先日VLMのモデルの実装法の概要、評価法を扱いました。 今回はVLMの収集すべきデータセットについて、選定の前の確認事項から、用途別のデータセットについて説明します。

選定の前に確認したいこと

VLMの学習用データセットを探し始める前に、 **「どのようなモデルを、何の目的で作りたいか」** という設計図を明確にする必要があります。

VLMの世界は非常に幅広く、目的が「写真の説明」なのか「図表の読み取り」なのかによって、選ぶべきデータの種類が全く異なるからです。

1. 学習のフェーズ（ステージ）

VLMの学習は通常、段階を踏んで行われます。どの段階のデータを探しているのかを明確にしましょう。

- **Pre-training（事前学習）:** 画像とテキストの基本的な対応関係を学ばせる段階。
- *必要なデータ:* CLIPのように、インターネットから収集された大量の画像とキャプションのペア（数億〜数十億規模）。
- **Instruction Tuning（指示調整）:** 「この画像には何が写っていますか？」といったユーザーの指示に従えるようにする段階。
- **必要なデータ:** 「画像・問い・答え」がセットになった、人間またはAIによって精緻に作られたデータ（LLaVAなど）。

2. ターゲットとするタスク（ドメイン）

汎用的なモデルを作りたいのか、特定の専門分野に特化させたいのかを決めます。

- **一般キャプショニング:** 日常の風景を説明する（COCO, LAIONなど）。
- **VQA（視覚的質疑応答）:** 画像の内容について推論して答える（VQA v2など）。
- **Document/OCR:** 請求書や論文、スライドなどの文字情報を読み取る（DocVQA, ChartQAなど）。
- **空間認識:** オブジェクトの位置（バウンディングボックス）を特定する。

3. データの「質」と「量」のトレードオフ

- **量重視（Web-scale）:** 質はそこそこ（ノイズが多い）だが、とにかく大量。モデルの基礎体力をつけるのに向いています。
- **質重視（Curated/Synthetic）:** GPT-4Vなどの強力なモデルを使って生成された高品質な解説データ。最近のトレンドは、量よりも「高品質な推論プロセス」が含まれたデータに移行しています。

4. モデルのアーキテクチャ（入力形式）

どのような形式で画像を入力するモデルなのかによって、データの準備方法が変わります。

- **シングル画像:** 1枚の画像に対して1つのテキスト。
- **マルチ画像/ビデオ:** 複数の画像や動画のフレームを入力し、それらの関係性を説明させる必要があるか。

- **インターリーブ形式:** テキストの途中に画像が挟まるような、より自然なドキュメント形式（MMC4など）。

5. ライセンスと倫理的制約

商用利用を考えている場合、これが最も重要になることもあります。

- **商用利用可否:** Apache 2.0やMITライセンスのものか、CC BY-NC（非商用限定）などの制限があるか。
- **プライバシー:** 顔のぼかし処理がされているか、不適切なコンテンツ（NSFW）がフィルタリングされているか。

公開データセット

VLMの学習に利用できるデータセットは、現在「事前学習用（膨大・低品質）」から「指示調整用（中規模・高品質）」まで多岐にわたります。

目的別に代表的なものをまとめました。

1. 事前学習用（大規模画像・テキストペア）

モデルに「画像と概念の対応」を教えるためのデータセットです。

データセット名	規模	特徴
LAION-5B	58億ペア	インターネットから収集された最大級のデータ。現在は法規制の関係でリンク集として公開。
COYO-700M	7億ペア	KakaoBrainが公開。LAIONよりフィルタリングが精度良く行われている。
Conceptual Captions (CC3M/12M)	300万/1200万	Googleが公開。比較的クリーンで、初期のVLM研究（BLIPなど）でよく使われる。
OBELICS	1.4億画像	**インターリーブ形式**（記事の中に画像が混ざる形式）。ドキュメント理解に強い。

2. 指示調整用（Instruction Tuning）

ユーザーの問いに答えたり、推論を行ったりするための高品質なデータセットです。

データセット名	規模	特徴
LLaVA-v1.5 / v1.6	約60万件	GPT-4を用いて生成された高品質な会話データ。現在のVLM学習の標準。
ShareGPT4V	120万件	非常に詳細な記述（Detailed Caption）が含まれており、モデルの語彙力向上に寄与。
LRV-Instruction	40万件	幻覚（ハルシネーション）を抑えるために、正しい指示と誤った指示を混ぜたデータ。

3. 特定タスク・ドメイン特化型

特定の能力（OCR、図表、数学など）を強化したい場合に必須となるデータセットです。

- **図表・グラフ理解:**
- **ChartQA:** グラフの数値を読み取り、推論する。
- **DocVQA:** 請求書や文書の文字位置と内容を理解する。
- **空間認識・グラウンディング:**
- **RefCOCO / RefCOCO+:** 「左から2番目の赤い椅子」といった具体的な物体指定に対応。
- **推論・科学:**
- **ScienceQA:** 科学的な知識が必要な多肢選択式問題。
- **MathVista:** 画像に基づいた数学的な推論が必要な超難関データセット。

