OpenAI

Menu

CLIP: Connecting text and images

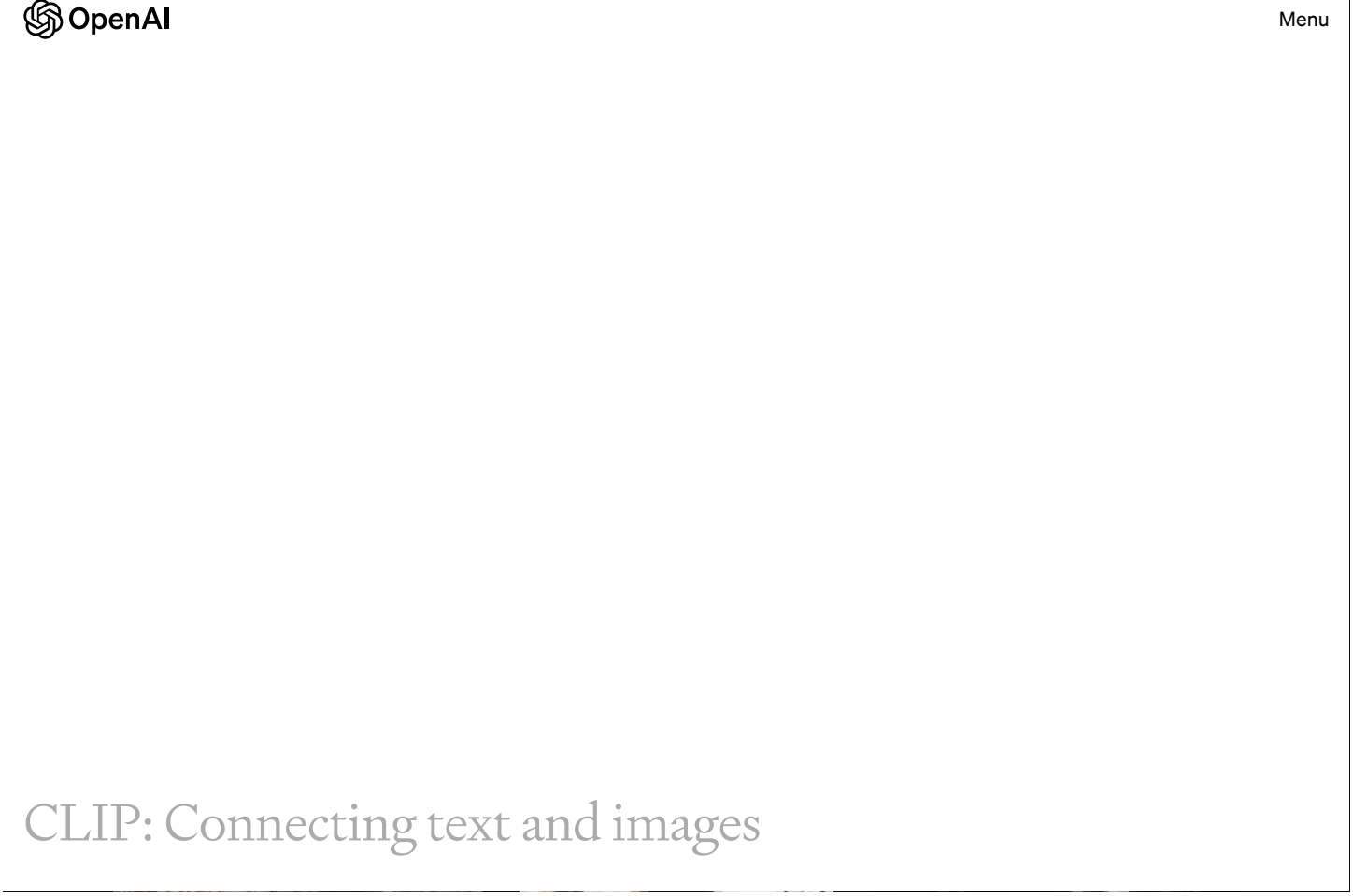


Illustration: Justin Jay Wang

We're introducing a neural network called CLIP which efficiently learns visual concepts from natural language supervision. CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the "zero-shot" capabilities of GPT-2 and GPT-3.

January 5, 2021

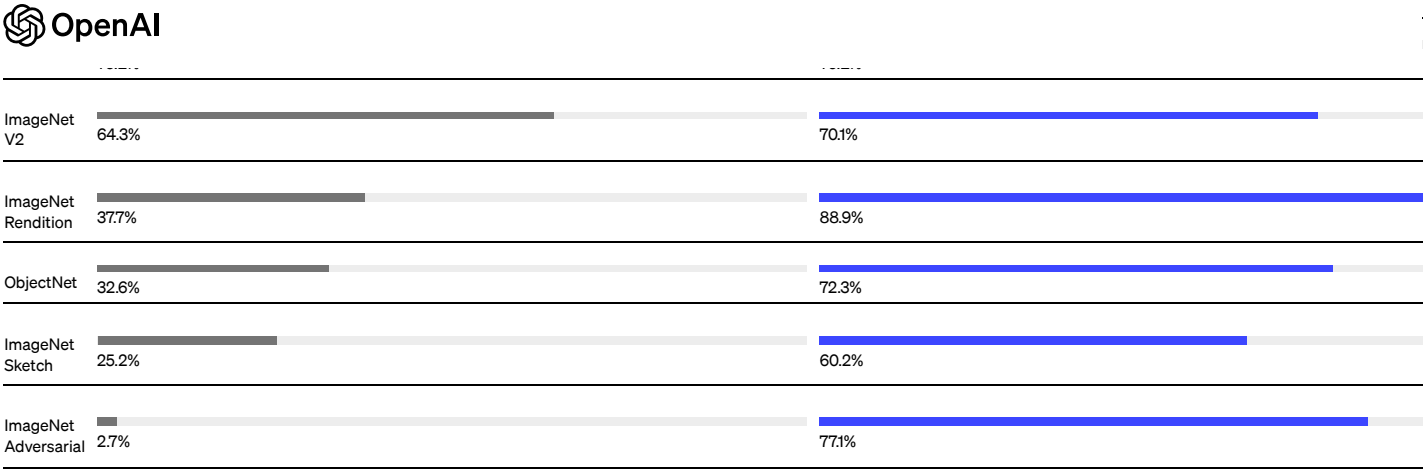
Read paper

View code

Computer vision, Representation learning, Transfer learning, Contrastive learning, Supervised learning, CLIP, Milestone, Publication, Release

Although deep learning has revolutionized computer vision, current approaches have several major problems: typical vision datasets are labor intensive and costly to create while teaching only a narrow set of visual concepts; standard vision models are good at one task and one task only, and require significant effort to adapt to a new task; and models that perform well on benchmarks have disappointingly poor performance on stress tests,[1, 2, 3, 4] casting doubt on the entire deep learning approach to computer vision.

