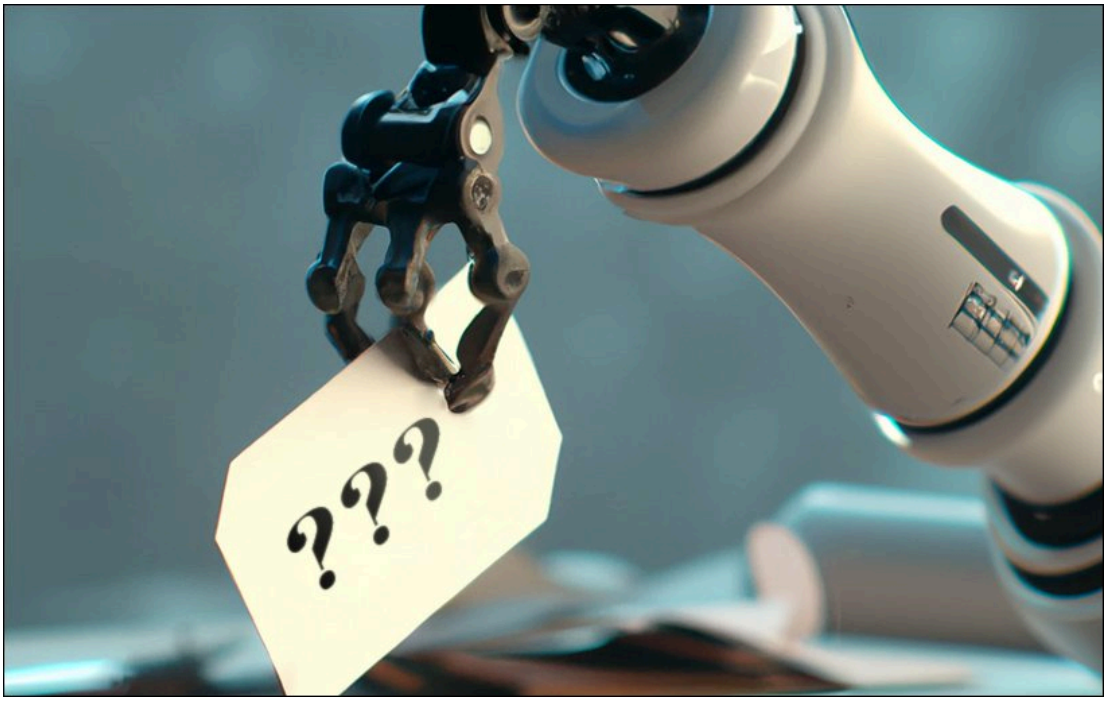
Image Synthesis Has an SEO Problem

Originally published at https://arquivo.pt/wayback/20240203190757oe_/https://blog.metaphysic.ai/image-synthesis-has-an-seo-problem/

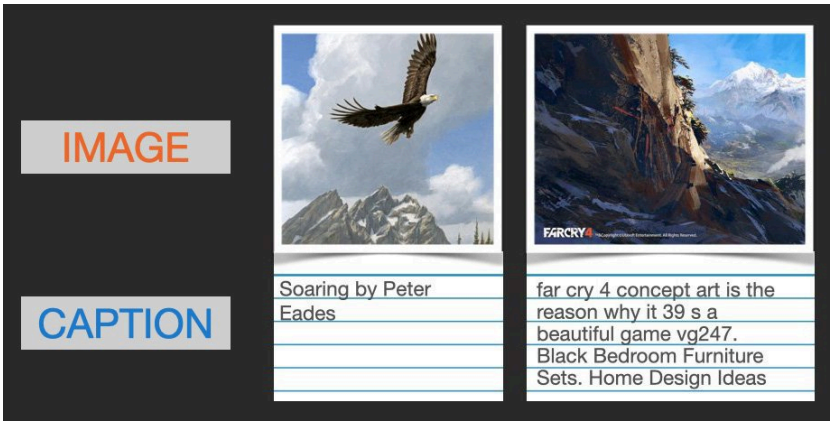
January 24, 2023



As training datasets get ever larger, it's becoming clear that domains in machine learning have an SEO problem. But it's not the usual SEO problem, and these are not the usual domains.

Rather, multimodal generative systems such as Stable Diffusion, which are trained on literally billions of images scraped from the internet, have to use whatever associated text (such as file-names or captions) the web-scraping program could find when the images were retrieved, because it's too expensive to have actual people write descriptions for so many images manually, for the purposes of training the dataset accurately.

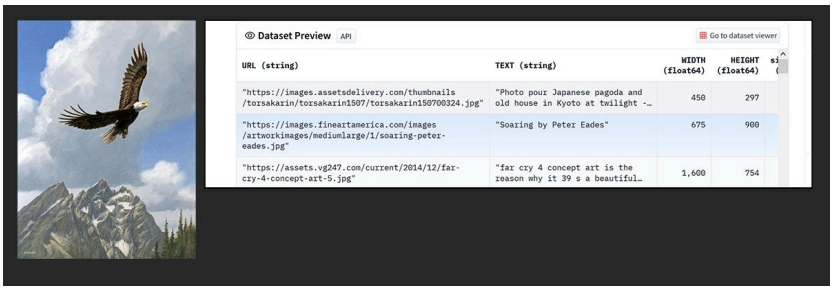
Frequently, the motivation behind those pre-existing names and captions is at odds with the kind of granular and methodical image description that would be most valuable to an image synthesis framework, and which would help it to extract useful features, and to learn to make more accurate and meaningful images in response to user input.



Examples of less-than-deal labels which end up as source material for generative AI systems. Source: https://jalammar.github.io/illustrated-stable-diffusion/

In the examples above, we see two of the many possible problems with using web-scraped captions. In the first, the caption is accurate, but unhelpful, in that it describes the action being performed, but not the creature performing it, the environment, nor any of the context, such as a description of the background (though arguably this caption would help Stable Diffusion users to imitate an artist's style, since it correctly names the artist who created the image).

