THÔNG TIN CHUNG CỦA NHÓM

- **Link YouTube video của báo cáo (tối đa 5 phút):**
- **https://youtu.be/l-zokgu3cls**
- **Link slides (dạng .pdf đặt trên Github của nhóm):**
  https://github.com/bach04py/CS519.O21.KHTN.git
- ***Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới***
- ***Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in***

• Họ và Tên: Nguyễn Xuân Bách • MSSV: 22520093 	• Lớp: CS519.O21.KHTN • Tự đánh giá (điểm tổng kết môn): 9/10 • Số buổi vắng: 1 • Số câu hỏi QT cá nhân: 3 • Số câu hỏi QT của cả nhóm: 15 • Link Github: https://github.com/bach04py/CS519.O21.KHTN.git • Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:Em làm mọi công việc.

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

NHẬN DẠNG ĐỐI TƯỢNG QUA DỰ ĐOÁN MÃ THÔNG BÁO TIẾP THEO

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

OBJECT RECOGNITION AS NEXT TOKEN PREDICTION

TÓM TẮT ***(Tối đa 400 từ)***

Chúng tôi trình bày một phương pháp đề xuất để áp dụng nhận dạng vật thể như một dự đoán mã thông tin tiếp theo. Ý tưởng của chúng tôi là sử dụng một bộ giải mã ngôn ngữ để dự đoán tự động các mã thông tin văn bản từ các vectơ nhúng hình ảnh để tạo ra các nhãn. Để gắn kết quá trình dự đoán này vào tính tự hồi quy, chúng tôi tùy chỉnh một mặt nạ chú ý không gây tác động cho bộ giải mã, với hai đặc điểm quan trọng: mô hình hóa các mã từ các nhãn khác nhau độc lập với nhau, và xem các mã hình ảnh như một tiền tố. Cơ chế che phủ này đã truyền cảm hứng cho một phương pháp hiệu quả - lấy mẫu một lần - để cùng một lúc lấy mẫu các mã từ nhiều nhãn khác nhau và xếp hạng các nhãn được tạo ra dựa trên xác suất trong quá trình suy luận. Để tăng tính hiệu quả, chúng tôi đề xuất một chiến lược đơn giản để xây dựng một bộ giải mã gọn nhẹ bằng cách loại bỏ các khối trung gian không cần thiết từ một mô hình ngôn ngữ đã được huấn luyện trước. Phương pháp này tạo ra một bộ giải mã có hiệu suất tương đương với mô hình đầy đủ và đồng thời tiết kiệm tài nguyên tính toán đáng kể.

GIỚI THIỆU ***(Tối đa 1 trang A4)***

Chúng tôi tập trung vào một vấn đề cơ bản trong thị giác máy tính - nhận dạng vật thể - dịch một hình ảnh thành các nhãn vật thể. Nói chung, khung nhận dạng bao gồm một bộ mã hóa hình ảnh và một bộ giải mã. Bộ mã hóa hình ảnh, có thể là một mạng tích chập (CNN) hoặc một bộ biến đổi thị giác (ViT) , tạo ra các vectơ embedding, trong khi bộ giải mã sử dụng chúng để dự đoán các nhãn vật thể.

Nếu bộ giải mã là một bộ phân loại tuyến tính (linear classifier) nó cần được khởi tạo với bộ trọng số cố định. Ví dụ, ResNet khởi tạo lớp tuyến tính cuối cùng của nó với 1K vectơ nhúng, còn được gọi là trọng số, để đại diện cho 1K ảnh trong ImageNet . Tuy nhiên, các trọng số tĩnh này hạn chế khả năng của mô hình nhận dạng bất kỳ vật thể nào. Hạn chế này có thể được giảm bớt bằng cách sử dụng một mô hình ngôn ngữ làm bộ giải mã để tạo ra một tập hợp linh hoạt các vectơ nhúng vật thể từ các mô tả đầu vào. Ví dụ, CLIP mã hóa các mô tả vật thể thành các trọng số động thông qua lời nhắc "một bức ảnh của {L}", trong đó L có thể là bất kỳ tên vật thể nào, và so khớp các trọng số này với các vectơ nhúng hình ảnh để nhận dạng các vật thể.

