều văn bản, nhằm khám phá sơ bộ tính khả thi và giới hạn của phương pháp nén dựa trên ngữ cảnh. Chúng tôi sử dụng phần tài liệu tiếng Anh trong bộ dữ liệu Fox, áp dụng công cụ token hóa của DeepSeek-OCR (với bộ từ vựng khoảng 129.000 từ) để chia nhỏ văn bản thành các đơn vị token, sau đó lựa chọn các tài liệu có từ 600 đến 1.300 token để thử nghiệm – tương ứng với khoảng 100 trang. Vì số lượng token không quá lớn, chúng tôi chỉ cần kiểm tra hiệu suất của công cụ ở hai chế độ: Tiny (64 token) và Small (100 token). Chúng tôi sử dụng câu hỏi mẫu để thực hiện việc này.

without layout: " Free OCR." to control the model's output format. Nevertheless, the output format still can not completely match Fox benchmarks, so the actual performance would be somewhat higher than the test results.

Như được thể hiện trong Bảng 2, với tỷ lệ nén lên đến 10×, độ chính xác khi giải mã của mô hình có thể đạt khoảng 97%, đây là một kết quả rất đáng khích lệ. Trong tương lai, có thể sẽ khả thi thực hiện việc nén văn bản mà không làm mất thông tin với tỷ lệ gần 10× thông qua các phương pháp chuyển đổi văn bản thành hình ảnh. Tuy nhiên, khi tỷ lệ nén vượt quá 10×, hiệu suất sẽ bắt đầu giảm sút; có hai lý do có thể giải thích điều này: thứ nhất là cấu trúc của các văn bản dài trở nên phức tạp hơn; thứ hai là các văn bản dài có thể bị mờ đi khi được hiển thị ở độ phân giải 512×512 hoặc 640×640. Vấn đề thứ nhất có thể được khắc phục bằng cách hiển thị toàn bộ nội dung văn bản trên cùng một trang, còn vấn đề thứ hai chúng tôi tin rằng sẽ được giải quyết thông qua cơ chế "quên lãng" dữ liệu. Khi tỷ lệ nén văn bản đạt gần 20×, độ chính xác vẫn có thể đạt khoảng 60%. Những kết quả này cho thấy rằng việc nén văn bản mà không làm mất thông tin là một hướng nghiên cứu rất triển vọng và đáng giá; đồng thời, phương pháp này không gây ra bất kỳ gánh nặng nào vì có thể tận dụng được cơ sở hạ tầng VLM, trong khi các hệ thống đa phương thức vốn đã cần đến bộ mã hóa hình ảnh.

Bảng 3 | Chúng tôi sử dụng công cụ OmniDocBench [27] để đánh giá hiệu suất của mô hình DeepSeek-OCR trong các tác vụ phân tích tài liệu thực tế. Tất cả các chỉ số trong bảng này đều là khoảng cách biên dịch; giá trị nhỏ hơn chứng tỏ hiệu suất tốt hơn. "Tokens" đại diện cho số lượng các ký hiệu được sử dụng trung bình trên mỗi trang, và "**_LATEXInline_0**_200dpi" có nghĩa là việc sử dụng thuật toán fitz để điều chỉnh độ phân giải hình ảnh gốc lên mức 200dpi. Đối với mô hình DeepSeek-OCR, các giá trị nằm trong dấu ngoặc đơn ở cột "Tokens" đại diện cho số lượng các ký hiệu hợp lệ được tính toán theo Phương trình 1.

Mô hình	Thông số đo lường	Tiếng Anh (tổng thể)	Tiếng Anh (văn bản)	Tiếng Anh (công thức)	Thứ tự bảng	Tiếng Trung (tổng thể)	Tiếng Trung (văn bản)	Tiếng Trung (công thức)	Thứ tự bảng
Dolphin [11]	-	0.356	0.352	0.465	0.258	0.35	0.44	0.44	0.604
Marker [1]	-	0.296	0.085	0.374	0.609	0.116	0.497	0.293	0.688
Mathpix [2]	-	0.191	0.105	0.306	0.243	0.108	0.364	0.381	0.454
MinerU-2.1.1 [34]	-	0.162	0.072	0.313	0.166	0.097	0.244	0.111	0.581
MonkeyOCR-1.2B [18]	-	0.154	0.062	0.295	0.164	0.094	0.263	0.179	0.464
PPstructure-v3 [9]	-	0.152	0.073	0.295	0.162	0.077	0.223	0.136	0.535
Các mô hình end-to-end									
Nougat [6]	2352	0.452	0.365	0.488	0.572	0.382	0.973	0.998	0.941
SmolDocling [25]	392	0.493	0.262	0.753	0.729	0.227	0.816	0.838	0.997
InternVL2-76B [8]	6790	0.44	0.353	0.543	0.547	0.317	0.443	0.29	0.701
Qwen2.5-VL-7B [5]	3949	0.316	0.151	0.376	0.598	0.138	0.399	0.243	0.5
OLMOCR [28]	3949	0.326	0.097	0.455	0.608	0.145	0.469	0.293	0.655
GOT-OCR2.0 [38]	256	0.287	0.189	0.360	0.459	0.141	0.411	0.315	0.528
OCRVerse-3B [3]	5949	0.238	0.112	0.447	0.269	0.126	0.349	0.256	0.716