The second is *initially* accurate (the image has some relationship to the game *Far Cry 4*, and to 'concept art'); but it then takes a deceptive turn, appending irrelevant labels related to *furniture retail* – a classic misuse of image labeling and abuse of SEO practices that's designed to hijack a high-traffic SEO trend and boost both the domain that's re-hosting the images and which has added the deceptive tags.

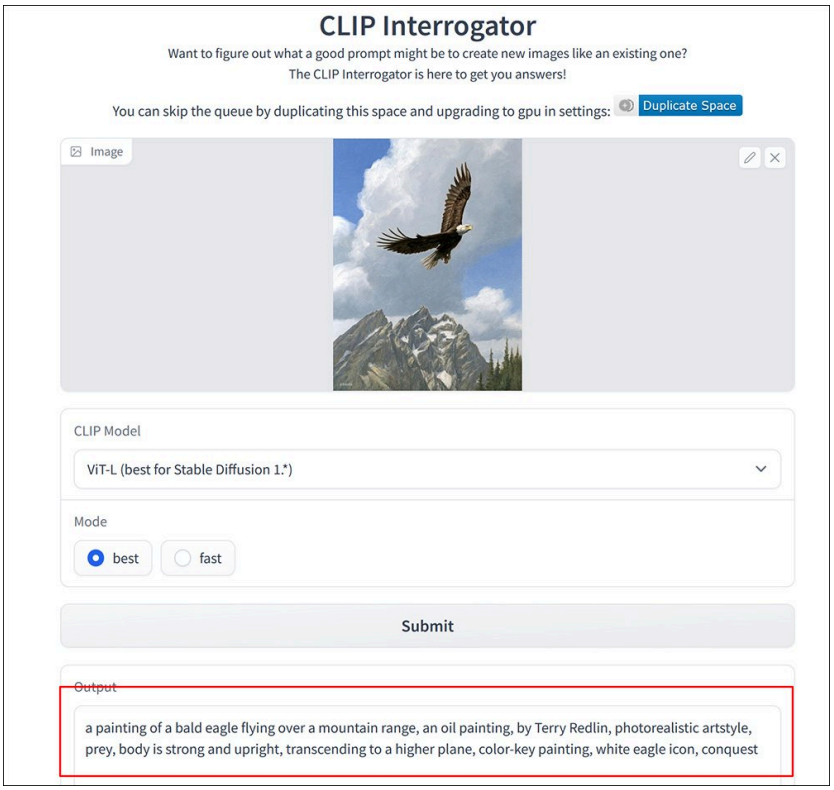


The origin of what Stable Diffusion knows, from training, about the Peter Eades painting of a Mountain Eagle titled 'Soaring'. Sources: https://huggingface.co/datasets/ChristophSchuhmann/improved_aesthetics_6.5plus and Fine Art America/Pinterest

In the image above, we can see that someone (or some algorithm) at the website Fine Art America has created the caption for this image, and that this caption has subsequently been ingested into the dataset labels for a sub-class of LAION, the database that powers Stable Diffusion.

That does not mean that more relevant data can't be extracted for inference purposes: cross-category recognition and pixel-based feature recognition may enable facets of this painting to enter into other categories and classes (such as *'bird*', *'eagle'*, *'mountains'* and *'art'*), but it's no thanks to the sparse caption.

In the image below, we can see the multimodal interrogator CLIP devising a better description of the above image (even though. among other issues, it gets the artist wrong, and adds the anomalous term *'conquest'*, possibly based on other similar artwork that is more relevant to that term):

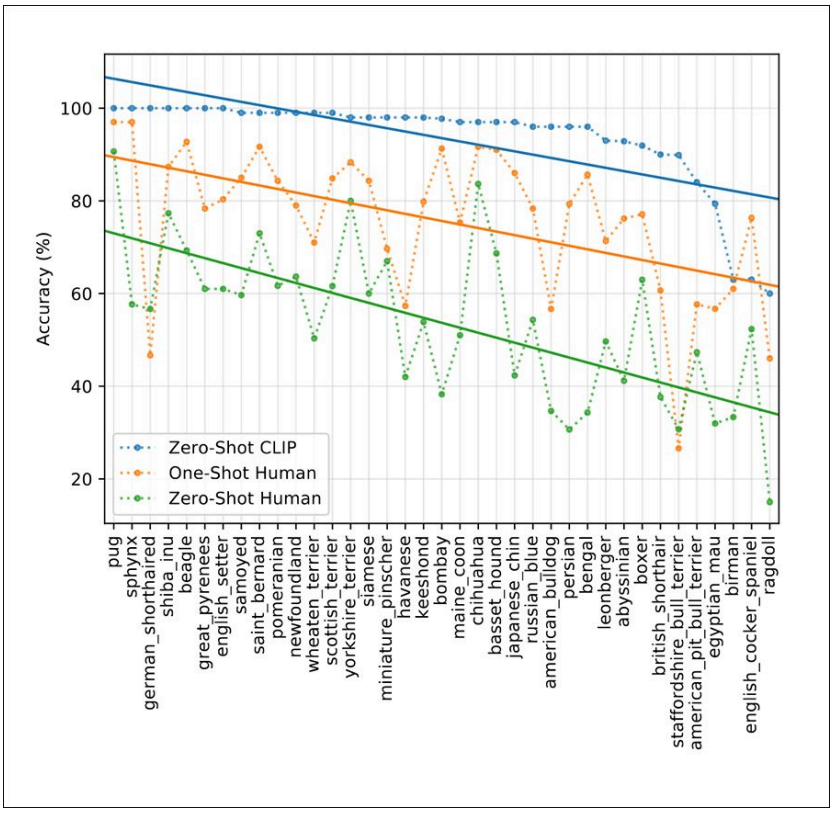


Source: https://huggingface.co/spaces/pharma/CLIP-Interrogator

In general, this does not signify that the bad labels are all that's on offer during training; the process of CLIP interpretation occurs also during core training of a Stable Diffusion model, via a masked self-attention Transformer; so the labels form only part of the 'sense' that the architecture makes of the data.

