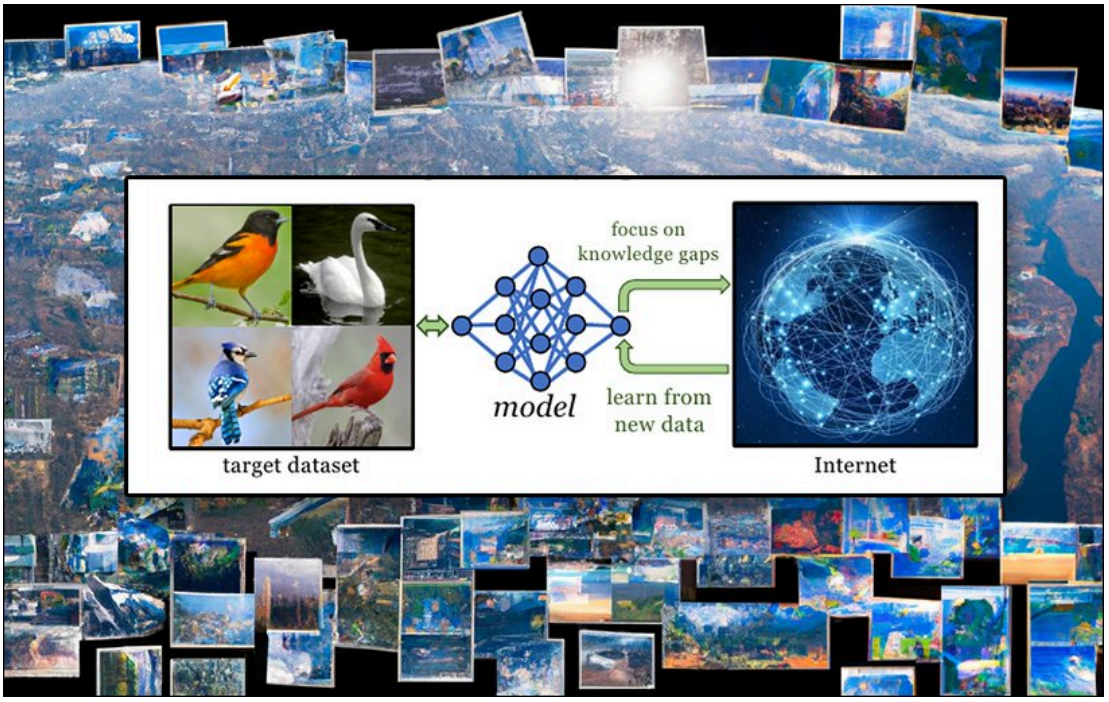
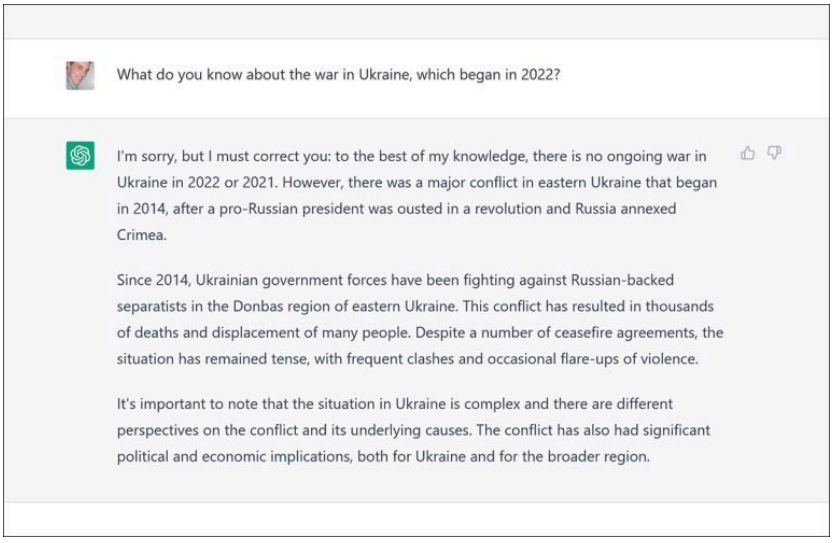
Using the Internet as a ‘Live’ Dataset for AI Training

Originally published at https://blog.metaphysic.ai/using-the-internet-as-a-live-dataset-for-ai-training/