Mô hình	Thông số đo lường	Tiếng Anh (tổng thể)	Tiếng Anh (văn bản)	Tiếng Anh (công thức)	Thứ tự bảng	Tiếng Trung (tổng thể)	Tiếng Trung (văn bản)	Tiếng Trung (công thức)	Thứ tự bảng
GPT4o [26]	-	0.233	0.144	0.425	0.234	0.128	0.399	0.409	0.606
InternVL3-78B [42]	6790	0.218	0.117	0.38	0.279	0.095	0.296	0.21	0.533
Qwen2.5-VL-72B [5]	3949	0.214	0.092	0.315	0.341	0.106	0.261	0.18	0.434
dots.ocr [30]	3949	0.182	0.137	0.320	0.166	0.182	0.261	0.229	0.468
Gemini2.5-Pro [4]	-	0.148	0.055	0.356	0.13	0.049	0.212	0.168	0.439
Mineru2.0 [34]	6790	0.133	0.045	0.273	0.15	0.066	0.238	0.115	0.506
dots.ocr†200dpi [30]	5545	0.125	**0.032**	0.329	**0.099**	**0.04**	0.16	**0.066**	0.416
DeepSeek-OCR (end2end)									
Tiny	**64**	0.386	0.373	0.469	0.422	0.283	0.361	0.307	0.635
Small	100	0.221	0.142	0.373	0.242	0.125	0.284	0.24	0.53
Base	256(182)	0.137	0.054	0.267	0.163	0.064	0.24	0.205	0.474
Large	400(285)	0.138	0.054	0.277	0.152	0.067	0.208	0.143	0.461
Gundam	795	0.127	0.043	0.269	0.134	0.062	0.181	0.097	0.432
Gundam-M†200dpi	1853	**0.123**	0.049	**0.242**	0.147	0.056	**0.157**	0.087	**0.377**

Bảng 4 | Khoảng cách chỉnh sửa đối với các loại tài liệu khác nhau trong OmniDocBench. Kết quả cho thấy một số loại tài liệu có thể đạt được hiệu suất tốt chỉ với 64 hoặc 100 “token thị giác”, trong khi những loại khác lại yêu cầu phải sử dụng chế độ “Gundam”.

Loại/Hình thức	Trang trình bày sách giáo khoa	Báo cáo tài chính	Sách giáo khoa điện tử	Đề thi kiểm tra	Tạp chí	Các bài báo học thuật	Ghi chú	Báo chí	Tổng thể
Nhỏ	0.147	0.116	0.207	0.173	0.294	0.201	0.395	0.297	0.94
Vừa	0.085	0.111	0.079	0.147	0.171	0.107	0.131	0.187	0.744
Cơ bản	0.037	0.08	0.027	0.1	0.13	0.073	0.052	0.176	0.645
Lớn	0.038	0.108	0.022	0.084	0.109	0.06	0.053	0.155	0.353
Gundam	0.035	0.085	0.289	0.095	0.094	0.059	0.039	0.153	0.122
Gundam-M	0.052	0.09	0.034	0.091	0.079	0.079	0.048	0.1	0.099

4.2. Hiệu suất thực tế của công nghệ OCR

DeepSeek-OCR không chỉ là một mô hình thử nghiệm mà còn sở hữu những khả năng thực tiễn mạnh mẽ, đồng thời có thể được sử dụng để xây dựng dữ liệu phục vụ việc huấn luyện các mô hình LLM/VLM. Để đánh giá hiệu suất của DeepSeek-OCR, chúng tôi đã thử nghiệm nó trên bộ dữ liệu OmniDocBench [27]; kết quả được trình bày trong Bảng 3. Chỉ cần sử dụng 100 token hình ảnh (với độ phân giải 640×640), DeepSeek-OCR đã vượt trội hơn so với GOT-OCR2.0 [38] – mô hình yêu cầu sử dụng tới 256 token. Khi sử dụng 400 token (trong đó 285 token hợp lệ, độ phân giải 1280×1280), DeepSeek-OCR đạt được hiệu suất ngang ngửa với các mô hình tiên tiến nhất trên bài kiểm thử này. Với lượng token ít hơn 800 token (chế độ Gundam), DeepSeek-OCR còn vượt trội hơn so với MinerU2.0 [34] – mô hình cần tới gần 7.000 token hình ảnh. Những kết quả này chứng minh rằng mô hình DeepSeek-OCR của chúng tôi rất mạnh mẽ trong các ứng dụng thực tế; đồng thời, nhờ khả năng nén dữ liệu hiệu quả, mô hình này còn tiềm năng phát triển cao hơn nữa trong tương lai.