The weights extracted from other images can be utilized to improve the semantic integrity of under-labeled or mislabeled data – *if* there are enough common features or lexical terms between the image/text pairs; otherwise, pixel and text information that could have been used to help quantify elements in the badly-labeled image will not be called into service.

When an image is both poorly-labeled and visually distinct enough to be an outlier, it is effectively isolated from integration into the classes and labels to which it belongs, even if – as in hyperscale datasets such as LAION – such 'compatible' image/text pairs are abundant.



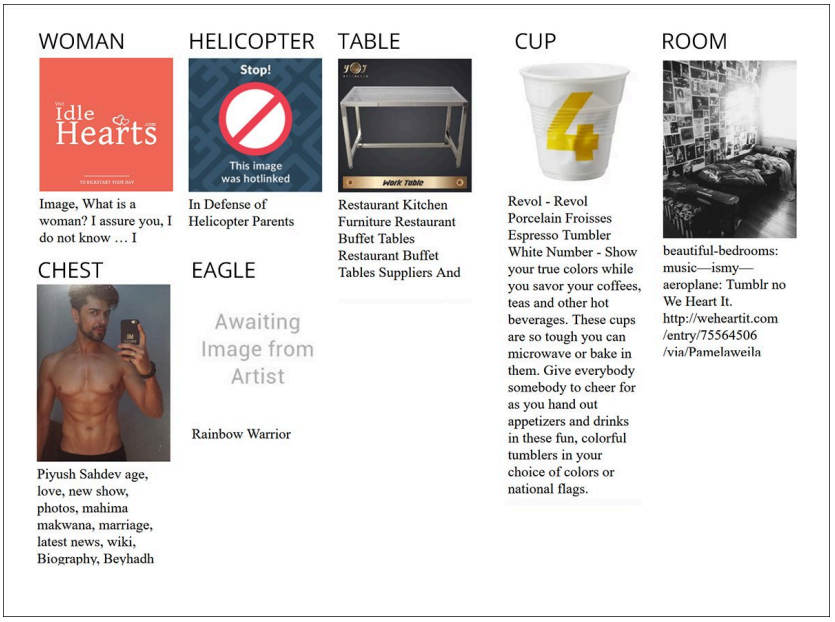
Bad labels affect performance for both humans and automated systems such as CLIP. Here, in the official Stable Diffusion release paper, we see humans and CLIP alike struggling to identify and codify badly-labeled images (the term 'zero-shot' refers to predictions made without labeled data; 'one-shot' refers to predictions made after exposure to prior examples with labels). Source: https://arxiv.org/pdf/2103.00020.pdf

By example, we can see, in the earlier images above, that the *post facto* CLIP interpretation of the eagle painting has misidentified the artist (who is actually *correctly* labeled in the dataset) whilst providing a much more useful and applicable description of the scene itself. If all that information had been present in the original label, errors and misassociations would be reduced.

Therefore the training process and the model itself would have benefited from a more apposite description, both for retrieval of the image concepts in themselves, and for the greater usefulness and accuracy of the generative system.

This is a problem that feeds upwards, as it were, since most of the non-human solutions (including CLIP) devised to address it are themselves algorithmic, and at the mercy of 'available' labels, due to the scale of the challenge and the lack of money to address it.

Though there are *not* likely to be enough instances of *'mountains with furniture'*-style errant labels in a dataset to produce a model that will spit out an IKEA showroom for the text-prompt *'a mountain view'*, this kind of misassociation and mislabeling compromises the integrity of a latent diffusion system – and it's easy to see how a determined mislabeler could actively poison future datasets through this approach.



Some tangential captions, and their associated images, in various LAION sub-sets. Source: https://rom1504.github.io/clip-retrieval/?

SEO Motivations

In practice, if there is any ill-intent at work here, it has little to do with machine learning, and all to do with getting low-value websites to rank higher in search engine result lists.

The slew of guides to SEO-friendly image-labeling, naming and captioning, constitute a broad schema for what a dataset's annotations are likely to resemble, while interest in optimizing images for SEO has only deepened since 2019, when Google's Gary Illyes revealed that Google considered media-based search 'too ignored', and would be increasing the search focus on images and video.

The previous year, Google removed the 'View image' button from default image search results, forcing the average user to visit the host site in order to view images found in their web searches, and increasing SEO industry attention on the use of 'collateral text' in web images.

Since the 'click-through funnel' in search results is increasingly from image-based web searches, there is currently a broad perception that an image's text attributes (names of images, alt-text, and associated captions) are a potential way to rise up the rankings.

Google's take on how these aspects should be handled is, unsurprisingly, quite friendly to ML-facing dataset logic and semantics; after all, the company has a vested interest in this, and the core principles are not very different to best practices in general SEO (such as using tags, names and captions that describe the image accurately).

Stuff That

So that's what Google wants, and what it's hoping to leverage in offerings such as multisearch. What the putative traffic-hound wants, often, is to exploit or even subvert these principles to gain undeserved SEO traction. Thus, some of the edgier SEO blogs continue to skirt or even break through the intended guidelines of good practice in image/text usage, and promote the use of specious or irrelevant tags and names (hijacking), and tedious repetition of terms for which good rank is desired (keyword-stuffing).

Though keyword-stuffing (as in the *'mountain furniture'* image example above) can, according to Google, harm your search-engine ranking, the web-scrapers that feed machine learning datasets are not intended for direct consumer usage, and are less discriminating; additionally, a web-scraper may not be using SEO ranking as a discriminating metric.