March 2, 2023

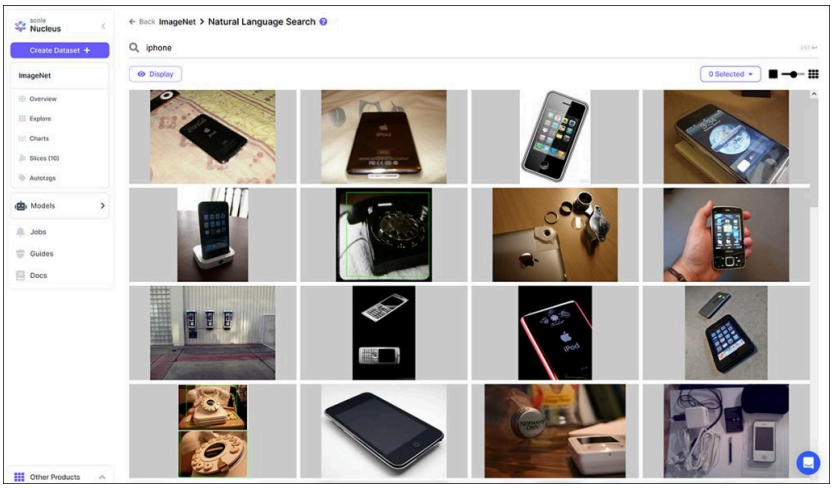


For cutting-edge technology, trained AI models can be surprisingly ‘historical’. Asking the much-feted ChatGPT about the war in Ukraine, for instance, is a hopeless task:



ChatGPT 'corrects' me about the non-existence of the Ukraine war (Wednesday, March 1, 2023).

OpenAI’s revolutionary Large Language Model (LLM) was trained in 2021, prior to this major news event.
Like language models, other types of generative AI models are essentially ‘frozen’ at the point of training, and can only create material relevant to their training time.
For instance, one of the most influential image datasets of all time, ImageNet, was compiled around 2007 (though the project was released in 2009), and therefore knows very little about non-‘antique’ iPhones, or other innovations, celebrities or events that have risen to prominence since.



Searching ImageNet for 'iPhone' reveals only the earliest model/s, with most search results featuring iPods or just plain analogue telephones. Source: https://dashboard.scale.com/nucleus/ds_bwk18y3e696g05rk8gyg?similar_text_query=iphone&explorer_display=image&display=natural_language_search&nlp_query=iphone

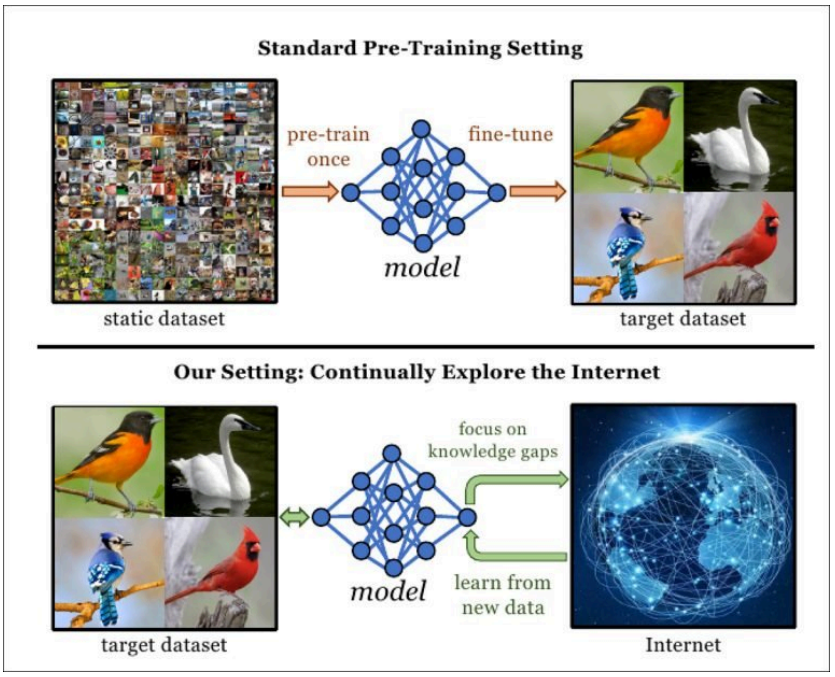
Considering the huge influence that ImageNet continues to have on the computer vision research sector, shortcomings like these are a notable liability, and the inevitable result of using historical data in the context of a modern resource – similar to expecting a 1978 set of Encyclopedias to comprehensively and usefully cover the modern computing and AI scenes.
There is very little one can do to remedy this: fine-tuning an existing model is more like painting the building long after the construction crew have departed; though it can adapt the model to include new data, it tends to favor the new data and make the existing weights less effective overall.
Then there is the new generation of *oracles* – AI systems designed to incorporate new data into trained matrices, so that results can be processed without time-consuming retraining or fine-tuning. However, these laggard 'foreign' facts and factors cannot benefit from true neural integration at training time, which produces the deepest and most intrinsic feature extractions and co-relationships between the trained data points.
Also, considering the enormous expense of training a hyperscale model, and the extent to which changes in the data could radically change the performance of a popular model, it is both risky and costly, even for well-heeled companies, to consider regularly training their best products from scratch, just to get the latest wrinkles in culture and data into the system.