We present a neural network that aims to address these problems: it is trained on a wide variety of images with a wide variety of natural language supervision that's abundantly available on the internet. By design, the network can be instructed in natural language to perform a great variety of classification benchmarks, without directly optimizing for the benchmark's performance, similar to the "zero-shot" capabilities of GPT-2[5] and GPT-3.[6] This is a key change: by not directly optimizing for the benchmark, we show that it becomes much more representative: our system closes this "robustness gap" by up to 75% while matching the performance of the original ResNet-50[7] on ImageNet zero-shot without using any of the original 1.28M labeled examples.



Although both models have the same accuracy on the ImageNet test set, CLIP's performance is much more representative of how it will fare on datasets that measure accuracy in different, non-ImageNet settings. For instance, ObjectNet checks a model's ability to recognize objects in many different poses and with many different backgrounds inside homes while ImageNet Rendition and ImageNet Sketch check a model's ability to recognize more abstract depictions of objects.

Background and related work

CLIP (*Contrastive Language–Image Pre-training*) builds on a large body of work on zero-shot transfer, natural language supervision, and multimodal learning. The idea of zero-data learning dates back over a decade[8] but until recently was mostly studied in computer vision as a way of generalizing to unseen object categories.[9, 10] A critical insight was to leverage natural language as a flexible prediction space to enable generalization and transfer. In 2013, Richer Socher and co-authors at Stanford[11] developed a proof of concept by training a model on CIFAR-10 to make predictions in a word vector embedding space and showed this model could predict two unseen classes. The same year DeVISE[12] scaled this approach and demonstrated that it was possible to fine-tune an ImageNet model so that it could generalize to correctly predicting objects outside the original 1000 training set.

Most inspirational for CLIP is the work of Ang Li and his co-authors at FAIR[13] who in 2016 demonstrated using natural language supervision to enable zero-shot transfer to several existing computer vision classification datasets, such as the canonical ImageNet dataset. They achieved this by fine-tuning an ImageNet CNN to predict a much wider set of visual concepts (visual n-grams) from the text of titles, descriptions, and tags of 30 million Flickr photos and were able to reach 11.5% accuracy on ImageNet zero-shot.

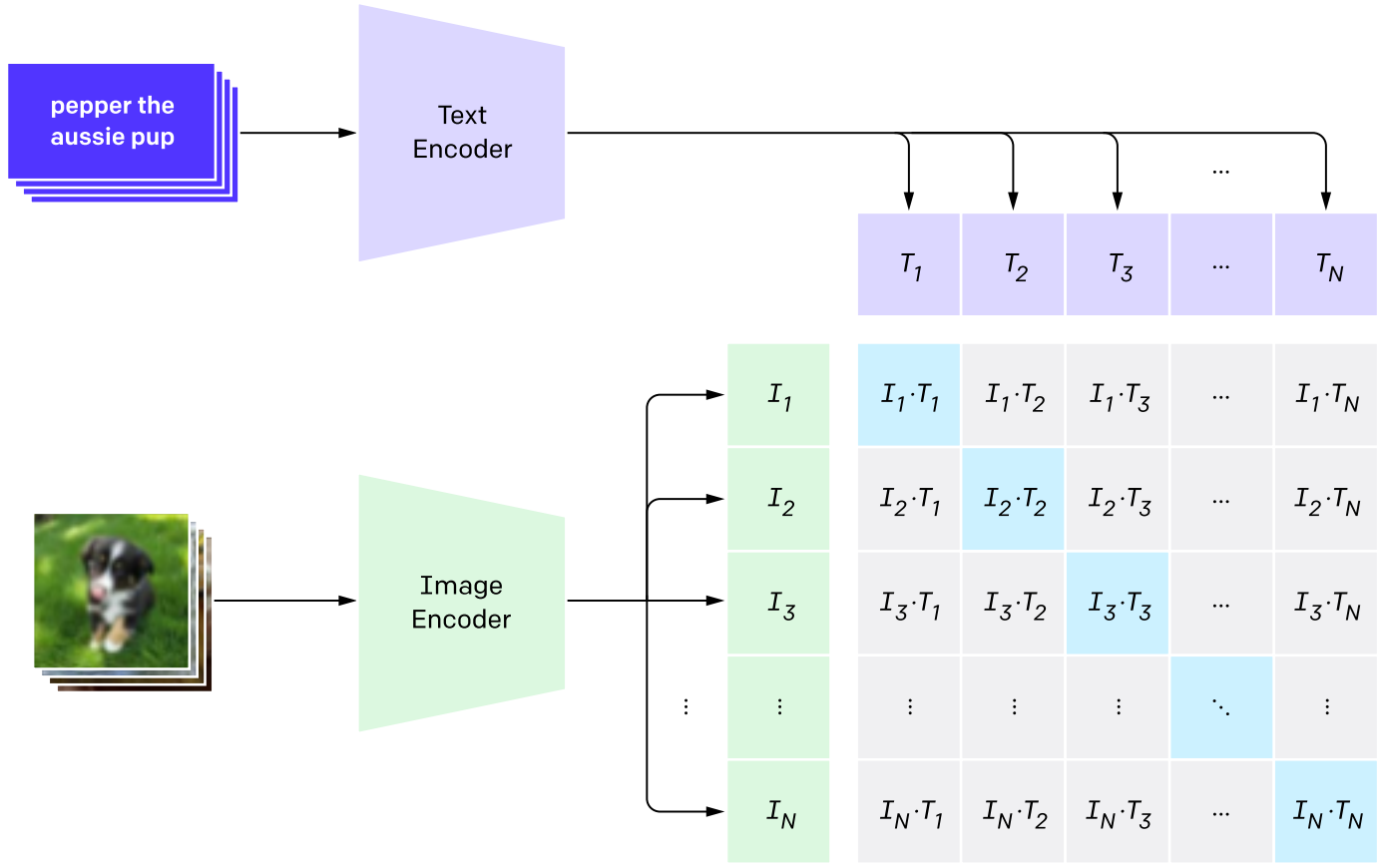
Finally, CLIP is part of a group of papers revisiting learning visual representations from natural language supervision in the past year. This line of work uses more modern architectures like the Transformer[14] and includes VirTex,[15] which explored autoregressive language modeling, ICMLM,[16] which investigated masked language modeling, and ConVIRT,[17] which studied the same contrastive objective we use for CLIP but in the field of medical imaging.