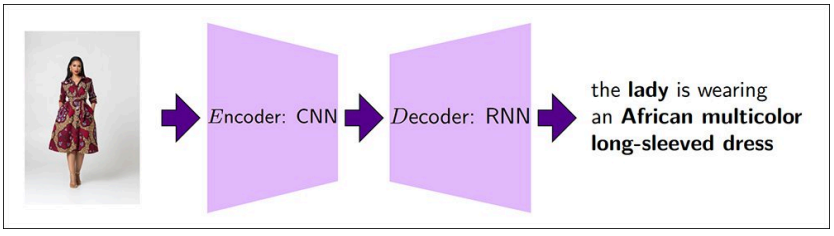
Nor should it, some argue; the current, growing conviction that high-value content may not be ranking well under Google's current approaches means that scrapers can't afford to ignore material just because it is poorly-ranked; and, by a coincidence, the backwaters of any web search is also where a great deal of trashy and low-value content lives.

For this reason, the keyword-stuffed images and poorly-labeled photos that are ostracized in ranking results are likely to get a 'free ride' into the average AI dataset – and at that point may even have a lexical advantage in the gauntlet of pre-training filtering (because the spammed text content is tantalizingly longer, as if it were an effortful description – and usually mixed with valid keywords as well, as in the *Far Cry 4*/furniture example earlier).

AI-Based Captions and Image Names

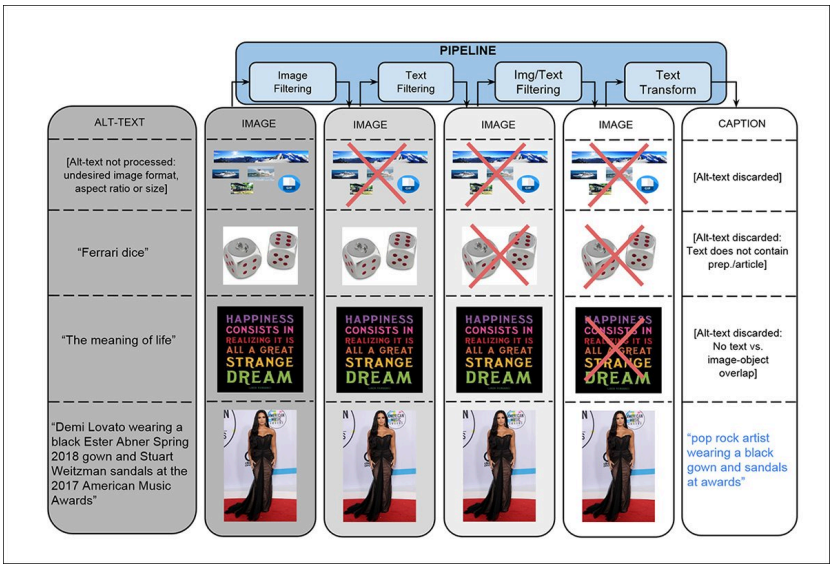
It's worth mentioning that not all bad or apparently 'subversive' image captions constitute a conspiracy; like AI research scientists, website owners with a high volume of daily image uploads have limited captioning resources, and either rely on the kindness of uploaders in this respect, or else on automated and even AI-driven systems.

The fashion industry is currently in the vanguard of AI-driven captioning, since it's an affluent sector that's well-represented in computer vision research publications. Because it deals with a very limited domain (clothing), fashion-focused AI systems are accurate enough to produce workable auto-captioned images – which will, inevitably, help to feed future AI datasets.



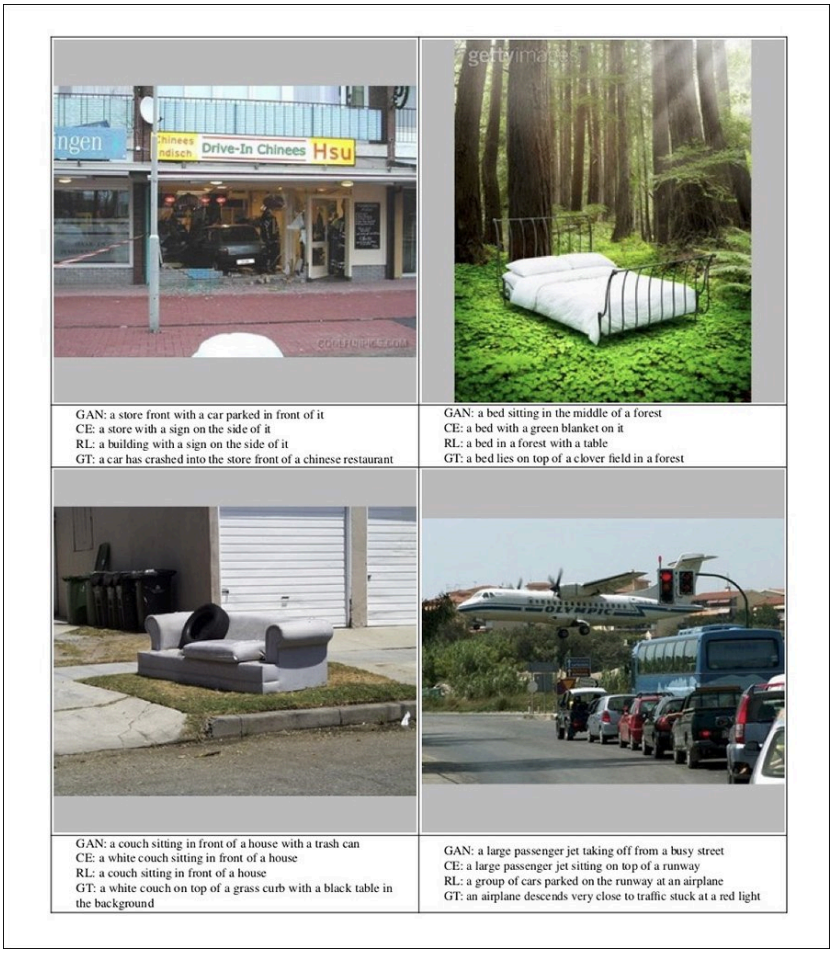
A diversity-aware automatic fashion captioning system. Source: https://files.osf.io/v1/resources/hwtpq/providers/osfstorage/60d17cc6f44078004575cef7

Some of the most ingrained and persistent captions in machine learning research datasets can potentially be improved and essentially retrofitted or rewritten to be more granular, accurate and useful. To this end, Google Research released the Conceptual Captions dataset in 2018, which is capable of recognizing non-apposite captions (such as text inside an image), and of creating broader labels that are more widely useful and likely to be related to existing classes in image-centric machine learning systems.