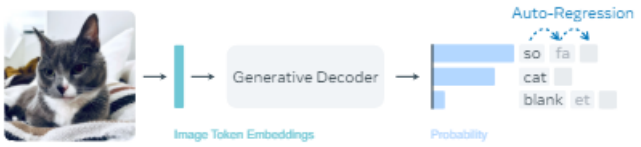

Figure 1. **Object recognition as next token prediction** using a generative decoder such as a transformer-based language model to auto-regressively predict object labels. *Photo authorized with CC BY 4.0.*

Mục tiêu của bài toán này là sử dụng phương pháp đề xuất để loại bỏ việc định trước các nhãn vật thể trong mô hình CLIP và thực hiện nhận dạng vật thể chỉ bằng cách sử dụng mô hình ngôn ngữ lớn (LLM) và mô hình CLIP.

Input của bài toán này là một hình ảnh, trong đó không có thông tin cho trước về các nhãn vật thể .

Output của bài toán là các nhãn vật thể được giải mã từ vectơ nhúng hình ảnh bằng cách sử dụng một bộ giải mã ngôn ngữ. Điều này đòi hỏi sử dụng vector embedding của mô hình ngôn ngữ lớn và vectơ nhúng hình ảnh đã được điều chỉnh với văn bản từ bộ mã hóa hình ảnh CLIP. Các vectơ nhúng này được biến đổi tuyến tính để phù hợp với chiều nhúng của bộ giải mã ngôn ngữ, và từ đó, bộ giải mã có thể tạo ra các nhãn vật thể cụ thể của hình ảnh.

Tóm lại, bài toán này nhằm mục đích nhận dạng các nhãn vật thể trong hình ảnh chỉ bằng cách sử dụng mô hình ngôn ngữ lớn và mô hình hình ảnh CLIP, mà không cần thông tin về các nhãn được định trước.

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- **Xây dựng thuật toán nhận dạng đối tượng từ hình ảnh dựa trên dự đoán mã thông báo tiếp theo và so sánh độ chính xác, tốc độ xử lý, và khả năng tổng quát hóa với các phương pháp khác.**
- **Đánh giá hiệu suất thuật toán trên các bộ dữ liệu đa dạng, bao gồm đối tượng phổ biến và hiếm, bằng cách đo độ chính xác, độ phủ, và khả năng xử lý đối tượng hiếm.**

- **Tối ưu hóa thuật toán để đạt hiệu suất tốt với tài nguyên tính toán hạn chế, đo lường thông qua thời gian huấn luyện, kích thước mô hình, và hiệu suất nhận dạng đối tượng**

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

Cụ thể cách thực hiện:.

Nội dung chính

Chúng tôi trình bày một phương pháp mới trong nhận diện đối tượng thông qua việc dự đoán token tiếp theo. Ý tưởng chính là sử dụng một bộ giải mã ngôn ngữ (language decoder) để tự động dự đoán các token văn bản từ các embedding ảnh, từ đó tạo thành các nhãn đối tượng.

Related work:

1.Aligning Images and Text:bao gồm các câu, cụm từ hoặc từ, trong không gian chung đã trở nên phổ biến trong các phương pháp khớp hình ảnh-văn bản và là nền tảng của các khung học tương phản, đồng thời cũng được sử dụng để tạo mô tả văn bản từ hình ảnh. Việc tích hợp nhận thức thị giác với các mô hình ngôn ngữ lớn (LLM) như GPT và LLaMA bằng cách coi các embedding của hình ảnh như các embedding của token ngôn ngữ đang ngày càng phổ biến, giúp hợp nhất thông tin thị giác và văn bản một cách liền mạch. Các phương pháp này được áp dụng vào các nhiệm vụ như phát hiện, nhận dạng few-shot, giải thích bằng văn bản, lý giải phân loại, mô hình bottleneck, suy luận, và các mô hình chat để chú thích và trả lời câu hỏi dựa trên hình ảnh (VQA).

