Image Captioning - Data layer

Datasets — Benchmarks & Sources

MS COCO Captions (2014/2017)

- **What it is:** ~123k images (train/val/test) with 5 human-written captions per image; the de-facto captioning benchmark (use "Karpathy splits" for comparability).
- **Why it matters:** Standard for training+evaluation (BLEU, METEOR, CIDEr, SPICE); wide scene variety.
- **Quirks:** Multiple caption references; long-tail objects; some noisy/underspecified text. Common practice: lowercase, strip rare tokens, or just rely on tokenizer.
- **Where:** Hugging Face (`coco_captions`), TFDS (`coco_captions`).

Flickr8k / Flickr30k

- **What it is:** 8k / 31k images with 5 captions each; photostream photos.
- **Why it matters:** Smaller/easier for quick baselines, teaching, or ablation studies.
- **Quirks:** More "people-centric" content; smaller vocabularies; risk of overfitting on Flickr8k.
- **Where:** Hugging Face (`flickr8k`, `flickr30k`).

NoCaps

- **What it is:** COCO-style evaluation set focusing on *novel object captioning* with out-of-vocabulary categories (leverages Open Images).
- **Why it matters:** Tests generalization to unseen objects—crucial for real-world deployment.
- **Quirks:** Requires strong visual grounding; benefits from large pretraining (CC3M/12M, LAION).
- **Where:** Hugging Face (`nocaps`) for annotations; images align with Open Images.

TextCaps

- **What it is:** ~28k images emphasizing *scene text* with 5 captions each.
- **Why it matters:** Evaluates OCR-aware captioning (reading storefronts, signs, menus).
- **Quirks:** Needs OCR features or models with text tokens; vanilla captioners underperform.
- **Where:** Hugging Face (`textcaps`).

VizWiz-Captions

- **What it is:** Images captured by blind/low-vision users; 5 captions each.
- **Why it matters:** Tests robustness to blur, occlusion, poor framing; socially impactful.
- **Quirks:** Very noisy visuals; safety/harms considerations; shorter, pragmatic captions.
- **Where:** Hugging Face (`vizwiz_captions`).

Conceptual Captions 3M (CC3M)

- **What it is:** ~3M image–alt-text pairs harvested from the web (Google AI).
- **Why it matters:** Large-scale weak supervision for pretraining image–text encoders/decoders.
- **Quirks:** Noisy, diverse, sometimes non-descriptive or templated alt-text; heavy filtering recommended.
- **Where:** TFDS/HF mirrors often labeled `conceptual_captions` (availability may vary; follow dataset card instructions).

Conceptual 12M (CC12M)

- **What it is:** ~12M web image–text pairs (Google AI).
- **Why it matters:** Scale helps zero-/few-shot generalization and long-tail vocabulary.
- **Quirks:** Higher noise; dedup + quality filters improve results; licensing carefulness required.
- **Where:** Hugging Face (`conceptual_captions_12m`) or instructions via dataset card.

SBU Captions (SBU1M)

- **What it is:** ~1M image–caption pairs (web alt-text).
- **Why it matters:** Classic pretraining set predating CC; still useful as supplemental pretrain data.
- **Quirks:** Web noise; domain shifts; shorter captions.
- **Where:** Hugging Face (`sbu_captions`).

Visual Genome (Region Captions)

- **What it is:** Dense annotations (region descriptions, attributes, relationships) for ~108k images.
- **Why it matters:** Enables *dense captioning* and grounding-aware pretraining.
- **Quirks:** Region captions are short/fragmented; alignment to full-image captions needs care.
- **Where:** Hugging Face (`visual_genome`).

Open Images – Localized Narratives (Google AI)

- **What it is:** Free-form spoken+text “narratives” with mouse traces grounding words to regions.
- **Why it matters:** Great for grounding/attention supervision and richer, paragraph-style captions.
- **Quirks:** Spoken → transcribed text; variable quality/length; needs alignment to images.
- **Where:** Google AI (TFDS: `localized_narratives`), some HF mirrors.

LAION-400M / LAION-5B (Alt-text Pairs)

- **What it is:** Massive web-scale image–text pairs created via CLIP filtering.

- **Why it matters:** Pretraining backbone for many SOTA captioners; strong coverage of rare entities.
- **Quirks:** Web noise; ethical/safety filtering essential; dedup strongly recommended.
- **Where:** Hugging Face (`laion` subsets; follow dataset cards).

> **Note on OpenAI:** OpenAI does not distribute proprietary caption datasets. You typically train/evaluate on the above public sets, and you may *evaluate/infer* via OpenAI APIs if desired.

Preprocessing (what to do and why)

Resizing & Cropping

We make images a consistent size while preserving content focus.

- **Train:** `RandomResizedCrop(224–384)` → Encourages robustness to scale/position; typical inputs 224–384px depending on backbone.
- **Eval:** `Resize(shorter=256/384) + CenterCrop(224/384)` → Deterministic evaluation; matches encoder's expected input size.

Light Augmentation

We add minimal perturbations that don't change semantics.

- **HorizontalFlip(p=0.5)** → Safe for most scenes; avoids label mismatch (captions shouldn't become wrong).
- **ColorJitter (very mild)** → If used at all, keep tiny deltas; aggressive color/blur can invalidate text in captions.
- **Avoid heavy RandAugment / Cutout** → Can contradict captions (e.g., removing objects the caption mentions).

Normalization

We adjust pixel values so they're centered and scaled, making training stable.

- **ImageNet stats:** `Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])` → Matches most pretrained vision encoders (ViT/CLIP/ResNet).
- **From scratch:** `Standardize per dataset` → If no pretrained weights, compute dataset mean/std; improves convergence.

Tokenization / Feature Extraction

We convert text to token IDs and images to encoder features that the decoder can attend to.

- **Processor unification:** Use model's `AutoProcessor` (e.g., BLIP/BLIP-2/ViT-GPT2) → Ensures identical image transforms + text tokenization to pretraining.

- **Text cleaning (optional):** lowercase, strip extra spaces; keep punctuation unless model card says otherwise → Modern tokenizers handle punctuation; over-cleaning can remove useful cues.
- **Max length:** `max_length=64–128` (dataset/model-dependent) → Long captions truncate; set `truncation=True`, and consider `min_length` for beam search at eval.

Special Cases (OCR / Novel Objects / Dense Captions)

- **OCR features (TextCaps):** Precompute OCR tokens/boxes (e.g., Tesseract/EasyOCR) and fuse as extra tokens → Improves reading text in images.
- **NoCaps:** Include open-vocab/CLIP-style pretraining and class name prompts → Helps recognize unseen categories.
- **Visual Genome (regions):** Sample multiple regions per image and train a region-caption head → Encourages fine-grained grounding.

Dataloading tips

We prepare the dataset so training is fast, reproducible, and efficient.

- **Prefetch & pin memory:** `DataLoader(pin_memory=True, prefetch_factor>1)` → Keeps GPUs busy; lowers host–device stalls.
- **Num workers & caching:** `num_workers=4–16`, cache decoded images/features when possible → Big I/O win on web-scale datasets.
- **Worker init functions:** `worker_init_fn=seed_all` → Makes random crops/flip reproducible across epochs/machines.
- **Deterministic validation:** `Resize + CenterCrop + Normalize` with fixed seeds → Fair, comparable metrics.
- **Caption sampling:** If multiple refs, sample 1 at train, keep *all* for eval → Matches standard metrics that average over multiple references.
- **Dedup & filtering:** Remove near-duplicates, extremely short/long or templated alt-texts → Boosts quality for noisy web corpora (CC/LAION/SBU).