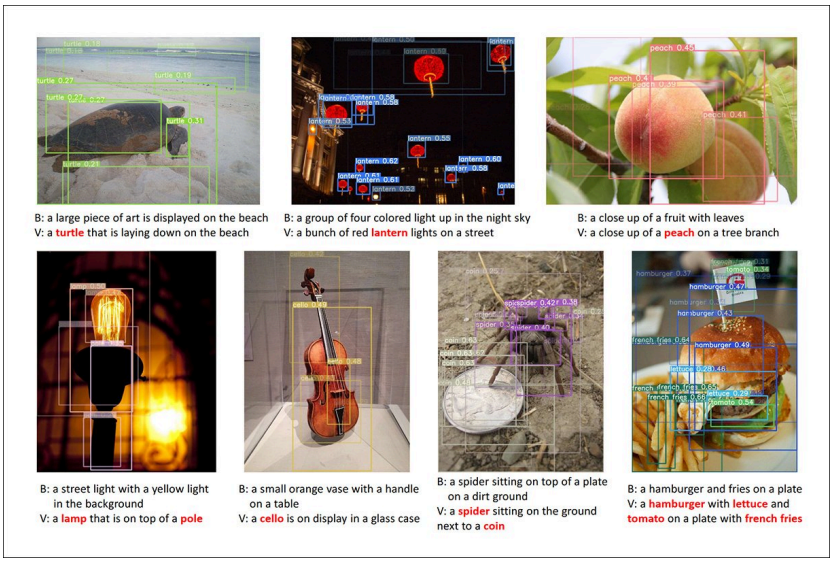
The Conceptual Captions pipeline amends some of the worst SEO or inadvertent mislabelings. Source: https://aclanthology.org/P18-1238.pdf

In June of the following year, IBM research pitched into AI-based image captioning with a system that leveraged Generative Adversarial Networks (GANs), though ultimately conceding that human-based captioning capabilities remain 'the gold standard'.



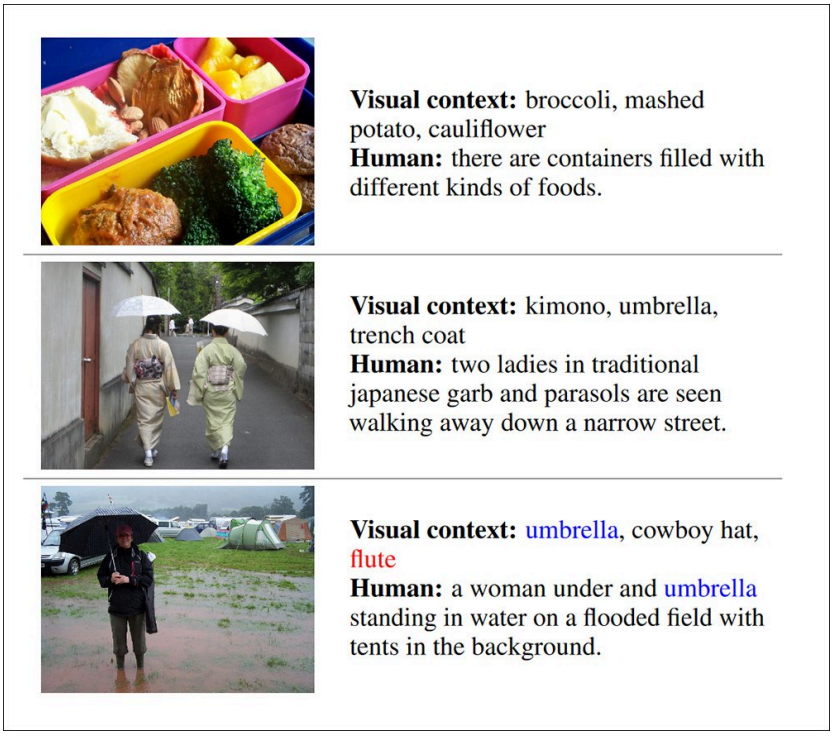
IBM's GAN-aided system is better than rival systems tested (middle two lines underneath each picture) on out-of-context images, but falls far short of the human-written ground truth (bottom-most line, 'GT', underneath each picture). Source: https://www.ibm.com/blogs/research/2019/06/image-captioning/

In 2019, Microsoft announced a breakthrough in AI-based image captioning, via the pretraining of a multimodal language model with a rich image dataset paired with word tags (originally labeled by humans), mapped to specific locations where the related object occurs in the image.