Dated Data

Therefore we have resigned ourselves, for the time being, to the fact that powerful trained AI systems can never have up-to-the-minute information about what exists in the world.
To make up for it, multimodal models such as DALL-E 2 and Stable Diffusion train a *lot* of data – hundreds of millions, even billions of text/image pairs, so that even if the resulting system will inevitably decline in currency, it covers as much as possible of the human experience.
As we've noted recently, the prospect of manually curating data at such volumes is almost inconceivable, while the automated solutions available are far from ideal, in terms of managing the frequent duplicates and great number of low-quality annotations in web-scraped data, in gargantuan collections such as LAION (which powers Stable Diffusion).
So, frequently, researchers that need targeted image data must choose between a) developing smaller but well-curated datasets that definitely suit the project, but which have inadequate volume to challenge the benefits of hyperscale systems; b) using existing high-volume datasets, with all their flaws and stale data, but creating relevant sub-sets using mechanisms such as CLIP; or c) attempting to obtain funds to create novel datasets which have suitable images (though this will still lead to datasets that quite quickly become outmoded).

Internet Explorer…Returns?

A better solution, if it were possible, would be to use the current, immediate state of the internet itself as an *ad hoc* image resource, and to download truly apposite and up-to-date images for your target dataset, instead of developing yet another framework to trawl through the vast expanses of older data in 'historical' datasets (though this is tempting, since, the hard work of obtaining the text/image pairs has been done by someone else).
This, in fact, is what has now been proposed by researchers from Carnegie Mellon and UC Berkeley. The new system, which uses elaborate Natural Language Processing (NLP) and broader data preprocessing techniques to identify and select images for custom datasets, is called *Internet Explorer* (though it remains to be seen if this title can usurp the outmoded but famous Microsoft web browser from which it has borrowed its name).



Internet Explorer is designed to find the right images for the right dataset – right now! Source: https://arxiv.org/pdf/2302.14051.pdf

The authors of the new work explain the concept:

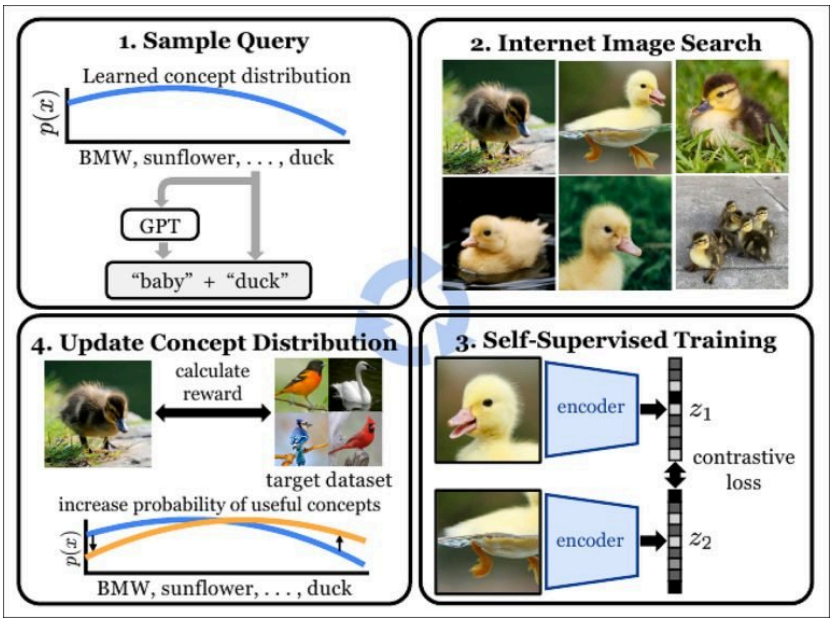
'[We] rethink the idea of a generic large-scale pretrained model and propose an alternate paradigm of training a rather small-scale but up-to-date model geared towards the specific downstream task of interest.
'To train such a model, we go beyond static datasets and treat the Internet itself as a dynamic, open-ended dataset. Unlike conventional datasets, which are expensive to increase and grow stale with time, the Internet is dynamic, rich, grows automatically, and is always up to date. Its continuously evolving nature also means we cannot hope to ever download it or train a model, whether large or small, on all of it.
'We propose that the Internet can be treated as a special kind of dataset—one that exists out there, ready to be queried as needed to quickly train a customized model for a desired task.'