Approach

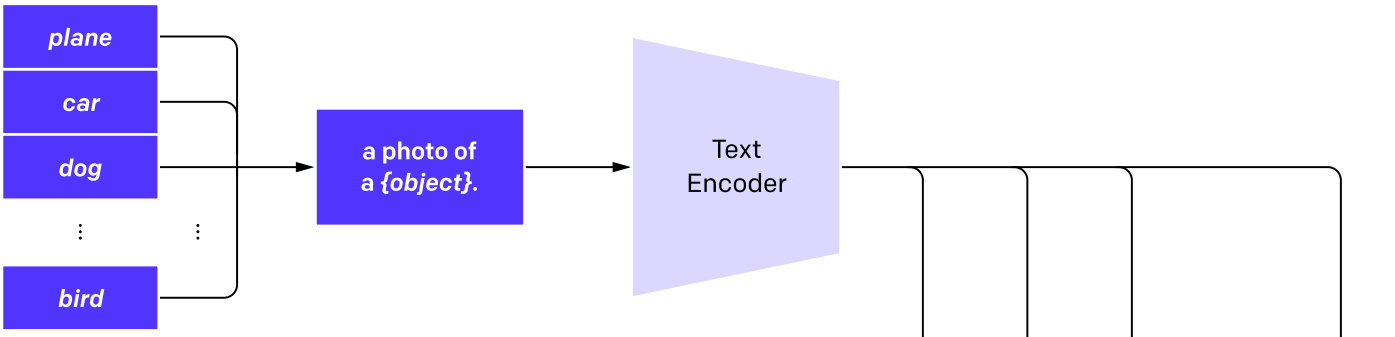
We show that scaling a simple pre-training task is sufficient to achieve competitive zero-shot performance on a great variety of image classification datasets. Our method uses an abundantly available source of supervision: the text paired with images found across the internet. This data is used to create the following proxy training task for CLIP: given an image, predict which out of a set of 32,768 randomly sampled text snippets, was actually paired with it in our dataset.

In order to solve this task, our intuition is that CLIP models will need to learn to recognize a wide variety of visual concepts in images and associate them with their names. As a result, CLIP models can then be applied to nearly arbitrary visual classification tasks. For instance, if the task of a dataset is classifying photos of dogs vs cats we check for each image whether a CLIP model predicts the text description "a photo of a *dog*" or "a photo of a *cat*" is more likely to be paired with it.

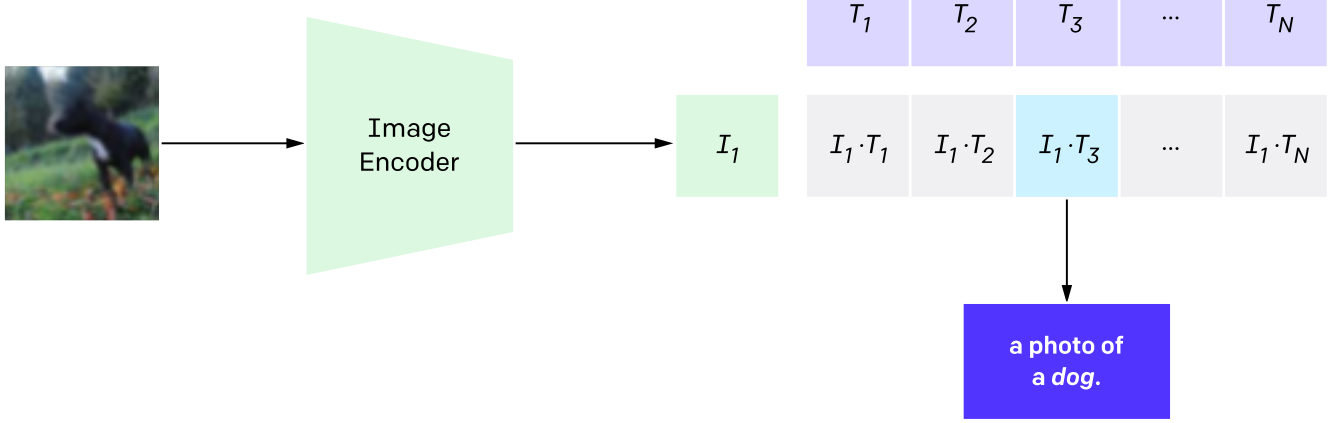
OpenAI



2. Create dataset classifier from label text



3. Use for zero-shot prediction



OpenAI

CLIP was designed to mitigate a number of major problems in the standard deep learning approach to computer vision:

Costly datasets: Deep learning needs a lot of data, and vision models have traditionally been trained on manually labeled datasets that are expensive to construct and only provide supervision for a limited number of predetermined visual concepts. The ImageNet dataset, one of the largest efforts in this space, required over 25,000 workers to annotate 14 million images for 22,000 object categories. In contrast, CLIP learns from text–image pairs that are already publicly available on the internet. Reducing the need for expensive large labeled datasets has been extensively studied by prior work, notably self-supervised learning,[18,19,20] contrastive methods,[21,22,23,24,25] self-training approaches,[26,27] and generative modeling.[28,29,30,31]

Narrow: An ImageNet model is good at predicting the 1000 ImageNet categories, but that's all it can do "out of the box." If we wish to perform any other task, an ML practitioner needs to build a new dataset, add an output head, and fine-tune the model. In contrast, CLIP can be adapted to perform a wide variety of visual classification tasks without needing additional training examples. To apply CLIP to a new task, all we need to do is "tell" CLIP's text-encoder the names of the task's visual concepts, and it will output a linear classifier of CLIP's visual representations. The accuracy of this classifier is often competitive with fully supervised models.