Human-led pretraining allowed Microsoft to develop a more accurate annotation and labeling system for images. Source: https://arxiv.org/pdf/2009.13682.pdf

In January of 2023, a research collaboration from Spain proposed a method (producing a publicly available dataset) to 'refit' the limited labels of the influential *COCO Captions* dataset to provide more descriptive detail:



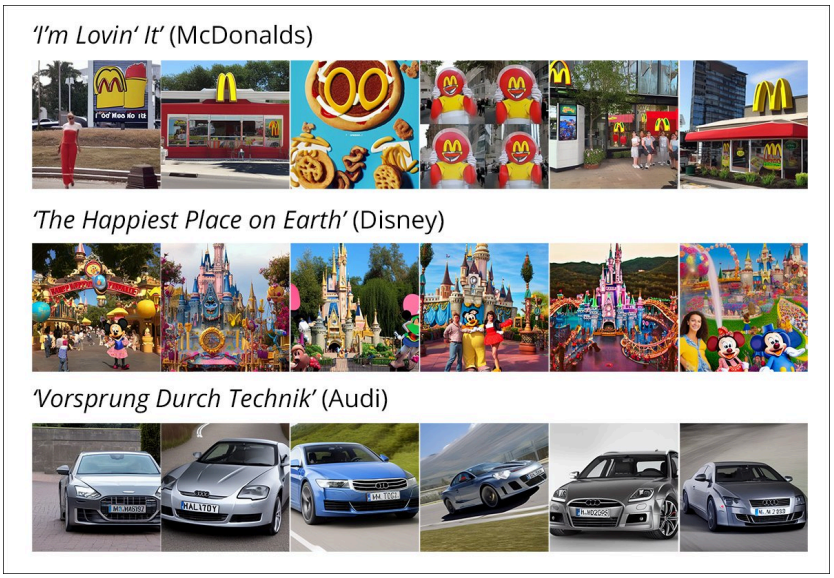
Revisions and upgrades of captions from the original COCO Captions dataset, with words common to both versions indicated in blue. Source: http://export.arxiv.org/pdf/2301.08784

The Spanish initiative is only one of a slew of current projects aiming to reform the minimal or inadequate captioning of 'historical' and influential computer vision datasets. An inhibiting factor on uptake, of course, is that engaging with the revised captions instead of the original captions requires comprehensive adoption within the broader research community, since the 'improved' captions do not have 5-10 years of provenance across tests in publications.

Why Not All Captions are Descriptive

Machine learning systems benefit most from multimodal data (images and text are the usual, though not the exclusive modalities here) where the text component is descriptive, and concisely describes what is in the scene. Any 'emotional', abstract, interpretive or poetic take on the image is likely to create implicit associations without actually stating what the facets of the image actually are. Such a relationship is valid, but unhelpful, and is part of the syndrome of *shortcut learning*.

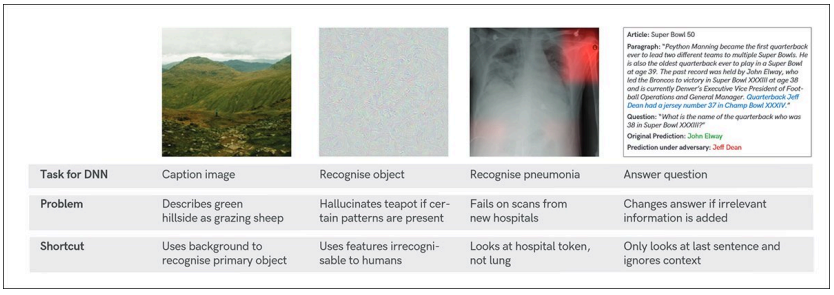
Thus, anomalies can occur where abstract prompts can become associated with very specific material, because there may be no other visual content that's as strongly associated with the term.



Various ultra-famous corporate slogans passed through Stable Diffusion (V1.5), non-cherrypicked results, no negative prompts used.

In the above examples, which were rendered in base V1.5 Stable Diffusion, and represent the sole and complete output of each prompt, the corporate slogans would clearly not be stumbled upon by accident. But the problem with non-descriptive associations is that, unlike in the above cases, it may never be clear how the association formed.

Shortcut learning will, for example, 'cheat' by learning the context for concepts, such as the likelihood that a pavement or floor will be the background for a picture of a dog (because most pictures of dogs are taken from above, and extensively feature floor surfaces).



Examples of shortcut learning, where non-relevant or unreproducible facets associated with a concept are used as a 'cheap' way of referencing, indexing or assimilating the object. Source: https://arxiv.org/pdf/2004.07780.pdf

Part of the problem with web-found images as AI data is that the intention and motivations for the names of image files, and their associated captions or annotation, are not always well-defined – even to the person creating the text content.

In the computer vision research sector there is a clear distinction between 'descriptions' and 'captions', in that a description can entirely describe (and effectively substitute) the image, whereas a caption is an *additional aide* to understanding the image (and therefore is not as semantically discrete or disentangled). One recent paper, studying caption quality and its relationship to CLIP, notes that the art historian Erwin Panofsky defined a caption as something that 'provides personal, cultural, or historical context for the image'.

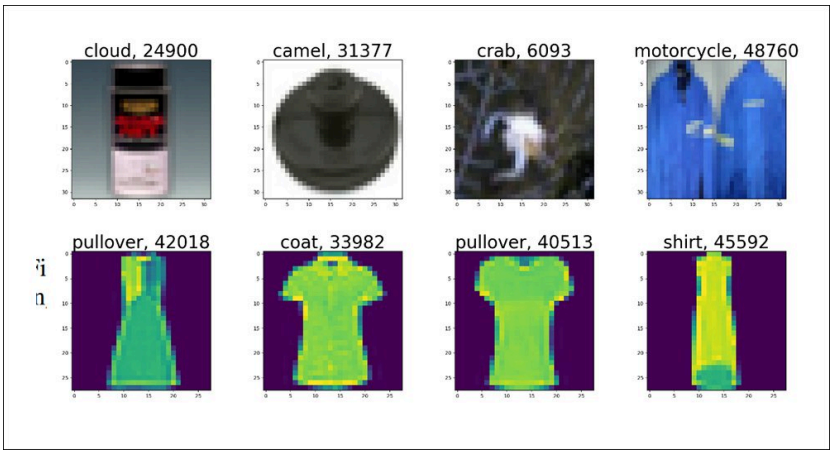
Since descriptions are more labor-intensive than captions, caption-style content tends to dominate AI-facing web-scraped image datasets, which means that the entanglement of image and text content could be said to be largely preexistent in the source material.

"I'm not sure for these generative models," says Nicolas Müller. "that mislabeled data is a huge problem. For me, the bigger problems are shortcuts – that classifiers learn to rely on spurious correlations, which don't translate to the real world."