2. Tackling Open-Vocabulary Tasks: trong nhận dạng, phát hiện và phân đoạn thường liên quan đến việc huấn luyện trên một tập hợp nhãn cơ bản và sau đó nhận dạng các nhãn hiếm chưa thấy. Học tương phản, như CLIP, sử dụng mô hình ngôn ngữ để mã hóa nhãn nhằm so khớp với hình ảnh, nhưng có những hạn chế do các nhãn cơ bản và nhãn hiếm được định trước. CaSED sử dụng

các chú thích thô để tạo ra một thư viện không cần từ vựng định trước, nhưng hiệu suất của nó phụ thuộc nhiều vào việc chọn lựa thư viện.

Phương pháp thực hiện:

1. . Revisiting Object Recognition:

Chúng tôi bắt đầu bằng cách xem xét ngắn gọn về nhận dạng đối tượng trong công thức chung. Giả sử rằng các hình ảnh 2D được đưa vào một mô hình backbone, ví dụ như ViT trong CLIP, tạo ra các embedding hình ảnh $X_v \in R^{M*D}$, trong đó M là kích thước không gian và D là kích thước embedding. Vấn đề nhận dạng nhằm giải mã nhãn đối tượng chỉ từ X_v, chuyển đổi các embedding hình ảnh thành không gian văn bản.

Trong những năm qua, thiết kế cốt lõi của sự chuyển đổi này sử dụng một tập hợp các embedding văn bản $W \in R^{N*D}$, để tìm kiếm sự căn chỉnh tối ưu với X_v:

$$\arg\max \ \sigma(\mathbf{W} f(\mathbf{X}_v)^\top),$$

trong đó σ là hàm softmax và f là hàm biến đổi X_v để căn chỉnh với W. Ví dụ, các bộ phân loại tuyến tính như ResNet sử dụng trung bình pooling như f để biến đổi X_v thành một đại diện vector duy nhất, và khởi tạo W bằng cách sử dụng một tập hợp các khái niệm định trước tương ứng với các nhãn đối tượng, ví dụ N=1000 cho ImageNet. Các khung tương phản như CLIP nhúng một tập hợp các mô tả đối tượng định trước vào WWW, và áp dụng một phép tổng hợp (như embedding [CLS]) và chiếu tuyến tính như f trên X_v. Phương trình trên nhằm tối đa hóa sự căn chỉnh giữa f(X_v) và W. Không gian của W đóng vai trò quan trọng trong sự căn chỉnh này vì sự đa dạng và phong phú của các embedding trong W ảnh hưởng trực tiếp đến khả năng của mô hình trong việc phân biệt các đối tượng. Tuy nhiên, các bộ phân loại tuyến tính và các khung tương phản giới hạn W trong một tập hợp con định trước, điều này có thể hạn chế khả năng của mô hình trong việc nhận dạng bất kỳ đối tượng nào. Mục tiêu của chúng tôi là loại bỏ sự hạn chế này và mở rộng W ra toàn bộ không gian văn bản.

2. Auto-Regression for Recognition:

Gần đây, các mô hình ngôn ngữ lớn (LLM) đã có những tiến bộ đáng kể trong việc hiểu và tạo ra văn bản. Xét rằng các embedding token của chúng được đào tạo để đại diện cho toàn bộ không gian văn bản, chúng tôi định nghĩa W với các embedding token từ một LLM được đào tạo trước, ví dụ như LLaMA, với N=32K token văn bản. Sau đó, phương trình trên thay đổi thành việc dự đoán token:

$$P(\mathbf{w}|\mathbf{X}_\mathrm{v}) = \arg\max\ \sigma(\mathbf{W} f(\mathbf{X}_\mathrm{v})^\top),$$

trong đó w đại diện cho token có khả năng cao nhất đối với X_v . Trong phương pháp của chúng tôi, f là sự kết hợp của phép chiếu tuyến tính và LLM được đào tạo trước để chiếu X_v vào không gian văn bản của W. Điều này có nghĩa là f là bộ giải mã ngôn ngữ của chúng tôi.