We show random, non-cherry picked, predictions of zero-shot CLIP classifiers on examples from various datasets below.

Food101

guacamole (90.1%)

Ranked 1 out of 101 labels

✓ a photo of **guacamole**, a type of food.

× a photo of **ceviche**, a type of food.

× a photo of **edamame**, a type of food.

× a photo of **tuna tartare**, a type of food.

× a photo of **hummus**, a type of food.

SUN397

television studio (90.2%)

Ranked 1 out of 397 labels

OpenAI

× a photo of a **conference room**.

× a photo of a **lecture room**.

× a photo of a **control room**.

Youtube-BB

airplane, person (89.0%)

Ranked 1 out of 23 labels

✓ a photo of a **airplane**.

× a photo of a **bird**.

× a photo of a **bear**.

× a photo of a **giraffe**.

× a photo of a **car**.

EuroSAT

annual crop land (46.5%)

Ranked 4 out of 10 labels

× a centered satellite photo of **permanent crop land**.

× a centered satellite photo of **pasture land**.

× a centered satellite photo of **highway or road**.

✓ a centered satellite photo of **annual crop land**.

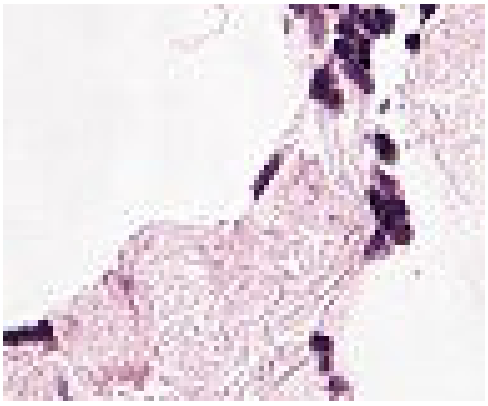
× a centered satellite photo of **brushland or shrubland**.

PatchCamelyon (PCam)

healthy lymph node tissue (77.2%)

Ranked 2 out of 2 labels

OpenAI



ImageNet-A (Adversarial)

lynx (47.9%)

Ranked 5 out of 200 labels

× a photo of a **fox squirrel**.

× a photo of a **mongoose**.

× a photo of a **skunk**.

× a photo of a **red fox**.

✓ a photo of a **lynx**.

Show more

Poor real-world performance: Deep learning systems are often reported to achieve human or even superhuman performance[32,A] on vision benchmarks, yet when deployed in the wild, their performance can be far below the expectation set by the benchmark. In other words, there is a gap between "benchmark performance" and "real performance." We conjecture that this gap occurs because the models "cheat" by only optimizing for performance on the benchmark, much like a student who passed an exam by studying only the questions on past years' exams. In contrast, the CLIP model can be evaluated on benchmarks without having to train on their data, so it can't "cheat" in this manner. This results in its benchmark performance being much more representative of its performance in the wild. To verify the "cheating hypothesis", we also measure how CLIP's performance changes when it is able to "study" for ImageNet. When a linear classifier is fitted on top of CLIP's features, it improves CLIP's accuracy on the ImageNet test set by almost 10%. However, this classifier does *no better* on average across an evaluation suite of 7 other datasets measuring "robust" performance.[34]

