PRE-TRAINING CONCEPT FREQUENCY IS PREDICTIVE OF CLIP ZERO-SHOT PERFORMANCE

Vishaal Udandarao[1,2,*], **Ameya Prabhu**[2,3,*], **Philip H.S. Torr**[3], **Adel Bibi**[3], **Samuel Albanie**[1,†], **Matthias Bethge**[2,†]
[1]University of Cambridge, [2]University of Tübingen, Tübingen AI Center, [3]University of Oxford

ABSTRACT

Web-crawled pre-training datasets are speculated to be key drivers of zero-shot generalization abilities of Vision-Language Models (VLMs) like CLIP, across a range of downstream classification and retrieval tasks, spanning diverse visual concepts. However, it is unclear how meaningful the term "zero-shot" generalization is for CLIP, as its pre-training datasets (*e.g.*, YFCC-15M, LAION-2B etc.) likely contain many samples of the "zero-shot" concept. To study this, for the first time, we analyze the composition of concepts in the pre-training datasets of CLIP. We robustly demonstrate that far from being "zero-shot", CLIP's zero-shot classification performance is strongly predictable by the frequency of a concept seen during pre-training. Precisely, the downstream zero-shot performance improves linearly as the pre-training concept frequency grows exponentially *i.e.*, they follow a log-linear scaling trend. Our data-centric investigation further highlights two key findings: (1) The extreme "data-hunger" of CLIP, *i.e.*, growing inability of "zero-shot" prediction on long-tailed concepts, and (2) A surprising degree of mis-alignment across image-text pairs in the pre-training datasets.

1 INTRODUCTION

Vision-Language Models (VLMs) like CLIP (Radford et al., 2021) are qualitatively different from the large-scale pre-trained models of the past era of computer vision (*e.g.*, ImageNet-21K-trained BiT (Kolesnikov et al., 2020) and BeIT (Bao et al., 2021))—CLIP is now the de-facto standard for downstream tasks like image recognition (Zhai et al., 2023; Li et al., 2021; Yang et al., 2022; Goel et al., 2022; Zhai et al., 2022) and image-text retrieval (Gadre et al., 2023; Kim et al., 2021; Castro & Heilbron, 2022; Udandarao et al., 2020; Yu et al., 2022). This is attributed to the robust "zero-shot" generalisation of CLIP to a wide variety of downstream tasks containing diverse visual concepts (Udandarao et al., 2023; Zhang et al., 2022; 2021; Zhou et al., 2022; Prabhu et al., 2023).

What properties underlie this remarkable concept generalisation of CLIP? A major differentiating factor for CLIP is its pre-training on vast web-crawled datasets encompassing several diverse visual concepts (Schuhmann et al., 2022; Byeon et al., 2022). Past work suggests that this mammoth data-scale is a key driver of such generalisation (Nguyen et al., 2022; Mayilvahanan et al., 2023; Fang et al., 2022; 2023; Nguyen et al., 2023). Nevertheless, it remains unclear how the different properties of the pre-training data distribution affect the downstream performance of CLIP models.

In this work, we deconstruct 3 popular image-text pre-training datasets (CC-3M (Sharma et al., 2018), CC-12M (Changpinyo et al., 2021), and YFCC-15M (Thomee et al., 2016)), to better understand their underlying composition—we first showcase that their constituent concept distribution is extremely long-tailed, comprising several rare concepts. To quantify the impact of this concept distribution on CLIP's performance, we perform a correlation study between the frequency of concepts seen in the pre-training datasets and CLIP's zero-shot accuracy, across 17 downstream datasets. Our findings reveal that CLIP's performance scales linearly as the concept frequency grows exponentially *i.e.*, they follow a log-linear scaling trend. Our analysis further uncovers a high degree of *concept mis-alignment* in pre-training datasets—several paired images and texts do not capture the same concepts. Taken together, our findings point out key issues with current web-scale image-

*Equal contribution. Correspondence to `vu214@cam.ac.uk` and `ameya.prabhu@bethgelab.org`
†Joint senior authors.