Để hướng dẫn bộ giải mã ngôn ngữ trong nhiệm vụ nhận dạng, chúng tôi kích hoạt nó bằng một hướng dẫn ngắn - "the objects in the image are" - được token hóa thành $X_p \in R^{P*D}$. Sau đó, chúng tôi nối X_v và X_p để tạo thành embedding token đầu vào của chúng tôi:

$$\mathbf{X} = \mathbf{X}_\mathrm{v} \oplus [\mathrm{IMG}] \oplus \mathbf{X}_\mathrm{p},$$

trong đó $\oplus$ là phép nối và [IMG] là một token đặc biệt để chỉ ranh giới.

Thông thường, một nhãn bao gồm nhiều token, ví dụ, "sofa" có hai token [so] và [fa]. Không mất tính tổng quát, chúng tôi giả sử một nhãn L có T token. Bây giờ dự đoán L tương đương với việc tự động dự đoán các token của nó:

$$P(L) = P(\mathbf{w}_1, \ldots, \mathbf{w}_T|\mathbf{X}_\mathrm{v}, \mathbf{X}_\mathrm{p}) = \prod_{t=1}^{T} P(\mathbf{w}_t|\mathbf{w}_{<t}, \mathbf{X}),$$

rong đó w_t là token thứ t của L, và $w_{<t}$ là chuỗi các token trước token thứ t. Để tính xác suất có điều kiện trong phương trình trên, LLM dựa trên transformer trong f sử dụng một mặt nạ nguyên nhân M trên sự chú ý cặp để mô hình hóa sự phụ thuộc giữa các token:

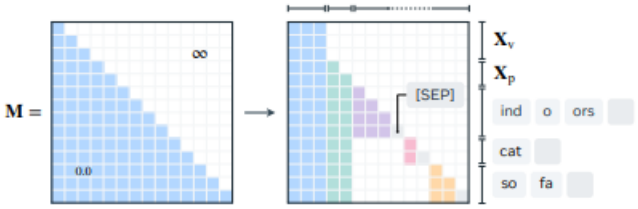

Figure 2. **Non-causal attention mask** for prefixing image tokens $\mathbf{X}_\mathrm{v}$ and decoupling tokens from different labels L_k to be independent at the [SEP] token.

trong đó tril(∞) là ma trận tam giác dưới với các giá trị không trong tam giác dưới và các giá trị vô cùng trong tam giác trên. Điều này buộc token wtw_twt chỉ tham gia vào các token trước $w_{<t}$, tức là làm cho w_t phụ thuộc vào $w_{<t}$, như được hiển thị bên trái Hình 2.

3. Non-causal Masking:

Thông thường, một hình ảnh chứa nhiều đối tượng, và mục tiêu của chúng tôi là dự đoán tất cả chúng. Giả sử có K đối tượng, và chúng tôi ký hiệu tập hợp các nhãn đầu ra cho hình ảnh là L={L1,...,LK}, trong đó nhãn thứ k có $T_k + 1$ token, bao gồm cả token đặc biệt [SEP] để phân tách. Khi đó, khả năng xuất hiện của tập hợp nhãn này trong hình ảnh là tích của các xác suất của chúng:

$$P(\mathcal{L}) = \prod_{k=1}^{K} P(L_k) = \prod_{k=1}^{K} \prod_{t=1}^{T_k+1} P(\mathbf{w}_t^k | \mathbf{w}_{<t}^k, \mathbf{X}). \quad (6)$$

Bây giờ, phương trình (6) không phải là một quy trình tự hồi quy tiêu chuẩn được áp dụng trong các mô hình ngôn ngữ lớn (LLMs) bởi vì wtkw_{t}^{k}wtk chỉ cần tham gia vào các token đầu vào X và các token trước đó w_t^k từ cùng một nhãn L_k. Điều này dựa trên hiểu biết rằng các nhãn cùng tồn tại trong cùng một hình ảnh do bối cảnh hình ảnh, nhưng chúng độc lập với nhau. Ngoài ra, các token hình ảnh X_v thể hiện mối tương quan không gian vốn có, trái ngược với mối tương quan thời gian của các token ngôn ngữ tự nhiên. Vì vậy, chúng tôi tùy chỉnh một non-casual mask M với hai thiết kế, được minh họa bên phải Hình 2: a) Chúng tôi tách rời mối tương quan giữa các token từ các nhãn khác nhau tại token [SEP] để ngăn các token này tham gia vào nhau; b) Chúng tôi coi các token hình ảnh X_v như một tiền tố, cho phép các token hình ảnh nhìn thấy nhau. Điều thú vị là mặt nạ chú ý không nhân quả của chúng tôi có thiết kế tương tự như mặt nạ cột trong [95] nhưng được phát triển từ một góc độ khác, trong đó mặt nạ cột cụ thể dành cho chú ý từ hình ảnh đến hình ảnh.