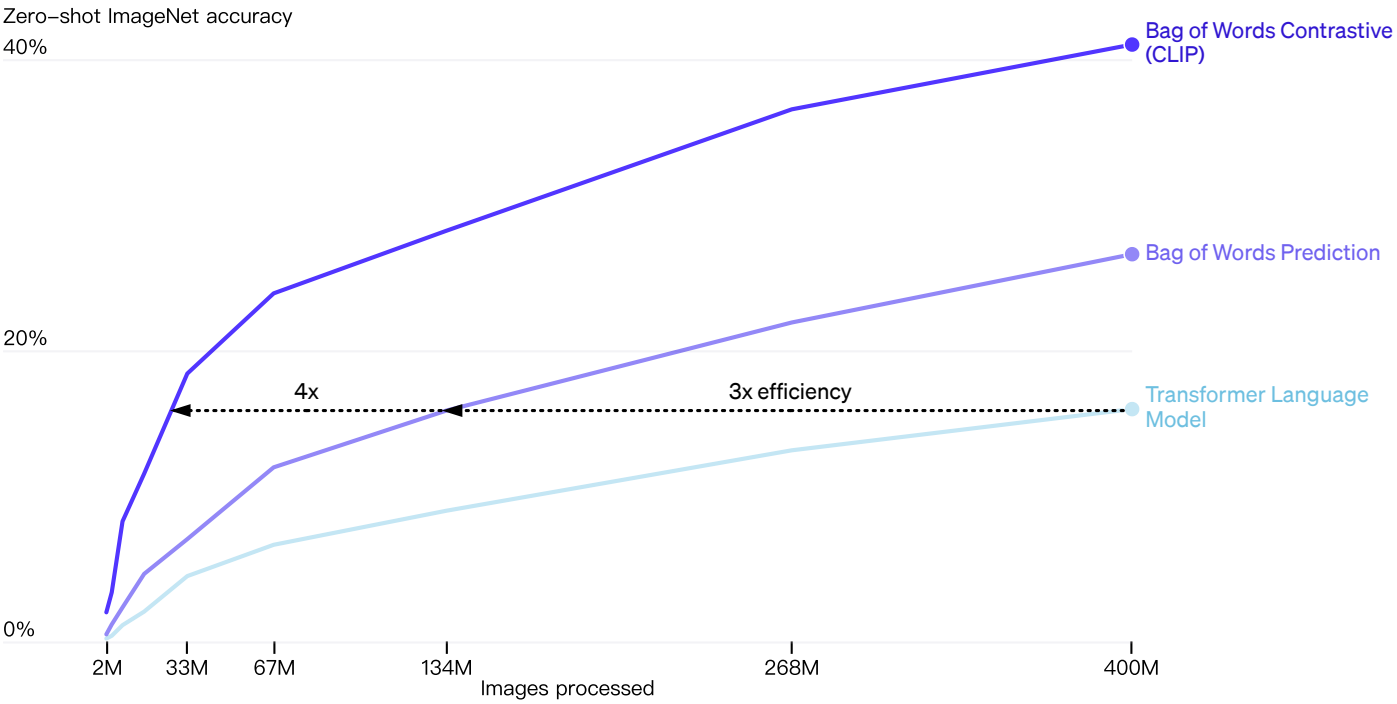
Key takeaways

1. CLIP is highly efficient

CLIP learns from unfiltered, highly varied, and highly noisy data, and is intended to be used in a zero-shot manner. We know from GPT-2 and 3 that models trained on such data can achieve compelling zero shot performance; however, such models require significant training compute. To reduce the needed compute, we focused on algorithmic ways to improve the training efficiency of our approach.

OpenAI

scaling this to achieve state-of-the-art performance. In small to medium scale experiments, we found that the contrastive objective used by CLIP is 4x to 10x more efficient at zero-shot ImageNet classification. The second choice was the adoption of the Vision Transformer,[36] which gave us a further 3x gain in compute efficiency over a standard ResNet. In the end, our best performing CLIP model trains on 256 GPUs for 2 weeks which is similar to existing large scale image models.[37,27,38,36]

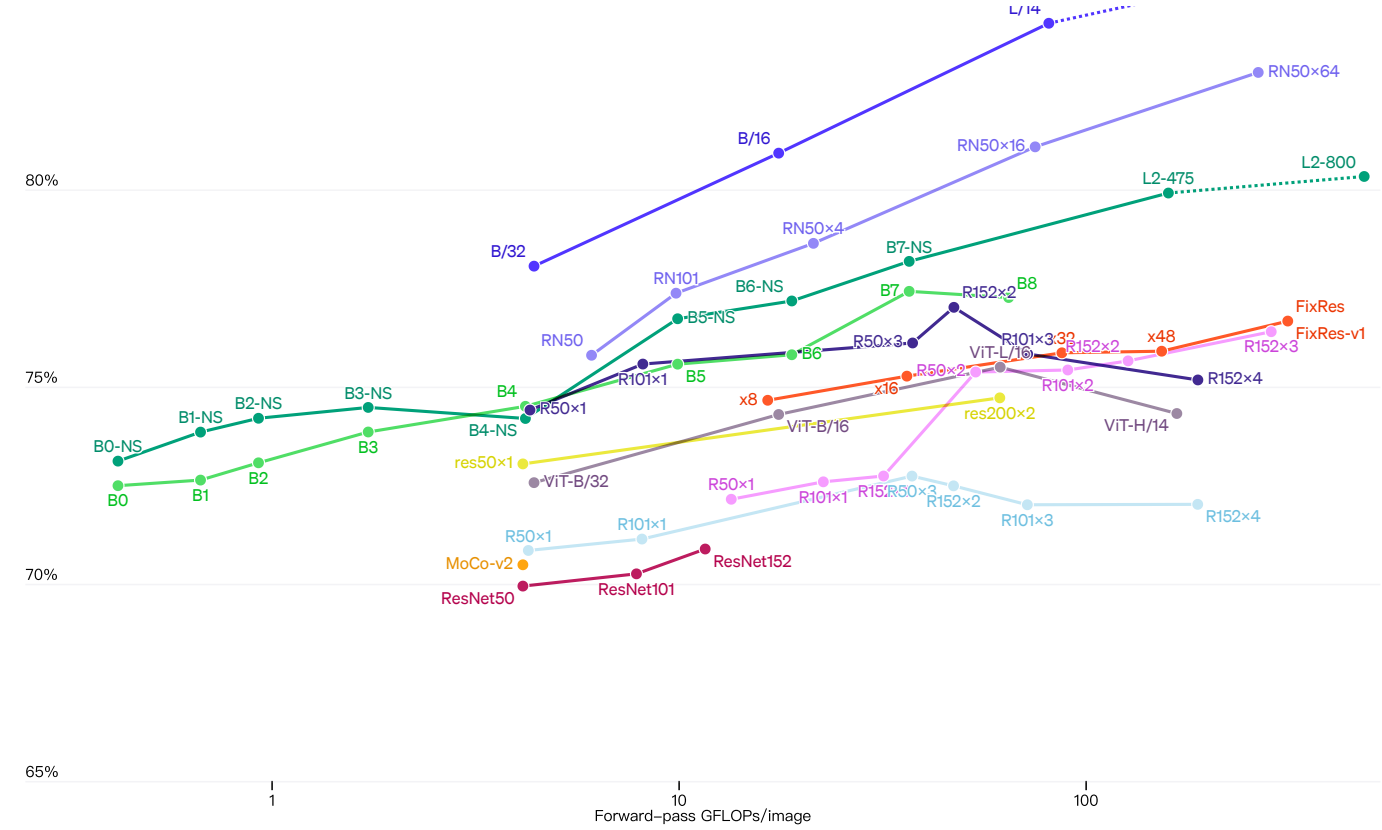


We originally explored training image-to-caption language models but found this approach struggled at zero-shot transfer. In this 16 GPU day experiment, a language model only achieves 16% accuracy on ImageNet after training for 400 million images. CLIP is much more efficient and achieves the same accuracy roughly 10x faster.