Như được thể hiện trong Bảng 4, một số loại tài liệu chỉ cần rất ít các “token” để đạt được hiệu suất xử lý tốt; ví dụ, các slide chỉ cần 64 “token” là đủ. Đối với các tài liệu sách hoặc báo cáo, công cụ DeepSeek-OCR cũng có thể hoạt động hiệu quả với chỉ 100 “token”. Kết hợp với phân tích ở Phần 4.1, điều này có thể được giải thích là do hầu hết các “token” trong những loại tài liệu này có số lượng trong khoảng 1.000, nên tỷ lệ nén các “token” trong những trường hợp này không vượt quá 10 lần. Trong khi đó, đối với các tài liệu báo chí hoặc các tài liệu liên quan đến bộ phim Gundam, cần sử dụng chế độ xử lý đặc biệt mới có thể đạt được độ chính xác xử lý đủ cao; lý do là số lượng các “token” trong những loại tài liệu này lên tới 4.000–5.000, vượt xa mức 10 lần nén mà các chế độ khác có thể đạt được. Những kết quả thí nghiệm này càng làm rõ các giới hạn của phương pháp nén dữ liệu dựa trên ngữ cảnh, đồng thời cung cấp những thông tin hữu ích cho các nghiên cứu về việc tối ưu hóa cách sử dụng các “token” trong các hệ thống trí tuệ nhân tạo, cũng như các cơ chế nén dữ liệu và ghi nhớ thông tin trong các hệ thống này.

4.3.1. Phân tích sâu

DeepSeek-OCR vừa sở hữu khả năng phân tích cấu trúc trang văn bản, vừa có thể thực hiện công việc nhận dạng ký tự quang học theo phiên bản OCR 2.0. Nhờ đó, mô hình này có thể tiếp tục phân tích các chi tiết trong hình ảnh bên trong tài liệu thông qua việc gọi các mô hình phụ trợ, tính năng này được chúng tôi gọi là "phân tích sâu". Như được minh họa trong Hình 7, 8, 9 và 10, mô hình của chúng tôi có thể thực hiện công việc phân tích sâu trên các loại dữ liệu như biểu đồ, công thức hóa học, thậm chí cả các hình ảnh tự nhiên, và chỉ cần một yêu cầu đầu vào thống nhất là đủ.

Hình 7 | Trong lĩnh vực báo cáo nghiên cứu tài chính, chế độ phân tích sâu của công cụ DeepSeek-OCR có thể được sử dụng để trích xuất các biểu đồ có cấu trúc rõ ràng từ các tài liệu. Biểu đồ là hình thức biểu diễn dữ liệu quan trọng trong lĩnh vực tài chính và khoa học; do đó, khả năng trích xuất các biểu đồ một cách có tổ chức là yếu tố thiết yếu đối với các mô hình OCR trong tương lai.

Storybook Reading for Young Dual Language Learners

Cristina Gillanders and
Dina C. Castro

In a community of practice meeting, teachers discuss their experiences reading aloud to dual language learners.

Susan: When I am reading a story, the Latino children in my class just sit there. They look at me, but you can tell that they are not engaged in the story.

Lisa: That happens in my class too. The little girls play with their hair, and the boys play with their shoes.

Beverly: And when you ask questions about the story, children who speak English take over and you can't get an answer from the Latino children.

Facilitator: What do you think is happening here?

Lisa: I think they just don't understand what the story is about.

Facilitator: How can we help them understand the story so they can participate?

RESEARCHERS WIDELY RECOMMEND storybook reading for promoting the early language and literacy of young children. By listening to stories, children learn about written syntax and vocabulary and develop phonological awareness and concepts of print, all of which are closely linked to learning to read and write (National Early Literacy Panel 2008). Teachers usually know a read-aloud experience has been effective because they see the children maintain their interest in the story, relate different aspects of the story to their own experiences, describe the illustrations, and ask questions about the characters and plot.

However, listening to a story read aloud can be a very different experience for children who speak a language other than English. What happens when the children are read to in a language they are just beginning to learn? What happens when an English-speaking teacher reads a story to a group of children who are learning English as a second language?

As illustrated in the vignette at the beginning of this article, teachers often describe young dual language learners in their class as distracted and unengaged during read-aloud sessions in English. In this article, we describe teaching strategies that English-speaking teachers can use when reading aloud to young dual language learners. These strategies are part of the Nuestros Niños Early Language and Literacy Program, a professional development intervention designed to improve the quality of teaching practices in prekindergarten classrooms to support Spanish-speaking dual language learners (Castro et al. 2006). The intervention was developed and evaluated in a study funded by

Cristina Gillanders, PhD, is a researcher at the FPG Child Development Institute at the University of North Carolina–Chapel Hill. She was an investigator in the Nuestros Niños study, and has worked with dual language learners as a bilingual preschool teacher, teacher educator, and researcher. cristina.gillanders@unc.edu

Hướng dẫn ôn tập cho bài viết này sẽ được đăng tải trực tuyến vào giữa tháng Giêng tại địa chỉ www.naecyc.org/.

Storybook Reading for Young Dual Language Learners

Cristina Gillanders and Dina C. Castro

In a community of practice meeting, teachers discuss their experiences reading aloud to dual language learners.

Susan: When I am reading a story, the Latino children in my class just sit there. They look at me, but you can tell that they are not engaged in the story.