4. データセットを探すためのプラットフォーム

最新のデータセットは日々更新されるため、以下のサイトで「VLM」や「Vision-Language」タグをチェックするのが最も効率的です。

1. **Hugging Face Datasets:** `Multimodal` フィルタを使用。モデルとデータセットがセットで公開されていることが多いです。
2. **Papers with Code:** 各タスク（VQA, Image Captioning等）のリーダーボードから、SOTAモデルが何で学習されたかを確認できます。

5. 最近のトレンド：合成データ (Synthetic Data)

最近は、人間が作ったデータだけでなく、**「強力なモデル（GPT-4o等）に画像の詳細を書き出させたデータ」**で学習するのが主流です。

- **メリット:** ノイズが少なく、推論プロセスが丁寧。
- **注意点:** 元となるモデル（GPT-4など）の利用規約により、そのデータで学習したモデルの商用利用が制限される場合があります。

前処理

VLM（視覚言語モデル）の学習における前処理は、**「画像の処理」**と**「テキストの処理」**、そしてその両方を**「位置（空間）や文脈で紐付ける処理」**の3つに大別されます。

モデルが画像と文字を正しく関連付けられるかどうかは、この前処理の設計にかかっています。

1. 画像側の前処理 (Vision Preprocessing)

ビジョンエンコーダー（CLIPなど）に入力するために、生の画像をモデルが理解しやすい形式に変換します。

- **リサイズとアスペクト比の維持:** 多くのモデルは や といった固定サイズを要求します。ただし、単純なリサイズはアスペクト比を歪ませるため、**パディング（余白追加）やランダムクロップ（切り抜き）**が一般的に行われます。
- **正規化 (Normalization):** 画素値（0-255）を、モデルが学習された際と同じ平均値と標準偏差（例：ImageNet統計量）で正規化します。
- **パッチ分割 (Patch Embedding):** ViT（Vision Transformer）ベースのVLMでは、画像を などの小さな正方形（パッチ）に分割し、それぞれをトークンとして扱い

ます。

2. テキスト側の前処理 (Language Preprocessing)

画像に関連するキャプションや指示文を処理します。

- **トークナイズ:** LLMのボキャブラリに基づいてテキストを数値IDに変換します。
- **特殊トークンの挿入:** VLM特有の処理として、画像情報がどこに入るかを示すプレースホルダー（例：`<image>`）を挿入します。

> 例: `USER: <image>\nこの画像について説明して。 ASSISTANT:`

- **パディングとマスキング:** バッチ内の文章の長さを揃え、パディング部分をアテンション計算から除外します。

3. 空間情報の処理 (Visual Grounding / OCR)

特定の場所を指し示す能力（グラウンディング）を持たせる場合、追加の処理が必要です。

- **バウンディングボックスの正規化:** 画像内の位置情報を `[xmin, ymin, xmax, ymax]` の形式で抽出し、それを 0 から 1000 などの数値にスケーリングして、テキスト形式のトークンとして扱えるようにします。

> 例: `{"point": [250, 500], "label": "dog"}`

- **OCRエンジンの適用:** 文書理解（DocVQAなど）が目的の場合、あらかじめ外部のOCRエンジンでテキストの位置と内容を抽出し、それをプロンプトに含める処理を行います。

4. VLM特有の高度なテクニック

最近の高性能なVLMで行われている特殊な前処理です。

- **AnyRes (多解像度処理):** 高解像度の画像をそのままリサイズすると文字が潰れるため、画像を複数のタイルに分割して処理し、最後に統合する手法（LLaVA-v1.6などで採用）です。

- **データの増強 (Data Augmentation):** モデルの堅牢性を高めるために、画像にノイズを加えたり、反転させたりします。ただし、VLMの場合「右にあるものは何か？」という問いに対して画像を反転させると正解が変わるため、**テキストとの整合性を保った増強**が必要です。

5. 前処理のパイプライン（まとめ）

1. **データのクリーニング:** 画像の破損チェック、不適切なテキストのフィルタリング。
2. **サンプリング:** 画像1枚に対して複数のキャプションがある場合、どれを採用するか決定。
3. **トークナイズとリサイズ:** 画像とテキストをそれぞれのエンコーダーが受け取れる形式に変換。
4. **アライメント:** 画像トークンとテキストトークンを連結し、一つの系列（Sequence）にする。

まとめ

今回はVLMの学習に必要なデータセットについて説明しました。 VLMのデータセットはLLMに比べると、画像と文章両方の処理が必要となります。 前処理も含めてご参考頂ければと思います。