Cuối cùng, phương trình (6) là mục tiêu huấn luyện cuối cùng của chúng tôi. Chúng tôi sử dụng hàm mất mát cross-entropy để tối ưu hóa, với các nhãn được giám sát yếu được trích xuất từ các chú thích hình ảnh tương ứng.

4. One-Shot Sampling:

Cơ chế non-casual masking tách biệt các token từ các nhãn khác nhau, chỉ ra rằng token đầu tiên của bất kỳ nhãn nào có thể là token tiếp theo sau X trong vòng lấy mẫu đầu tiên. Nói cách khác, xác suất cao cho token đầu tiên, được lấy mẫu sau đầu vào X, sẽ dẫn đến mức độ liên quan cao hơn của nhãn với hình ảnh. Điều này gợi ý chúng ta lấy mẫu các token của nhiều nhãn song song, như được hiển thị trong Hình 3.

Figure 3. **One-shot sampling** for generating tokens of top-k labels in parallel. Once the model samples the [SEP] token, the label is completed. Otherwise, the model continues for unfinished labels.

Cho các token đầu vào X, chúng ta truyền chúng vào bộ giải mã và xếp hạng các đầu ra logits theo xác suất softmax. Các token top-k, được gọi là token khởi đầu, quyết định các nhãn top-k sẽ được tạo ra. Hiệu quả của việc liên kết các token khởi đầu với các nhãn cuối cùng được khám phá trong Bảng 8, nhấn mạnh tiềm năng của phương pháp tiếp cận đơn giản này. Sau đó, chúng ta lấy mẫu token tiếp theo cho các token khởi đầu top-k song song, sử dụng phương pháp lấy mẫu top-1, để tạo ra k nhãn. Nếu token được lấy mẫu là [SEP], nhãn đó được hoàn thành. Ngược lại, mô hình tiếp tục lấy mẫu token tiếp theo cho các nhãn chưa hoàn thành. Cuối cùng, chúng ta báo cáo xác suất của mỗi nhãn như là tích của các xác suất token của nó. Chúng tôi gọi phương pháp này là lấy mẫu một lần, cho phép lấy mẫu song song nhiều nhãn trong một lần. Chìa khóa cho tính song song của nó nằm ở cơ chế mặt nạ không nhân quả, điều này cũng tránh được vấn đề lặp lại thường gặp trong tìm kiếm tham lam và beam search, vì nó khiến mô hình tập trung đồng đều vào các token đầu vào X trên các nhãn khác nhau.

Tóm lại, lấy mẫu một lần khác với các phương pháp lấy mẫu khác ở hai khía cạnh chính: a) Nó hoạt động song song trên nhiều nhãn đối tượng, với mỗi nhánh song song xử lý một số lượng nhỏ token (khoảng ít hơn mười token), trái ngược với việc lấy mẫu tuần tự của các phương pháp khác; b) Nó tự nhiên phù hợp với nhiệm vụ nhận dạng hình ảnh bằng cách đại diện cho hình ảnh như một thực thể có tương quan không gian, trong khi các phương pháp lấy mẫu khác mô tả hình ảnh như một chuỗi các token.

5. Truncating the Decoder

Bây giờ, xem xét mô hình ngôn ngữ LLaMA trong bộ giải mã f của chúng tôi, chúng tôi cho rằng một tập hợp con cụ thể về hiểu biết ngôn ngữ trong các tham số phong phú của nó là rất quan trọng cho việc nhận dạng. Ý tưởng này thúc đẩy chúng tôi tập trung vào việc tối đa hóa hiệu quả bằng cách không sử dụng toàn bộ mô hình. Chúng tôi xây dựng bộ giải mã ngôn ngữ của mình, ban đầu dựa trên LLaMA 7B (phiên bản 1 hoặc 2), bằng cách cắt ngắn nó thành 6 khối transformer đầu tiên cùng với lớp đầu ra cuối cùng, như được mô tả trong Hình 4, đồng thời giữ nguyên bộ mã hóa từ và 32K token embedding đã được huấn luyện trước để mã hóa đầu vào. Chúng tôi gọi phiên bản sửa đổi này là bộ giải mã ngôn ngữ cắt ngắn, ký hiệu là Langtruncated trong các thí nghiệm của chúng tôi.