2. CLIP is flexible and general

Because they learn a wide range of visual concepts directly from natural language, CLIP models are significantly more flexible and general than existing ImageNet models. We find they are able to zero-shot perform many different tasks. To validate this we have measured CLIP's zero-shot performance on over 30 different datasets including tasks such as fine-grained object classification, geo-localization, action recognition in videos, and OCR.[B] In particular, learning OCR is an example of an exciting behavior that does not occur in standard ImageNet models. Above, we visualize a random non-cherry picked prediction from each zero-shot classifier.

This finding is also reflected on a standard representation learning evaluation using linear probes. The best CLIP model outperforms the best publicly available ImageNet model, the Noisy Student EfficientNet-L2,[27] on 20 out of 26 different transfer datasets we tested.



Across a suite of 27 datasets measuring tasks such as fine-grained object classification, OCR, activity recognition in videos, and geo-localization, we find that CLIP models learn more widely useful image representations. CLIP models are also more compute efficient than the models from 10 prior approaches that we compare with.

Limitations

While CLIP usually performs well on recognizing common objects, it struggles on more abstract or systematic tasks such as counting the number of objects in an image and on more complex tasks such as predicting how close the nearest car is in a photo. On these two datasets, zero-shot CLIP is only slightly better than random guessing. Zero-shot CLIP also struggles compared to task specific models on very fine-grained classification, such as telling the difference between car models, variants of aircraft, or flower species.

CLIP also still has poor generalization to images not covered in its pre-training dataset. For instance, although CLIP learns a capable OCR system, when evaluated on handwritten digits from the MNIST dataset, zero-shot CLIP only achieves 88% accuracy, well below the 99.75% of humans on the dataset. Finally, we've observed that CLIP's zero-shot classifiers can be sensitive to wording or phrasing and sometimes require trial and error "prompt engineering" to perform well.

Broader impacts

CLIP allows people to design their own classifiers and removes the need for task-specific training data. The manner in which these classes are designed can heavily influence both model performance and model biases. For example, we find that when given a set of labels including Fairface[39] race labels[C] and a handful of egregious terms such as "criminal", "animal," etc., the model tends to classify images of people aged 0–20 in the egregious category at a rate of ~32.3%. However, when we add the class "child" to the list of possible classes, this behaviour drops to ~8.7%.

Additionally, given that CLIP does not need task-specific training data it can unlock certain niche tasks with greater ease. Some of these tasks may raise privacy or surveillance related risks and we explore this concern by studying the performance of CLIP on celebrity identification. CLIP has a top-1 accuracy of 59.2% for "in the wild" celebrity image classification when choosing from 100 candidates and a

OpenAI

challenges that CLIP poses in our paper and we hope that this work motivates future research on the characterization of the capabilities, shortcomings, and biases of such models. We are excited to engage with the research community on such questions.

Conclusion

With CLIP, we've tested whether task agnostic pre-training on internet scale natural language, which has powered a recent breakthrough in NLP, can also be leveraged to improve the performance of deep learning for other fields. We are excited by the results we've seen so far applying this approach to computer vision. Like the GPT family, CLIP learns a wide variety of tasks during pre-training which we demonstrate via zero-shot transfer. We are also encouraged by our findings on ImageNet that suggest zero-shot evaluation is a more representative measure of a model's capability.

Footnotes

A In 2015, a group of researchers from Microsoft first trained a model which achieved a top-5 accuracy on ImageNet that surpassed reported human top-5 accuracy.[33] ↩

B While CLIP's zero-shot OCR performance is mixed, its semantic OCR representation is quite useful. When evaluated on the SST-2 NLP dataset rendered as images, a linear classifer on CLIP's representation matches a CBoW model with direct access to the text. CLIP is also competitive at detecting hateful memes without needing ground truth text. ↩

C FairFace is a face image dataset designed to balance age, gender, and race, in order to reduce asymmetries common in previous face datasets. It categorizes gender into 2 groups: female and male and race into 7 groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. There are inherent problems with race and gender classifications, as e.g. Bowker and Star (2000)[40] and Keyes (2018)[41] have shown. While FairFace's dataset reduces the proportion of White faces, it still lacks representation of entire large demographic groups, effectively erasing such categories. We use the 2 gender categories and 7 race categories defined in the FairFace dataset in a number of our experiments not in order to reinforce or endorse the use of such reductive categories, but in order to enable us to make comparisons to prior work. ↩

References

1 Dodge, S., & Karam, L. (2017, July). "A study and comparison of human and deep learning recognition performance under visual distortions." In ICCCN 2017. ↩

2 Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." In ICLR 2019. ↩

3 Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W. S., & Nguyen, A. (2019). "Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects." In CVPR 2019. ↩

4 Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., ... & Katz, B. (2019). "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models." In NeurIPS 2019. ↩

5 Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). "Language Models are Unsupervised Multitask Learners." Technical Report, OpenAI. ↩

6 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). "Language Models are Few-Shot Learners." In NeurIPS 2020. ↩

7 He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep residual learning for image recognition." In CVPR 2016. ↩

8 Larochelle, H., Erhan, D., & Bengio, Y. (2008, July). "Zero-data learning of new tasks." In AAAI 2008. ↩

9 Lampert, C. H., Nickisch, H., & Harmeling, S. (2009, June). "Learning to detect unseen object classes by between-class attribute transfer." In CVPR 2009. ↩

10 Lei Ba, J., Swersky, K., & Fidler, S. (2015). "Predicting deep zero-shot convolutional neural networks using textual descriptions." In ICCV 2015. ↩

11 Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. (2013). "Zero-shot learning through cross-modal transfer." In NeurIPS 2013. ↩

12 Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., & Mikolov, T. (2013). "Devise: A deep visual-semantic embedding model." In NeurIPS 2013. ↩

13 Li, A., Jabri, A., Joulin, A., & van der Maaten, L. (2017). "Learning visual n-grams from web data." In Proceedings of the IEEE International Conference on Computer Vision 2017. ↩

14 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). "Attention is all you need." In NeurIPS 2017. ↩

15 Desai, K., & Johnson, J. (2020). "VirTex: Learning Visual Representations from Textual Annotations." arXiv preprint. ↩ ↩