Table 1: **Summary of pre-training and downstream datasets, and models used in experiments.**

Category	Dataset/Model
Pre-training Datasets	CC-3M (Sharma et al., 2018) CC-12M (Changpinyo et al., 2021) YFCC-15M (Thomee et al., 2016)
Downstream Datasets	ImageNet (Deng et al., 2009) StanfordCars (Krause et al., 2013) CUB (Wah et al., 2011) UCF101 (Soomro et al., 2012) Caltech101 (Fei-Fei et al., 2004) SUN397 (Xiao et al., 2010) Caltech256 (Griffin et al., 2007) Flowers102 (Nilsback & Zisserman, 2008) DTD (Cimpoi et al., 2014) OxfordPets (Parkhi et al., 2012) Food101 (Bossard et al., 2014) EuroSAT (Helber et al., 2019) FGVCAircraft (Maji et al., 2013) Birdsnap (Berg et al., 2014) Country211 (Radford et al., 2021) CIFAR-10, CIFAR100 (Krizhevsky et al., 2009)
Models	ResNet50 (He et al., 2016) ResNet101 (He et al., 2016) ViT-B-16 (Dosovitskiy et al., 2020)

text datasets, and show how these issues percolate into current CLIP training strategies, leading to inherent challenges in learning rare, long-tailed concepts and thereby hampering true generalisation.

2 Concepts in Pre-training Data and Quantifying Frequency

We first describe our notion of "concepts", followed by our procedure for obtaining concept frequencies from both images and text captions of pre-training datasets, and how we combine them to get image-text matched frequencies.

Defining Concepts. To analyze concept frequencies within pre-training datasets, we first establish the specific concepts we aim to examine. In the context of zero-shot classification, we define concepts as the class names of the classification datasets. For example, for ImageNet, concepts are the set of 1000 classes *e.g.*, "tench", "goldfish", "stingray" etc.

Concept Frequency from Text Captions. To efficiently run text searches for each concept, we pre-cache and index all captions of the pre-training datasets. Our pre-caching first involves lemmatization to normalize word forms in captions (Koskenniemi, 1984). Subsequently, we perform part-of-speech tagging using Spacy (Honnibal & Montani, 2017) to extract common and proper nouns. These extracted nouns are broken down into unigrams and cached in inverted unigram dictionaries. In these dictionaries, each unique noun is linked to a list of sample indices, where each index corresponds to a pre-training sample containing the caption with that particular unigram. This allows $\mathcal{O}(1)$ searching for all sample indices of a given unigram across the pre-training dataset. To determine the frequency of each concept, we decompose it into individual unigrams and query these within our text dictionary. For each unigram, we first get the list of sample indices from our pre-cached dictionary and then perform an intersection of these lists to identify samples containing all unigrams. The samples in the intersection give us the frequency of the concept in the text captions.

Concept Frequency from Images. For images, we do not pre-cache concepts but search them on-the-fly. For efficient searching, we rely on a pre-trained open-set image tagging model, RAM++ (Huang et al., 2023). We collate the set of all concepts present in our downstream evaluation datasets and pass them in as possible outputs from the RAM++ model. The model then tags each image in our pre-training dataset with all relevant concepts from our concept list in a multi-label fashion. This is compiled into a list of all pre-training images that have been tagged with each specific concept, allowing calculation of concept frequencies from the pre-training image set.

Image-Text Matched Concept Frequencies. Having obtained frequencies using both text-based search and image-based search, we compute an image-text matched frequency by identifying pre-training samples where both the image and its caption align with the query concept—we intersect the lists obtained from our image and text searches and determine the sample count in the intersection.

3 Correlation Between Pre-training Frequency and Zero-Shot Performance Across Concepts

In this section, we examine how the frequencies of concepts in pre-training datasets (computed using methods from Sec. 2) influence the zero-shot performance of CLIP models on those same concepts.

Experimental Setup. We analyze the concept frequencies within 3 popular pre-training datasets. In each case, we examine their correlation with zero-shot classification performance on 17 downstream