Müller is a research scientist at Fraunhofer AISEC, and the co-author of a recent paper, *Identifying Mislabeled Instances in Classification Datasets*. Among other lines of research, Müller also investigates deepfake audio, and has found evidence of the extent to which the training of classification systems can be subverted by irrelevant or specious signals in the data.

"There," he says "we found an error in a very popular data set. The label corresponded massively with the amount of *leading silence* in the audio file. So, if there was a lot of silence, the audio was supposed to be benign. Of course, this isn't a relevant signal, and assigning importance to it breaks the whole model – and they have to redo everything."

"Shortcut learning is definitely under-studied," he observes. "Often, in machine learning, scientists just like to take data and train models on it, and they don't care about *why* it works."



Top, instances of mislabeling in the CIFAR-100 training dataset. Bottom, similar mistakes in the Fashion-MNIST dataset. Source: https://arxiv.org/pdf/1912.05283.pdf

Trying to redress the balance with AI is a tautological challenge, since most of the upstream libraries one could possibly use are in themselves susceptible to shortcut learning and the other vagaries of reliance on web-scraped, unsupervised or under-supervised data.

"So," says Müller, "considering an ideal system you might want to build, you could use these systems. But at the same time, that's the system you would *need* to correct a given data set."

"People sometimes propose that you have a significantly smaller dataset, and you train the system, fix the labels, and then iterate. But from an information theory background, I'm not sure if this really helps, because then the same information that you're learning during the training has been input previously by this very same model. So, you're trying to create information out of thin air, if you will."

Regarding the sparseness of detail so often found in web-scraped captions, Müller believes that much more thorough captioning/annotation is needed, notwithstanding how it's generated, and also that part of the way forward lies in choosing better-annotated source data, such as material from Flickr, Pinterest and Instagram – or biting the bullet and paying for the content legitimately (though legitimacy is currently a moot point) .

"One possibility would be to pay for high quality data, for example stock photos" Müller says, "Compared to crawling images from the web, this could yield higher-quality, but at the same time costly training data."

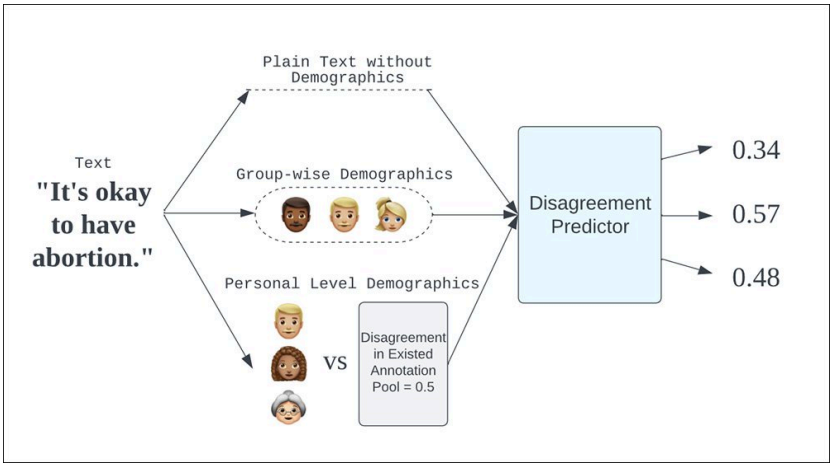
The Quality of Crowdsourced Labeling

Image access is not the only cost being avoided by cash-strapped researchers, because the 'nuclear option' for improving label and annotation quality is to pay people to create the labels, at high volume.

At LAION-scale (i.e., Stable Diffusion), that's a notable financial commitment; even if you could get the lowest possible price for each HIT (i.e., a single image annotation costing 1 cent), the requisite budget would be a minimum $23 million USD for around 2.3 billion images. For that money, which is very unlikely to be obtainable for all images, and which is considered peanuts, you might get mostly monkeys (since finding good annotators at any price is challenging).

But even if you do get competent people, are they qualified to make the value judgements that are being asked of them?

Research scientist Ruyuan Wan of the University of Notre Dame in Indiana is the lead author of a recent paper examining the possible necessity of demographic vetting of crowdsourced annotators – with a particular emphasis on ensuring that 'minority voices' are not dismissed in cases where annotators' opinions differ on the same data.



Annotator disagreements are not always solved in the fairest or most rational way, and the new University of Notre Dame paper suggests ways forward through the problem. Source: http://export.arxiv.org/pdf/2301.05036

"The central idea," Wan told us, "is that very often, when there is a disagreement among annotators, a straightforward consensus system can silence the voice of the person with the most informed take on the data."

The system Wan supports is called *collaborative annotation*.

“When there’s a dispute, it gets prioritized rather than sidelined or abandoned, because conflict can be hiding interesting insights into the data, or even reveal that the schema for the annotation may need revision.”

“For instance,” she continues. “many people think that ‘simple’ annotation is an obvious task such as saying whether an animal you’re being shown is a cat or a dog. But if you mix in real and synthetic or stylized images, the response of the annotator might be *‘that’s not a cat, that’s a cartoon’*, which you may not have anticipated in the allowed list of responses.”

“In cases like that, you could be getting a deceptive negative or positive. Only studying conflict across annotators can bring that kind of misinterpretation or bad survey design to light.”

Understanding more about the background of an annotator can also reveal that they may be the only person in a dispute-group of three with an informed opinion on the data being evaluated, and that the consensus system is drowning out their (correct) take on the data in the same way as any other unfortunate signal-to-noise ratio.

Another risk in traditional consensus systems for crowd-sourced annotation is that the under-resourced researchers may simply throw out any cases where disagreement arises. As a 2021 Google Research paper has noted, *‘disagreement between annotators, which is often viewed as negative, can actually provide a valuable signal’*.

Wan also believes that the research community’s growing dependence on Amazon Mechanical Turk needs reconsidering.