Lisa: That happens in my class too. The little girls play with their hair, and the boys play with their shoes.

Beverly: And when you ask questions about the story, children who speak English take over and you can't get an answer from the Latino children.

Facilitator: What do you think is happening here?

Lisa: I think they just don't understand what the story is about.

Facilitator: How can we help them understand the story so they can participate?

RESEARCHERS WIDELY RECOMMEND storybook reading for promoting the early language and literacy of young children. By listening to stories, children learn about written syntax and vocabulary and develop phonological awareness and concepts of print, all of which are closely linked to learning to read and write (National Early Literacy Panel 2008). Teachers usually know a read-aloud experience has been effective because they see the children maintain their interest in the story, relate different aspects of the story to their own experiences, describe the illustrations, and ask questions about the characters and plot.

However, listening to a story read aloud can be a very different experience for children who speak a language other than English. What happens when the children are read to in a language they are just beginning to learn? What happens when an English-speaking teacher reads a story to a group of children who are learning English as a second language?

As illustrated in the vignette at the beginning of this article, teachers often describe young dual language learners in their class as distracted and unengaged during read-aloud sessions in English. In this article, we describe teaching strategies that English-speaking teachers can use when reading aloud to young dual language learners. These strategies are part of the Nuestros Niños Early Language and Literacy Program, a professional development intervention designed to improve the quality of teaching practices in prekindergarten classrooms to support Spanish-speaking dual language learners (Castro et al. 2006). The intervention was developed and evaluated in a study funded by the US Department of Education. Teachers from the North Carolina More at Four Pre-Kindergarten

Cristina Gillanders, PhD, is a researcher at the FPG Child Development Institute at the University of North Carolina–Chapel Hill. She was an investigator in the Nuestros Niños study, and has worked with dual language learners as a bilingual preschool teacher, teacher educator, and researcher. cristina.gillanders@unc.edu

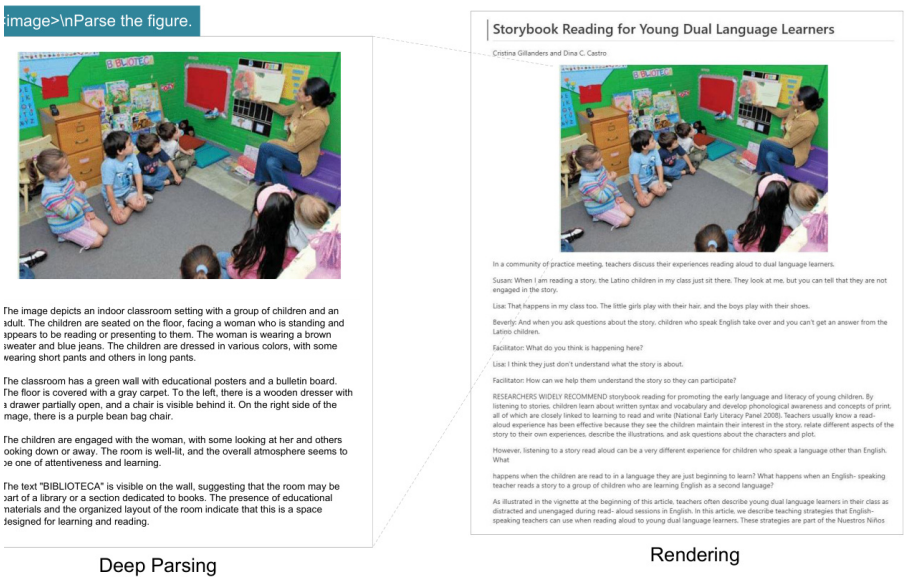
Dina C. Castro, PhD, is a senior scientist at the FPG Child Development Institute. She was the principal investigator for the Nuestros Niños study. Her research focuses on improving the quality of early education for children from diverse cultural and linguistic backgrounds. dina.castro@unc.edu

Photos courtesy of the authors.

A study guide for this article will be available in mid-January online at **www.naeyc.org/yc**