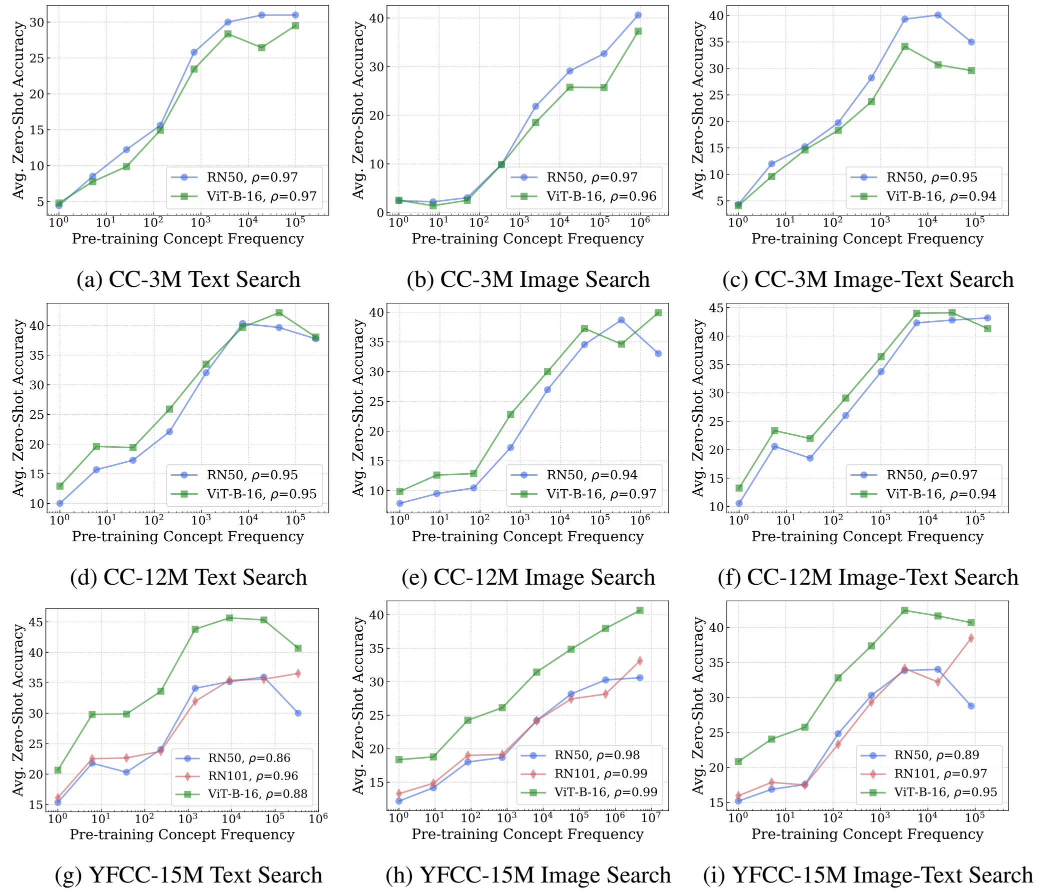


Figure 1: **Log-linear relationships between concept frequency and CLIP zero-shot accuracy.** Across all tested models (RN50, RN101, ViT-B-16) and pre-training datasets (CC-3M, CC-12M, YFCC-15M), we observe the strong linear relationship between CLIP's zero-shot accuracy on a concept and the log of the concept's pre-training frequency.

datasets, across three CLIP models as detailed in Table 1. The downstream datasets cover a range of categories, including objects, scenes, and fine-grained distinctions. We show results with the three methods for computing concept frequencies: text search, image search, and image-text search.

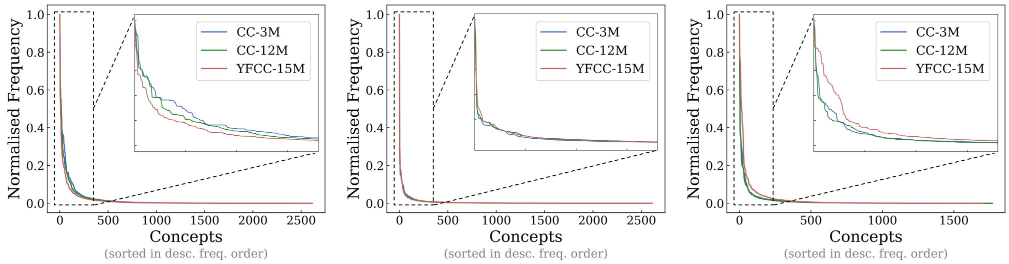
To assess CLIP's performance for each individual concept, we calculate the mean zero-shot classification accuracy for that concept based on its performance within the relevant dataset. For example, the accuracy for the concept "tench" is computed by averaging the zero-shot accuracy scores for all "tench" images within the ImageNet dataset (*i.e.*, across 1000 classes).

Plotting Style. For improved readability, we apply a logarithmic transformation to our concept frequencies. Additionally, we average zero-shot accuracy for all samples using equally spaced bins on the log(concept frequency) axis of our plots (as in Kandpal et al. (2023); Razeghi et al. (2022)).