Frequent Dataset Refreshes = Less 'Frozen' Oracles

The central idea behind the new paper appears to be to abandon the tendency to 'immortalize' monumental datasets, which are increasingly under-curated, and whose value will quickly decrease with time, in favor of a series of refreshed and more up-to-date datasets which feature better curation and more current data, and which are therefore feasible to train 'from zero' regularly, as new data comes in.
The risk, of course, is that V2.0 will perform differently from V1.0 – but if the data has altered since V1.0 (i.e., new data has been found), it could hardly be otherwise. One cannot, arguably, have long-term performance consistency and currency of data, since the new data will form new intra-relationships, and offer new features.

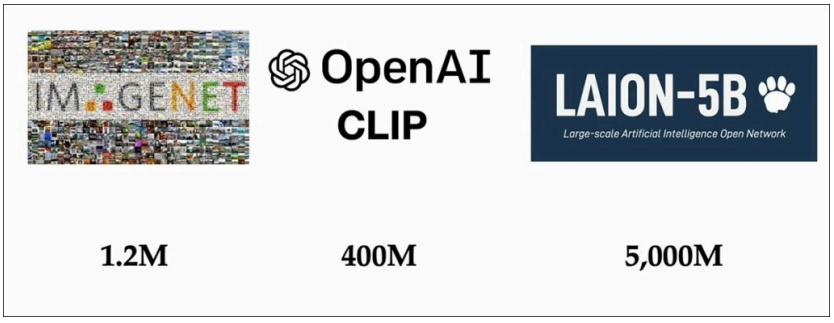
Bespoke Data Selections for New Image Sets

The advantage of this approach hinges on the ability of the system to identify images on the internet which conform to the *aims* of the dataset, instead of the 'shovel-and-sift' approach of trying to download vast acres of data from which the right images will later, somehow, be found.
The kind of (for instance) generative image system that could result from such an approach might be more specialized than the likes of DALL-E 2 and Stable Diffusion, but might well do its specific job better; and since the smaller but higher-quality and better-curated data will take less time to train, frequent 'from scratch' training and updates become a viable prospect, as does the creation of task-oriented bespoke generative systems.



The conceptual workflow for Internet Explorer.

The paper notes that the traditional method of obtaining new datasets is to fine-tune an existing large model, such as ImageNet, using transfer learning via other pre-existing datasets such as AlexNet, ResNet or CLIP.
However, the incredible and ever-growing scale of datasets in computer vision is beginning to make this traditional approach unworkable, because the resources needed to train and/or to pre-process such sets is becoming inaccessible to lesser-funded research entities.