16 Sariyildiz, M. B., Perez, J., & Larlus, D. (2020). "Learning Visual Representations with Caption Annotations." In ECCV 2020. ↩

18 Doersch, C., Gupta, A., & Efros, A. A. (2015). "Unsupervised visual representation learning by context prediction." In ICCV 2015. ↩

19 Zhai, X., Oliver, A., Kolesnikov, A., & Beyer, L. (2019). "S4l: Self-supervised semi-supervised learning." In ICCV 2019. ↩

20 Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., ... & Piot, B. (2020). "Bootstrap your own latent: A new approach to self-supervised learning." In NeurIPS 2020. ↩

21 Oord, A. V. D., Li, Y., & Vinyals, O. (2018). "Representation Learning with Contrastive Predictive Coding." arXiv preprint. ↩ ↩

22 Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., & Bengio, Y. (2018). "Learning deep representations by mutual information estimation and maximization." In ICLR 2019. ↩

23 Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). "Learning representations by maximizing mutual information across views." In NeurIPS 2019. ↩

24 He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). "Momentum contrast for unsupervised visual representation learning." In CVPR 2020. ↩

25 Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). "A simple framework for contrastive learning of visual representations." arXiv preprint. ↩

26 Lee, D. H. (2013, June). "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks." In Workshop on challenges in representation learning, ICML (2013). ↩

27 Xie, Q., Luong, M. T., Hovy, E., & Le, Q. V. (2020). "Self-training with noisy student improves imagenet classification." In CVPR 2020. ↩ ↩ ↩

28 Kingma, D. P., Mohamed, S., Jimenez Rezende, D., & Welling, M. (2014). "Semi-supervised learning with deep generative models." In NeurIPS 2014. ↩

29 Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). "Improved techniques for training gans." In NeurIPS 2016. ↩

30 Donahue, J., & Simonyan, K. (2019). "Large scale adversarial representation learning." In NeurIPS 2019. ↩

31 Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020, November). "Generative pretraining from pixels." In ICML 2020. ↩

32 He, K., Zhang, X., Ren, S., & Sun, J. (2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." In ICCV 2015. ↩

33 Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). "Imagenet large scale visual recognition challenge." In IJCV 2015. ↩ ↩

34 Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., & Schmidt, L. (2020). "Measuring robustness to natural distribution shifts in image classification." In NeurIPS 2020. ↩

35 Sohn, K. (2016). "Improved deep metric learning with multi-class n-pair loss objective." In NeurIPS 2016. ↩

36 Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Uszkoreit, J. (2020). "An image is worth 16×16 words: Transformers for image recognition at scale." arXiv preprint. ↩ ↩

37 Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., ... & van der Maaten, L. (2018). "Exploring the limits of weakly supervised pretraining." In ECCV 2018. ↩

38 Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2019). "Big Transfer (BiT): General Visual Representation Learning." arXiv preprint. ↩

39 Kärkkäinen, K., & Joo, J. (2019). "Fairface: Face attribute dataset for balanced race, gender, and age." arXiv preprint. ↩

40 Bowker, G., & Star, S. L. (1999). "Sorting things out. Classification and its consequences" Book. ↩ ↩

41 Keyes, O. (2018). "The misgendering machines: Trans/HCI implications of automatic gender recognition." In Proceedings of the ACM on Human-Computer Interaction. ↩ ↩