Hình ảnh đầu vào

Kết quả



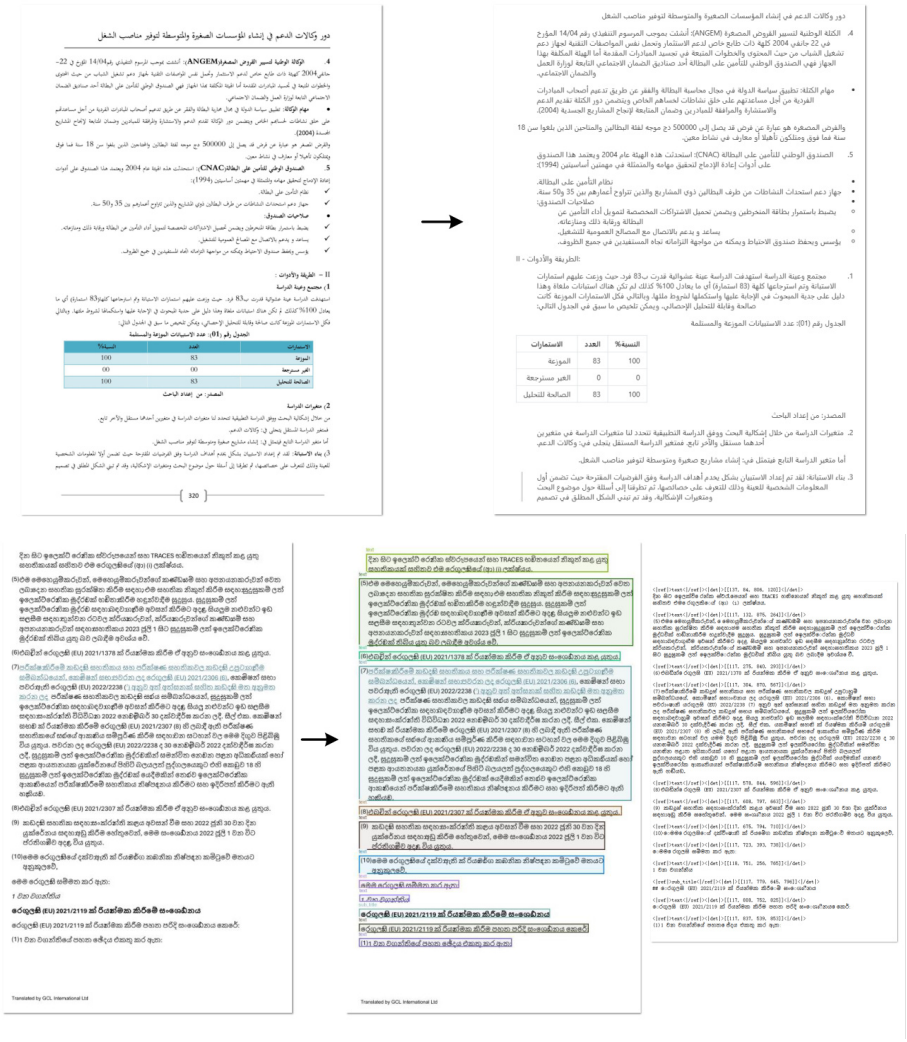
Hình 8: Đối với sách và bài viết, chế độ phân tích sâu có thể tạo ra các chú thích chi tiết cho các hình ảnh trong tài liệu. Chỉ cần một yêu cầu nhỏ, mô hình có thể tự động xác định loại hình ảnh đó và hiển thị kết quả cần thiết.

Hình 9 | Chương trình DeepSeek-OCR, khi hoạt động ở chế độ phân tích sâu, cũng có thể nhận diện các công thức hóa học trong các tài liệu khoa học và chuyển đổi chúng sang định dạng SMILES. Trong tương lai, công nghệ OCR 1.0+2.0 có thể đóng vai trò quan trọng trong sự phát triển của các mô hình trí tuệ nhân tạo dạng VLM/LLM trong các lĩnh vực khoa học, kỹ thuật và toán học.

Hình 10 | DeepSeek-OCR cũng có khả năng sao chép các hình dạng hình học đơn giản. Do mối quan hệ phức tạp giữa các đoạn thẳng trong các hình học đó, việc phân tích các dữ liệu hình học là một nhiệm vụ vô cùng thách thức và còn nhiều công việc phải làm.

4.3.2. Nhận dạng đa ngôn ngữ

Dữ liệu PDF trên Internet không chỉ bao gồm tiếng Trung và tiếng Anh, mà còn chứa rất nhiều tài liệu đa ngôn ngữ, điều này cũng vô cùng quan trọng trong quá trình huấn luyện các mô hình LLM. Đối với các tài liệu PDF, công cụ DeepSeek-OCR có thể xử lý gần 100 ngôn ngữ khác nhau. Giống như các tài liệu tiếng Trung và tiếng Anh, dữ liệu đa ngôn ngữ cũng được hỗ trợ cả trong các định dạng OCR có cấu trúc và không có cấu trúc. Kết quả xử lý được trình bày trong Hình 11; chúng tôi đã chọn hai ngôn ngữ là tiếng Ả Rập và tiếng Sinhala để minh họa kết quả.



Hình 11: Để có thể xử lý các tệp PDF được thu thập rộng rãi từ nhiều nguồn ngôn ngữ khác nhau, chúng tôi đã huấn luyện mô hình của mình với khả năng nhận dạng chữ viết tay (OCR) cho gần 100 ngôn ngữ. Các tài liệu thuộc các ngôn ngữ thiểu số cũng có thể được xử lý thành cả dạng có bố cục trang trí và dạng không có bố cục trang trí, thông qua các lệnh yêu cầu khác nhau.

4.3.3. Nhận thức tổng quát về hình ảnh

Chúng tôi cũng trang bị cho DeepSeek-OCR những khả năng nhận diện hình ảnh ở mức độ nhất định. Kết quả hiển thị của các chức năng này được trình bày trong Hình 12.