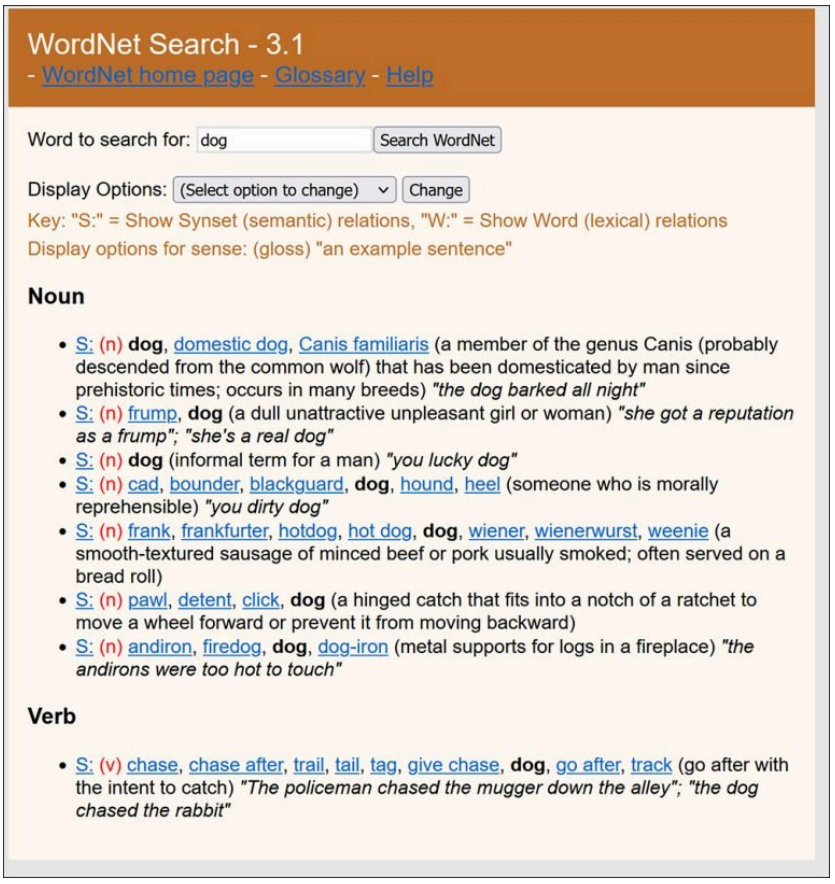


From the accompanying video for Internet Explorer, an indication of the way that hyperscale datasets are beginning to become unusable for all but the largest research groups. Numbers cited represent the approximate number of images in the sets. Source: https://internet-explorer-ssl.github.io/static/videos/internet_explorer.mp4

The idea of paying more attention to the *quality* of the data, rather than relying on volume to provide features at scale, is one that has been advocated over the last few years by influencers such as Andrew Ng.
However, methods for identifying better-quality data (and producing more productive labels and annotations) are difficult to develop. The NLP-based methodology of Internet Explorer therefore lies at the core of the system, offering a way of finding the most useful images from the entirety of the current state of the internet, and bringing them into a highly performant local dataset.

Approach

Internet Explorer (IE) operates by querying search engines that can return image-based results, and also operates on domain-specific online and 'live' resources such as Flickr. Additionally, its methods can be applied as necessary to static and historical datasets such as LAION.
Though it is possible to use images as search queries, in search engines such as Google Search and Yandex, such data is too specific for the purposes of dataset curation, which instead is looking for images that embody *concepts*.
Therefore the researchers for the new project were constrained to search the internet using only text, and to this end divided their core language tools into *concepts* (which outline semantic categories such as *'person'*, *'animals'*, etc.) and *descriptors* (which can modify appearance).
The concepts for IE were taken from the 1995 WordNet hierarchy.



Princeton's WordNet project fuels the initial web searches for IE. Source: http://wordnetweb.princeton.edu/perl/webwn

For the descriptors, the authors used Mesh Transformer JAX, a GPT-J language model, with examples of descriptor/concept pairs.

PROMPT TEMPLATE

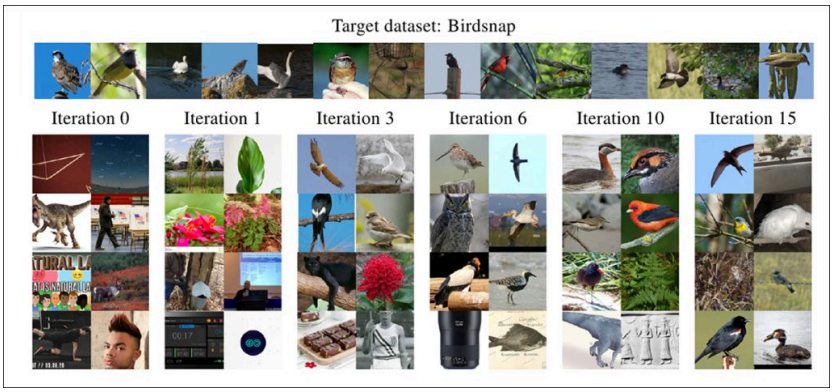
```
"What are some words that describe the quality of '{concept}'?

The {concept} is frail.

The {concept} is red.

The {concept} is humongous.

The {concept} is tall.

The {concept} is"
```