Authors

Alec Radford

Ilya Sutskever

Jong Wook Kim

Gretchen Krueger

Sandhini Agarwal

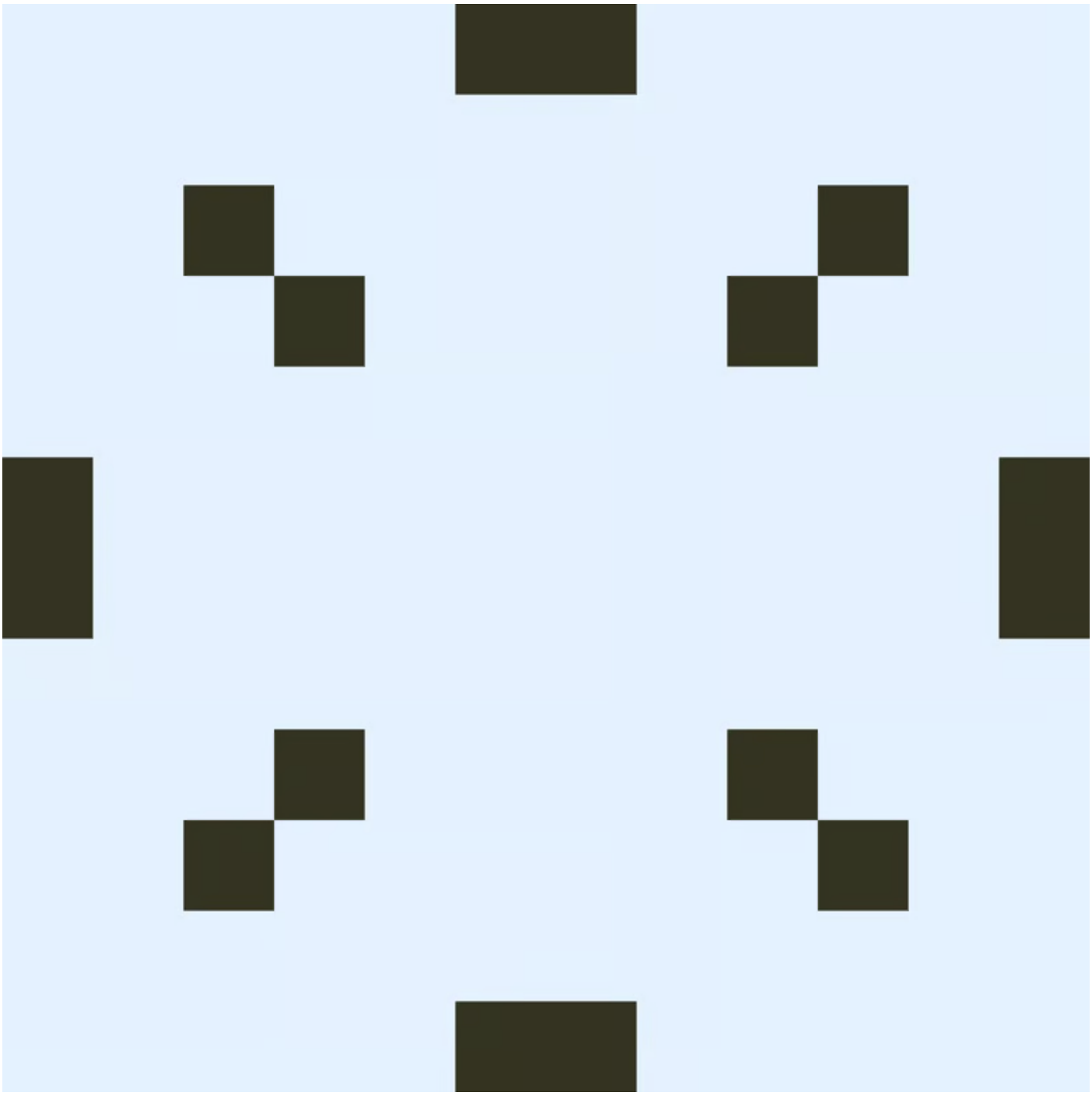
Acknowledgments

We'd like to thank the millions of people involved in creating the data CLIP is trained on. We also are grateful to all our co-authors for their contributions to the project. Finally, we'd like to thank Jeff Clune, Miles Brundage, Ryan Lowe, Jakub Pachocki, and Vedant Misra for feedback on drafts of this blog and Matthew Knight for reviewing the code release.

OpenAI

Related research

View all research



Point-E: A system for generating 3D point clouds from complex prompts
Dec 16, 2022

OpenAI

Learning to play Minecraft with Video PreTraining
Jun 23, 2022

OpenAI

Multimodal neurons in artificial neural networks
Mar 4, 2021

OpenAI

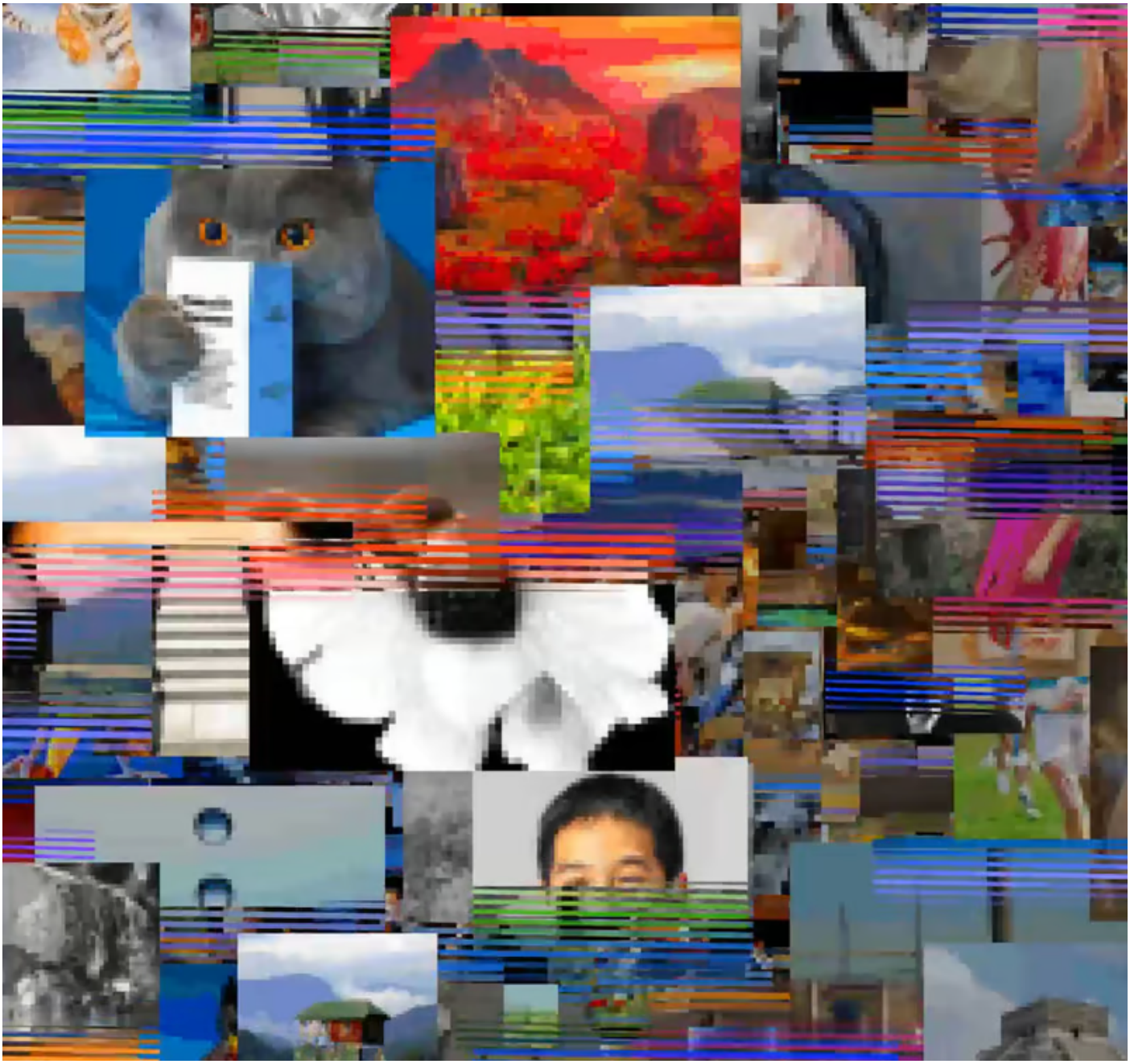


Image GPT
Jun 17, 2020

OpenAI

Careers
Charter
Security

Social
Twitter
YouTube
GitHub
SoundCloud
LinkedIn

Back to top