Hình 12 | Chúng tôi vẫn giữ nguyên các khả năng nhận diện hình ảnh của DeepSeek-OCR, bao gồm việc mô tả hình ảnh, phát hiện đối tượng, xác định vị trí đối tượng trên

hình, v.v. Đồng thời, nhờ việc sử dụng dữ liệu chỉ chứa văn bản, các khả năng ngôn ngữ của DeepSeek-OCR cũng được duy trì. Lưu ý rằng vì chúng tôi không áp dụng giai đoạn huấn luyện có giám sát (SFT – Supervised Fine-Tuning), nên mô hình này không phải là một chatbot; một số chức năng của nó chỉ có thể được kích hoạt khi có các yêu cầu cụ thể được đưa vào.

5. Thảo luận

Công trình nghiên cứu của chúng tôi đại diện cho những khám phá ban đầu về các giới hạn trong lĩnh vực nén dữ liệu hình ảnh kết hợp văn bản; chúng tôi đã tìm hiểu xem cần bao nhiêu thông tin hình ảnh thì mới có thể giải mã được một số lượng nhất định thông tin văn bản. Kết quả ban đầu khá đáng khích lệ: Phương pháp DeepSeek-OCR cho phép thực hiện việc nén dữ liệu với tỷ lệ khoảng 10 lần mà vẫn giữ được độ chính xác gần như 100%, trong khi tỷ lệ nén 20 lần vẫn đảm bảo độ chính xác khoảng 60%. Những phát hiện này mở ra những hướng đi hứa hẹn cho các ứng dụng tương lai, chẳng hạn như việc áp dụng các công nghệ xử lý hình ảnh trong các cuộc đối thoại đa vòng để đạt được hiệu quả nén cao hơn.

Hình 13 | Cơ chế "quên lãng" là một trong những đặc điểm cơ bản nhất của trí nhớ con người. Phương pháp nén dữ liệu dựa trên nguyên lý này có thể mô phỏng cơ chế này bằng cách hiển thị các đoạn văn cũ đã được sử dụng trong các lần nén trước lên các hình ảnh để thực hiện bước nén đầu tiên; sau đó, kích thước của các hình ảnh đó sẽ được điều chỉnh dần để đạt được mức độ nén đa tầng. Khi đó, số lượng các ký tự trong văn bản sẽ giảm dần và nội dung văn bản sẽ trở nên mờ nhạt hơn, từ đó thực hiện việc "quên lãng" thông tin đó.

Trong các trường hợp cũ hơn, chúng ta có thể dần giảm kích thước các hình ảnh được hiển thị để tiếp tục giảm lượng thông tin cần sử dụng. Giả định này được lấy cảm hứng từ mối tương đồng tự nhiên giữa quá trình suy giảm trí nhớ con người theo thời gian và sự suy giảm khả năng nhận thức thị giác khi khoảng cách tăng lên; cả hai hiện tượng đều tuân theo những quy luật tương tự về việc mất dần thông tin theo từng giai đoạn, như được minh họa trong Hình 13. Bằng cách kết hợp các cơ chế này, phương pháp nén hình ảnh này tạo ra một cơ chế "mất dần thông tin" phản ánh đúng quá trình quên lãng sinh học: thông tin mới vẫn được lưu trữ với độ chính xác cao, trong khi những thông tin cũ hơn sẽ tự nhiên bị mất đi do tỷ lệ nén tăng lên.

Mặc dù những nghiên cứu ban đầu cho thấy tiềm năng trong việc xử lý các dữ liệu ngữ cảnh siêu dài một cách hiệu quả, trong đó các thông tin ngữ cảnh gần đây được bảo toàn với độ chi tiết cao, trong khi các thông tin ngữ cảnh cũ tiêu tốn ít tài nguyên hơn, chúng tôi cũng nhận thức rằng đây vẫn là giai đoạn nghiên cứu sơ bộ và cần được tiếp tục khám phá thêm. Phương pháp này mở ra hướng đi để xây dựng các kiến trúc xử lý ngữ cảnh có khả năng xử lý lượng dữ liệu lớn mà vẫn duy trì sự cân bằng giữa việc lưu trữ thông tin và các hạn chế về mặt tính toán; tuy nhiên, những ứng dụng thực tế và những hạn chế của các hệ thống nén dữ liệu loại này vẫn đòi hỏi phải được nghiên cứu sâu hơn trong các công trình nghiên cứu tương lai.

6. Kết luận

Trong báo cáo kỹ thuật này, chúng tôi đề xuất mô hình DeepSeek-OCR và đã kiểm chứng sơ bộ tính khả thi của phương pháp nén văn bản dựa trên ngữ cảnh thông qua mô hình này. Kết quả cho thấy rằng mô hình này có thể giải mã hiệu quả các thông tin văn bản với số lượng gấp hơn 10 lần so với số lượng dữ liệu hình ảnh được sử dụng để huấn luyện mô hình. Chúng tôi tin rằng phát hiện này sẽ góp phần thúc đẩy sự phát triển của các mô hình VLM và LLM trong tương lai. Ngoài ra, DeepSeek-OCR còn là một mô hình rất hữu ích, có khả năng sản xuất dữ liệu huấn luyện quy mô lớn, đóng vai trò quan trọng trong việc hỗ trợ hoạt động của các mô hình LLM. Tất nhiên, việc áp dụng công nghệ OCR một mình là chưa đủ để chứng minh trọn vẹn tính hiệu quả của phương pháp nén văn bản dựa trên ngữ cảnh; chúng tôi sẽ tiếp tục tiến hành các thử nghiệm bổ sung như huấn luyện kết hợp dữ liệu kỹ thuật số và hình ảnh, hay các phương pháp đánh giá khác trong tương lai. Từ góc độ khác, lĩnh vực nén văn bản dựa trên ngữ cảnh vẫn còn nhiều tiềm năng để nghiên cứu và cải tiến, đây chính là một hướng đi đầy hứa hẹn cho tương lai.