EXAMPLES

```
"a good-looking dog", "very gentle", "a", "brown", "lovable", "a
strong runner", "a male or a female", "sturdy", "agile", "a strong",
"beautiful", "a male", "kind", "long-haired", "a male or a female", "a
good-looking dog", "gentle", "medium", "loyal", "very gentle", "blue-eyed",
"sturdy", "blue-eyed", "a retriever", "kind", "loyal", "large", "brown",
"good-natured", "gentle", "large", "small".
```

The prompt template and example outputs for the autoregressive Mesh Transformer JAX.

The top 100 images for each selected query are downloaded from Google Images using the Google Images Download framework.
IE uses self-supervised learning (SSL) to obtain actionable representations from the downloaded images, and in this regard is compatible with any SSL algorithm that uses image/text pairs. However, the researchers chose Facebook Research's MoCo-V3, for 'speed and stability reasons'.
Prior to looking for internet-based images, a ResNet-50 model is initialized on a MoCo-V3 checkpoint which has been trained for 100 epochs on ImageNet, and also fine-tuned on the target dataset. As the process iterates repeatedly, MoCo-V3 is fine-tuned again and again on the latest downloaded images vs. the target dataset images. The downloaded images are therefore ranked many times, in terms of their consistency in representation space with the target dataset images.



The process, coarse to fine, of iterating through downloaded images until the latest set is most apposite for the target dataset.

Since the initial results are very noisy (i.e., they contain many irrelevant or unhelpful images), at the end of each iteration the system saves the top 50% to a 'replay buffer', effectively setting them to one side for later rounds.

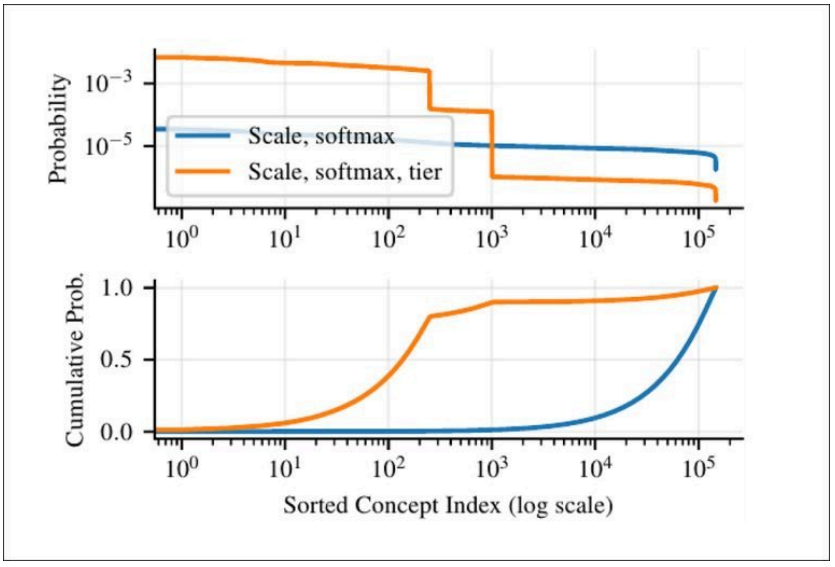
Estimating the extent to which related concepts may be helpful is a separate challenge for IE. For example, the classes 'person' or 'athlete' may be useful adjuncts to the class 'man' in Stable Diffusion's exploitation of the LAION database, since there are actions and concepts embodied in these classes that can be usefully transplanted to prompts which feature 'man'. Therefore IE averages out the top 10 image-level rewards from the results and utilizes this as a *concept score*.

The authors observe*:

'Since our vocabulary contains hundreds of thousands of concepts, it is inefficient to search to test whether a query yields relevant images. Luckily, we can estimate the quality of a query by using the observed rewards of the queries used so far. Humans can do this effortlessly due to our understanding of what each concept means.

'To us, it is obvious that if querying "golden retriever" yielded useful images for this dataset, then "labrador retriever" probably should as well.