Figure 4. **Encoder and truncated decoder**. We retain the first 6 transformer blocks along with the final output layer of the LLaMA 7B as our truncated decoder, and train with partial encoder blocks.

Kết quả

Phương pháp đề xuất cho thấy khả năng nhận diện đối tượng mà không cần các nhãn đối tượng hoặc mô tả được xác định trước, đồng thời cải thiện hiệu suất và tốc độ suy diễn.

- Chúng tôi nhận thấy rằng CLIP, một phương pháp phổ biến hiện nay, yêu cầu một tập hợp mô tả đối tượng cố định trước khi suy diễn, hạn chế khả năng của mô hình trong thực tế.

- Phương pháp mới đề xuất sử dụng một mô hình ngôn ngữ lớn để giải mã các nhãn từ các embedding ảnh mà không cần các mô tả cố định.

- Chiến lược bỏ các khối trung gian trong mô hình ngôn ngữ đã huấn luyện trước giúp tăng hiệu suất suy diễn lên gấp 4.5 lần.

Phương pháp mới này không chỉ nâng cao khả năng nhận diện đối tượng mà còn giảm thiểu các yêu cầu về cấu hình cố định, mở ra tiềm năng lớn cho việc ứng dụng trong các hệ thống nhận diện đối tượng tự động.

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- **Khả năng Nhận dạng Đối tượng Hiệu quả: Thuật toán đề xuất có khả năng nhận dạng các đối tượng từ hình ảnh một cách hiệu quả và chính xác. Điều này được đánh giá thông qua các thí nghiệm so sánh với các phương pháp nhận dạng đối tượng hiện tại, như mạng nơ-ron tích chập (CNN).**
- **Hiệu suất Trên Bộ Dữ liệu Đa Dạng: Thuật toán đã được thử nghiệm trên nhiều bộ dữ liệu khác nhau, bao gồm cả những đối tượng phổ biến và hiếm, và đạt được độ chính xác cao, khả năng phủ rộng và khả năng xử lý tốt các đối tượng hiếm.**
- **Tối ưu Hóa Tài Nguyên: Thuật toán được tối ưu hóa để hoạt động hiệu quả với tài nguyên tính toán hạn chế, chẳng hạn như giảm thời gian huấn luyện và kích thước mô hình, trong khi vẫn duy trì hiệu suất cao trong nhận dạng đối tượng.**

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur

Mensch, Katherine Millican, Malcolm Reynolds, et al.

Flamingo: A Visual Language Model for Few-Shot Learning. In NeurIPS, 2022. 1, 2, 13

[2] Jacob Andreas and Dan Klein. Reasoning About Pragmatics With Neural Listeners and Speakers. In EMNLP, 2016.2

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel,

Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan

Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo:

An Open-Source Framework for Training Large Autoregressive Vision-Language Models. In arXiv:2308.01390,

2023. 1, 5, 6, 9, 12, 13

[4] Kobus Barnard and David Forsyth. Learning the Semantics

of Words and Pictures. In ICCV, 2001. 14

[5] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando

De Freitas, David M Blei, and Michael I Jordan. Matching Words and Pictures. In JMLR, 2003. 14

[6] Abhijit Bendale and Terrance Boult. Towards Open World

Recognition. In CVPR, 2015. 2

[7] Rodrigo Benenson and Vittorio Ferrari. From Colouring-in

to Pointillism: Revisiting Semantic Segmentation Supervision. arXiv:2210.14142, 2022. 4

[8] Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing With Python: Analyzing Text With the

Natural Language Toolkit. O'Reilly Media, Inc., 2009. 12

Data. In ACM SIGIR, 2003. 2