“MTurk is the most famous platform,” she says, ” but it’s not the only one. Other platforms may be more focused on specific domains, and smaller groups, and may be a better and more representative solution for certain projects, with greater commitment to the task.”

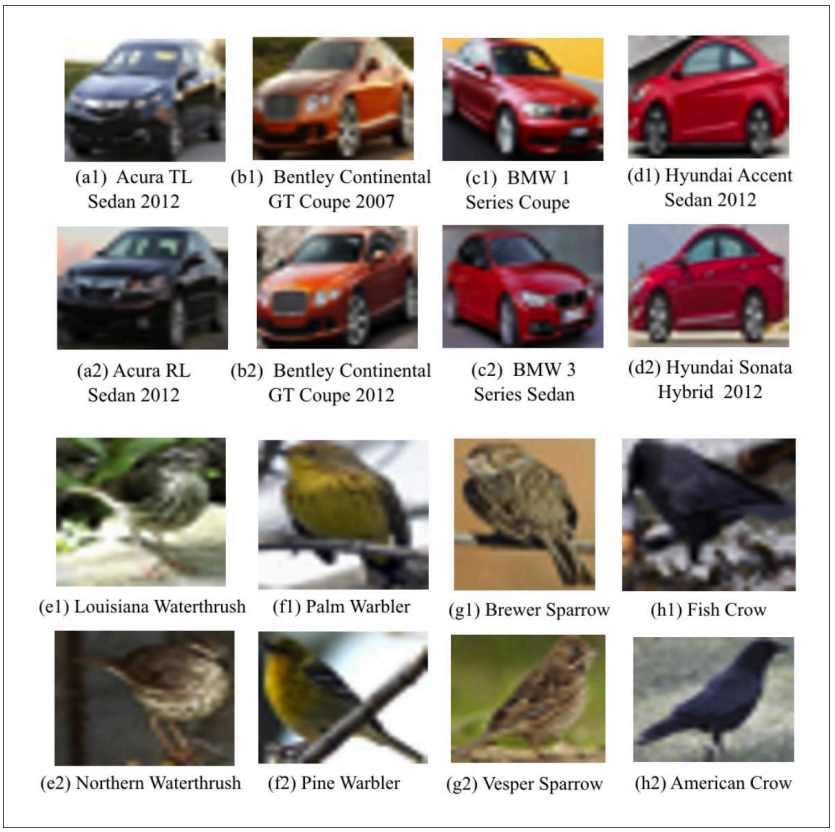
The Influence of Image Compression on Training

One final SEO consideration for web-scraped data is that images taken directly from public-facing internet sources may be extraordinarily compressed, with JPEG artefacts and over-sharp details.

At a certain level of optimization, images can be so degraded that classes and objects can't be efficiently extracted from them. Source: https://hal.science/hal-01196958/file/Chevalier_ICIP15_LRCNNvaryingresolution.pdf

Over-optimized images were traditionally used in order to decrease a site’s initial loading time, which has long been a factor in search engine ranking.

This problem crops up most in some of the oldest and most venerable computer vision datasets from the pre- or early broadband era, when optimized image sizes and resolutions were critical for a website’s basic usability.



Examples of very poor-quality and highly-compressed image sections, from the Stanford Cars and UCSD-Caltech Birds 200-2011 benchmarks. Source: https://arxiv.org/pdf/1703.05393.pdf | http://ai.stanford.edu/~jkrause/cars/car_dataset.html | https://authors.library.caltech.edu/27452/

However, even if the current standard and size of images has notably and consistently increased over the last ten years, many of those older and heavily optimized images are still there, and still likely to be ingested even into the newest datasets. To boot, they're most likely to still exist at domains with the greatest long-term resilience and authority, which are likely to be 'favored sources' for image datasets.

The BBC maintains unaltered and 'unimproved' historical posts, which provide abundant low-quality images that nonetheless may be available from no other source, or at better resolution. Source: http://news.bbc.co.uk/2/hi/uk_news/970354.stm

Ten or more years ago, in the earliest days of GPU-driven AI training, resolution and image quality was less important, since the broad shapes visible even in poor-quality images were adequate for conversion into the 64x64px training images that were typical of the architectures of the day. At the same time, the nature of generalization and upsampling could compensate for the poor quality, though only within certain limits in cases of very low-res or excessively optimized input data.

However, with 512x512px and 1024x1024px now being made possible with better GPUs and more efficient frameworks, researchers are increasingly expecting to extract and synthesize *detail* from web-scraped image data, rather than just representative shapes, while 'legacy' LQ data has given rise to a distinct sub-sector of image/video synthesis – super-resolution: the pursuit of high-resolution data from low-resolution originals.

Conclusion

The dictates and requisites of good (or at least ‘effective’, if not ‘good’) SEO practices are frequently at odds with the needs of computer vision datasets. Even where many of the pivotal players involved (such as Google, Facebook and Microsoft) have a notable stake in each camp, the companies struggle to reconcile the conflicting requirements and policies related to their search engine vs. machine learning needs.

Therefore the research sector continues to seek out economical ways to increase description and caption quality without the need for hyperscale manual labeling, and the expense and vision involved therein. A casual browse of recent literature, much of it related to CLIP and associated technologies, reveals that the sector continues to hope that effective re-captioning or re-interpretation systems will be developed without the need (or with minimal need) for expensive human labeling.

Worse, even with notable financial resources, many of the current standards and practices in crowd-sourced labeling indicate that throwing money at the problem without reconsidering the infrastructure, schemas and conditions involved, would not represent a definitive or totally effective solution.

Systems relying on the relationship between text and images have ballooned in importance and prominence in less than a year since the release of DALL-E 2, while the extraordinary and culturally far-reaching phenomenon of the open source Stable Diffusion framework has further magnified the significance of label and general annotation quality in computer vision datasets. Suddenly, it’s become very important that images have better labels, and there are no equally sudden solutions to the issue.