'To give our method the same understanding of concept meaning, we embed our 146,347 WordNet concepts into a 384-dimensional space using a pre-trained sentence similarity model…'



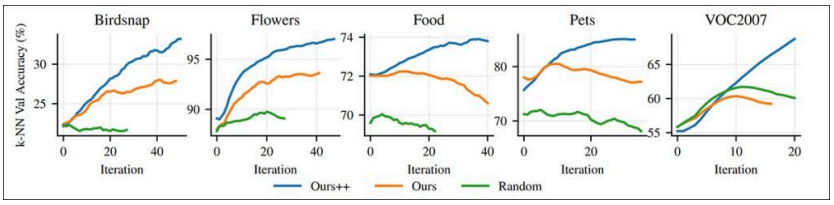
With estimated scores for each of the 146,347 concepts used in the projects, selections must be made to narrow down the space, and create the right balance between exploration (which otherwise would go on forever) and exploitation. In the lower part of the image above, the top 1000 concepts are only marginally sampled.

Tests

Internet Explorer is novel enough that there are no frameworks adequately similar against which tests could be conducted. The 2010 framework NELL extracts web-page text to infer 'candidate beliefs', while the 2013 follow-on project NEIL makes use of images obtained from Google Image Search to learn visual concepts. Neither of these methods modify their explorative behavior as IE does, and neither has the same very specific target objective – of auto-curating a current dataset designed for medium-term use.

Therefore the researchers instead tested three diverse methods within their own framework: sampling concepts randomly; sampling concepts from a learned distribution (IE-style); and sampling concepts from a learned distribution with the addition of GPT-generated descriptors.

IE was tested on four low-small scale fine-grained classification datasets: Birdsnap, Flowers-102, Food101, and Oxford-IIT Pets. Additionally the system was tested against Pascal visual object classes (VOC) 2007, a broad multi-label classification task. Large-scale datasets such as ImageNet were not tested, since their central philosophy and size does not align with the aims of Internet Explorer.



Results from the testing round.

Of these results, the authors comment:

'[Our] method significantly improves on the starting MoCo-v3 (ImageNet + target) checkpoint and can outperform a [CLIP] model of the same size while using much less compute and data. This is impressive as CLIP can be thought of as an oracle, since its training set contains up to 20k Bing image search results for each WordNet lemma (in addition to other queries). Using GPT-generated descriptors in "Ours++" also significantly improves performance by enabling Internet Explorer to generate diverse views of the most useful concepts.'

Further tests on linear probe accuracy (a facet of CLIP) also confirm the superior performance of Internet Explorer:

Model	Birdsnap	Flowers	Food	Pets	VOC2007	Images	GPU-hours
Fixed dataset, self-supervised							
MoCo-v3 (ImageNet + target)	39.9	94.6	78.3	85.3	58.0†	1.2M	84
Exploring the Internet							
Random exploration	39.6 (−0.3)	95.3 (+0.7)	77.0 (−1.3)	85.6 (+0.3)	70.2 (+12.2)	2.2M	124
Search labels only	47.1 (+7.2)	96.3 (+1.7)	80.9 (+2.6)	85.7 (+0.4)	61.8 (+3.8)	2.2M	124

Results for linear probe accuracy, a CLIP metric.

The researchers conclude:

'We show that interactively exploring the Internet is an efficient source of highly relevant training data—if one knows how to search for it. In just 30-40 hours of training on a single GPU, Internet Explorer either significantly outperforms or closely matches the performance of compute-heavy oracle models like [CLIP] trained on static datasets, as well as strong baselines that search the Internet in an undirected manner.'

Conclusion

The ultimate hope for real-time inference in machine learning systems is that new architectures could eventually be developed wherein new data could have a 'knock on' (but non-destructive) effect to the already-formed weights and relationships established at great length and cost during the training process. At the moment, this is science-fiction.

While image synthesis has developed a number of methods for modifying the inference effects from an existing 'frozen' model, or for fine-tuning existing models without the need for full retraining (textual inversion and LORA being just two examples, for Stable Diffusion), none of these 'satellite' solutions operate quite as well as data that has been comprehensively and thoroughly iterated through in a full-fledged training process, while fine-tuning impairs the original overall effectiveness of a large-scale model.

Therefore, a method, such as IE, that can make training models more of an 'everyday' rather than annual or bi-annual affair offer one possible solution to allow models to more accurately reflect current data. It remains to be seen whether IE's authors will fully release the code for their work, as promised in the new paper.

** My conversion of the authors' inline citations to hyperlinks.*