Main Result: *Log-linear scaling between concept frequency and zero-shot performance.* All our plots from Fig. 1 reveal a consistently strong linear correlation between zero-shot performance and log-scaled pre-training concept frequencies. This points towards the *extreme sample-inefficiency* of current CLIP models in learning concepts from pre-training datasets. Further, this strongly emphasises CLIP's *inability to perform well on long-tailed concepts*, which is a key limitation.

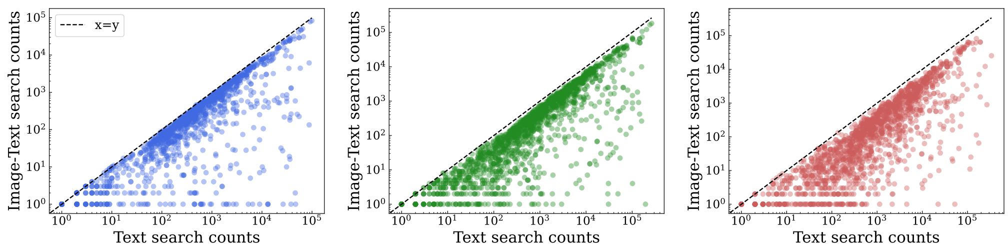
Auxiliary Finding 1: *Long-tailed concept distribution of pre-training datasets*. In Fig. 2, we plot the distribution of pre-training concept frequencies—evidently, it is severely long-tailed. More than two-thirds of the entire concept distribution accounts for almost negligible frequencies when compared to the pre-training dataset sizes. This analysis explains our previous result—VLMs inherit long-tailed biases of their pre-training datasets, and hence are inherently limited on long-tailed data.

Auxiliary Finding 2: *Mis-aligned image-text pairs*. Our analysis additionally allows us to determine *concept mis-alignment* in pre-training datasets *i.e.*, *how many of the paired pre-training images*

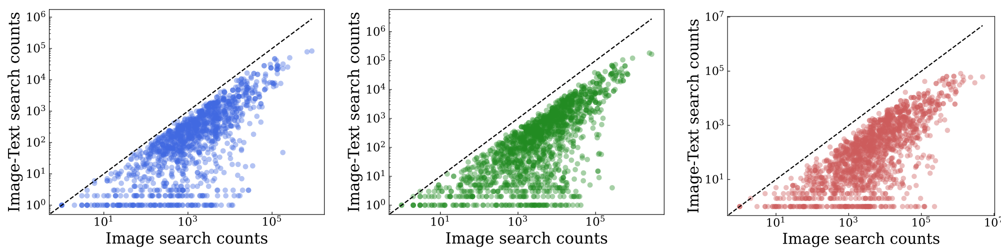


(a) Text search counts (b) Image search counts (c) Image-text search counts

Figure 2: **Concept distribution of pre-training datasets is extremely long-tailed.** We showcase the distribution of pre-training frequencies of all concepts aggregated across all our 17 downstream datasets (about 3200 in total). Across all three pre-training datasets, we observe very heavy tails. We normalise the concept frequencies and remove concepts with 0 counts for improved readability.



(a) CC-3M Text Mis-alignment (b) CC-12M Text Mis-alignment (c) YFCC-15M Text Mis-alignment



(d) CC-3M Image Mis-alignment (e) CC-12M Image Mis-alignment (f) YFCC-15M Image Mis-alignment

Figure 3: **High mis-alignment degree between paired images and texts.** Concept frequencies estimated with only text-search (*top row*) and only image-search (*bottom row*) are always over-estimated compared to those estimated with paired image-text search, suggesting a high degree of mis-alignment between image-text pairs present in the pre-training datasets.

and texts are actually paired? (contain the same concepts). In Fig. 3, we plot the pre-training frequencies obtained from image-text search as a function of frequencies obtained from text or image alone. For an ideal image-text dataset that has completely aligned paired image-text samples, all the points in the scatter plots would lie on the $x=y$ line. However, we see that the scatter plots largely deviate from this line, confirming the high *concept mis-alignment* in these pre-training datasets.