Tài liệu tham khảo

[1] Đối tượng đánh dấu. Địa chỉ URL: https://github.com/datalab-to/marker.

[2] Mathpix. Địa chỉ URL: https://mathpix.com/.

[3] Ocrflux, 2025. Địa chỉ URL: https://github.com/chatdoc-com/OCRFlux.

[4] G. AI. Gemini 2.5-pro, 2025. Địa chỉ URL: https://gemini.google.com/.

[5] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu và J. Lin. Báo cáo kỹ thuật Qwen2.5-v1. Bản thảo trên arXiv: arXiv:2502.13923, 2025.

[6] L. Blecher, G. Cucurull, T. Scialom và R. Stojnic. Nougat: Phương pháp hiểu dữ liệu văn bản thông qua công nghệ thị giác nhân tạo dành cho các tài liệu học thuật. Bản thảo trước khi công bố trên arXiv: arXiv:2308.13418, 2023.

[7] J. Chen, L. Kong, H. Wei, C. Liu, Z. Ge, L. Zhao, J. Sun, C. Han và X. Zhang. Onechart: Phương pháp tinh chỉnh quá trình trích xuất thông tin cấu trúc biểu đồ bằng cách sử dụng một ký hiệu hỗ trợ. Trong Tài liệu hội nghị ACM International Conference on Multimedia lần thứ 32, trang 147–155, năm 2024.

[8] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma và cộng sự. Chúng ta còn cách GPT-4V bao xa? Làm thế nào để thu hẹp khoảng cách với các mô hình đa□□ thương mại nhờ vào các bộ công cụ mã nguồn mở. Bản thảo trên arXiv: arXiv:2404.16821, 2024.

[9] C. Cui, T. Sun, M. Lin, T. Gao, Y. Zhang, J. Liu, X. Wang, Z. Zhang, C. Zhou, H. Liu và cộng sự. Báo cáo kỹ thuật PaddleOCR 3.0. Bản thảo trên arXiv: arXiv:2507.05595, 2025.

[10] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin và cộng sự. Patch n' Pack: Navit – một công cụ biến đổi hình ảnh phù hợp với mọi tỷ lệ khung hình và độ phân giải. Tạp chí Advances in Neural Information Processing Systems, tập 36, trang 3632–3656, năm 2023.

[11] H. Feng, S. Wei, X. Fei, W. Shi, Y. Han, L. Liao, J. Lu, B. Wu, Q. Liu, C. Lin và cộng sự. Dolphin: Phương pháp phân tích hình ảnh tài liệu thông qua các gợi ý đa dạng. Bản thảo trên arXiv: arXiv:2505.14059, 2025.

[12] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra và D. Parikh. Việc tăng tầm quan trọng của khả năng hiểu hình ảnh trong các bài toán trả lời câu hỏi hình ảnh: Nâng cao vai trò của công nghệ nhận diện hình ảnh trong lĩnh vực này. Trong tài liệu báo cáo hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 6904–6913, năm 2017.

[13] J. Gu, X. Meng, G. Lu, L. Hou, N. Minzhe, X. Liang, L. Yao, R. Huang, W. Zhang, X. Jiang và cộng sự. Wukong: Một bài kiểm thử chuẩn quy mô lớn dành cho ngôn ngữ Trung Quốc, với dữ liệu được huấn luyện đa modal trên quy mô 100 triệu dữ liệu. Tạp chí Advances in Neural Information Processing Systems, tập 35, trang 26418–26431, năm 2022.

[14] High-flyer. HAI-LLM: Công cụ đào tạo hiệu quả và gọn nhẹ dành cho các mô hình lớn, 2023. Địa chỉ URL: https://www.high-flyer.cn/en/blog/hai-llm.

[15] S. Iyer, X. V. Lin, R. Pasunuru, T. Mihaylov, D. Simig, P. Yu, K. Shuster, T. Wang, Q. Liu, P. S. Koura và cộng sự. Opt-imal: Phương pháp học máy meta để tối ưu hóa các hướng dẫn sử dụng mô hình ngôn ngữ thông qua khả năng tổng quát hóa. Bản thảo trên arXiv: arXiv:2212.12017, 2022.

[16] S. Kazemzadeh, V. Ordonez, M. Matten và T. Berg. Referitgame: Phương pháp định danh các vật thể trong những bức ảnh về cảnh quan tự nhiên. Trong tài liệu hội nghị EMNLP 2014 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên, trang 787–798, 2014.