4 Conclusion

In this work, we took a deep-dive into the pre-training datasets of CLIP models (specifically, CC-3M, CC-12M and YFCC-15M) to understand their constituent concepts and their composition. Our analysis showcased that all the considered datasets are extremely long-tailed in their concept distributions. We further studied how this long-tailed nature impacts the downstream performance of CLIP models, showcasing that the zero-shot performance of CLIP on a specific concept can be reliably predicted by the frequency of the concept in the pre-training dataset—this relationship scales log-linearly. Lastly, we uncovered a key abnormality of current pre-training datasets—high degree of concept mis-alignment between the paired images and texts. Our experiments suggest that current CLIP models suffer from extreme data-inefficiency, leading to poor performance on long-tailed concepts—hence questioning the validity of the term "zero-shot" in this context.

REFERENCES

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2011–2018, 2014.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

M. Byeon, B. Park, H. Kim, S. Lee, W. Baek, and S. C. Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

Santiago Castro and Fabian Caba Heilbron. Fitclip: Refining large-scale pretrained image-text models for zero-shot video understanding tasks. *arXiv preprint arXiv:2203.13371*, 2022.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). *arXiv preprint arXiv:2205.01397*, 2022.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks, 2023.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Open-set image tagging with multi-grained text supervision. *arXiv e-prints*, pp. arXiv–2310, 2023.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pp. 15696–15707. PMLR, 2023.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pp. 5583–5594. PMLR, 2021.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 491–507. Springer, 2020.

Kimmo Koskenniemi. A general computational model for word-form recognition and production. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*. The Association for Computational Linguistics, 1984.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. Does clip's generalization performance mainly stem from high train-test similarity? *arXiv preprint arXiv:2310.09562*, 2023.

George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *arXiv preprint arXiv:2208.05516*, 2022.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *arXiv preprint arXiv:2307.10350*, 2023.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Ameya Prabhu, Hasan Abed Al Kader Hammoud, Ser-Nam Lim, Bernard Ghanem, Philip HS Torr, and Adel Bibi. From categories to classifier: Name-only continual learning by exploring the web. *arXiv preprint arXiv:2311.11293*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 840–854, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Vishaal Udandarao, Abhishek Maiti, Deepak Srivatsav, Suryatej Reddy Vyalla, Yifang Yin, and Rajiv Ratn Shah. Cobra: Contrastive bi-modal representation algorithm. *arXiv preprint arXiv:2005.03687*, 2020.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2725–2736, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15671–15680, 2022.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023.

Renrui Zhang, Longtian Qiu, Wei Zhang, and Ziyao Zeng. Vt-clip: Enhancing vision-language models with visual-guided texts. *arXiv preprint arXiv:2112.02399*, 2021.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. *arXiv preprint arXiv:2207.09519*, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

ACKNOWLEDGEMENTS

The authors would like to thank (in alphabetic order): Jonathan Roberts, Karsten Roth, Mehdi Cherti, Prasanna Mayilvahanan, Shyamgopal Karthik and Thao Nguyen for helpful feedback and providing access to various resources throughout the project. AP is funded by Meta AI Grant No. DFR05540. VU thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS). VU also thanks the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program for support. PT thanks the Royal Academy of Engineering for their support. AB acknowledges the Amazon Research Award. SA is supported by a Newton Trust Grant. MB acknowledges financial support via the Open Philanthropy Foundation funded by the Good Ventures Foundation. This work was supported by the German Research Foundation (DFG): SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, TP4, project number: 276693517 and the UKRI grant: Turing AI Fellowship EP/W002981/1. MB is a member of the Machine Learning Cluster of Excellence, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

A NOTE ON PAPER VERSIONS

An extended version of this paper is now available on arxiv. It has two additional authors and several more experimental results.

B LIMITATIONS AND FUTURE DIRECTIONS

Our frequency counting methods currently suffer from several issues, which we hope can be accounted for in the future: (1) *Overly-restrictive exact-text-matching*—as our current method for measuring hits with the text-search relies on exact-matching of the concepts, our measured frequencies can be a gross underestimate of the true frequency, *e.g.*, for the "airplane" concept, we will miss counting all samples that contain the terms "aircraft", "jet", "airliner" etc. One simple solution to mitigate this would be to incoporate synonyms into the search process or performing a semantic search using an embedding model, (2) *Unreliability of image-tagging model*—despite the RAM++ model being a very strong image tagger, it still has some limitations: low image-resolutions, very small object scales, failure in modeling long-tailed concepts etc. Future work can mitigate this using better image tagging models or including ensembles of image taggers and object detectors. Another direction could be to incorporate hierarchies (*e.g.*, Wordnet (Miller, 1998)) in the search process to map long-tailed concepts to more frequent every-day concepts.