[17] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo và cộng sự. Phương pháp phân đoạn dữ liệu một cách linh hoạt. Bản thảo trước khi công bố trên arXiv: arXiv:2304.02643, 2023.

[18] Z. Li, Y. Liu, Q. Liu, Z. Ma, Z. Zhang, S. Zhang, Z. Guo, J. Zhang, X. Wang và X. Bai. Monkeyocr: Phương pháp phân tích văn bản dựa trên mô hình ba yếu tố – nhận diện cấu trúc, xác định mối quan hệ giữa các thành phần trong văn bản. Bản thảo trên arXiv: arXiv:2506.05218, 2025.

[19] A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Deng, C. Ruan, D. Dai, D. Guo và cộng sự. Deepseek-v2: Một mô hình ngôn ngữ loại "hỗn hợp chuyên gia" mạnh mẽ, tiết kiệm tài nguyên và hiệu quả. Bản thảo trước khi công bố trên arXiv: arXiv:2405.04434, 2024.

[20] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan và cộng sự. Báo cáo kỹ thuật Deepseek-v3. Bản thảo trên arXiv: arXiv:2412.19437, 2024.

[21] C. Liu, H. Wei, J. Chen, L. Kong, Z. Ge, Z. Zhu, L. Zhao, J. Sun, C. Han và X. Zhang. Khả năng hiểu nội dung các tài liệu gồm nhiều trang một cách chi tiết, bất kể vị trí nào được chọn làm điểm tập trung. Bản thảo trên arXiv: arXiv:2405.14295, 2024.

[22] I. Loshchilov và F. Hutter. Phương pháp hạ dốc đạo hàm ngẫu nhiên với cơ chế khởi động lại. Bản thảo trên arXiv: arXiv:1608.03983, 2016.

[23] I. Loshchilov và F. Hutter. Phương pháp điều chỉnh trọng số theo nguyên lý tách rời các yếu tố ảnh hưởng lẫn nhau. Trong hội nghị ICLR năm 2019.

[24] A. Masry, D. X. Long, J. Q. Tan, S. Joty và E. Hoque. Chartqa: Một bộ tiêu chuẩn đánh giá khả năng trả lời câu hỏi liên quan đến biểu đồ, dựa trên cả lý luận trực quan lẫn logic. Bản thảo trên arXiv: arXiv:2203.10244, 2022.

[25] A. Nassar, A. Marafioti, M. Omenetti, M. Lysak, N. Livathinos, C. Auer, L. Morin, R. T. de Lima, Y. Kim, A. S. Gurbuz và cộng sự. Smoldocling: Một mô hình học máy siêu nhỏ gọn dành cho việc chuyển đổi đa phương thức tài liệu một cách tự động. Bản thảo trên arXiv: arXiv:2503.11576, 2025.

[26] OpenAI. Báo cáo kỹ thuật về GPT-4, 2023.

[27] L. Ouyang, Y. Qu, H. Zhou, J. Zhu, R. Zhang, Q. Lin, B. Wang, Z. Zhao, M. Jiang, X. Zhao và cộng sự. Omnidocbench: Công cụ đánh giá hiệu suất các phương pháp xử lý tài liệu PDF với các bình luận chi tiết. Trong Tài liệu hội nghị Thị giác máy tính và Nhận dạng mẫu, trang 24838–24848, 2025.

[28] J. Poznanski, A. Rangapur, J. Borchardt, J. Dunkelberger, R. Huff, D. Lin, C. Wilhelm, K. Lo và L. Soldaini. OLMOC: Giải mã hàng nghìn tỷ token trong các tệp PDF bằng các mô hình ngôn ngữ hình ảnh. Bản thảo trên arXiv: arXiv:2502.18443, 2025.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark và cộng sự. Việc xây dựng các mô hình thị giác có khả năng áp dụng được trong nhiều tình huống thông qua việc học từ dữ liệu ngôn ngữ tự nhiên. Trong Hội nghị Quốc tế về Máy học, trang 8748–8763. PMLR, 2021.

[30] Rednote. dots.ocr, 2025. Địa chỉ URL: https://github.com/rednote-hilab/dots.ocr.

[31] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev và A. Komatsuzaki. Laion-400m: Bộ dữ liệu mở chứa 400 triệu cặp hình ảnh-khóa văn bản đã được lọc. Bản thảo trên arXiv: arXiv:2111.02114, 2021.

[32] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh và M. Rohrbach. Những nỗ lực nhằm phát triển các mô hình trả lời câu hỏi dựa trên hình ảnh có khả năng "đọc" thông tin từ hình ảnh. Trong tài liệu báo cáo hội nghị IEEE/CVF về thị giác máy tính và nhận diện mẫu, trang 8317–8326, năm 2019.

[33] T. Sun, C. Cui, Y. Du và Y. Liu. Pp-doclayout: Một mô hình thống nhất để phát hiện cấu trúc trang tài liệu, nhằm tăng tốc độ xây dựng dữ liệu quy mô lớn. Bản thảo trên arXiv: arXiv:2503.17213, 2025.