DETECTING CLOUD PRESENCE IN SATELLITE IMAGES USING THE RGB-BASED CLIP VISION-LANGUAGE MODEL

Mikolaj Czerkawski, Robert Atkinson, Christos Tachtatzis

University of Strathclyde
Glasgow, UK

The text medium has begun to play a prominent role in the processing of visual data over the last years, such as images [1, 2, 3, 4, 5, 6, 7], or videos [8, 9, 10]. The use of language allows human users to easily adapt the computer vision tools to their needs and so far, it has primarily been used for purely creative purposes. Yet, vision-language models could also pave the way for many remote sensing applications that can be defined in a zero-shot manner, without the need for extensive training or any training at all.

At the core of many text-based vision solutions stands CLIP, a vision-language model designed for measuring alignment between text and image inputs [1]. In this work, the capability of the CLIP model to recognize cloud-affected satellite images is investigated. The approach to this is not immediately obvious; the CLIP model operates on RGB images, while a typical solution to detect clouds in satellite imagery involves more than the RGB visible bands, such as infrared, and is often sensor-specific. Some past works have explored the potential of an RGB-only cloud detection model [11], but the task is considered significantly more challenging. Furthermore, the CLIP model has been trained on the general WebImageText dataset [1], so it is not currently obvious how well it could perform with a task as specific as classification of cloud-affected satellite imagery.

In this work, the capability of the official pre-trained CLIP model (ViT-B/32 backbone) is put to test. There are two important insights gained here: it allows to estimate the utility of representations learned by CLIP for cloud-oriented tasks (which can potentially lead to more complex uses such as segmentation or removal), and further, it can act as a tool for filtering datasets based on the presence of clouds.

The CLIP model [1] has been designed for zero-shot classification of images where labels can be supplied (and hence, specified as text) upon inference. The CLIP model consists of separate encoders for text and image input, with jointly learned embedding space. A relative measure of alignment between a given text-image pair can be obtained by computing the cosine similarity between the encodings.

The manuscript explores four variants of using CLIP for cloud presence detection, shown in Table 1, one (fully zero-shot) based on text prompts (1), and (2)-(4) based on minor (1,000 gradient steps with batch size of 10, on only the training dataset) fine-tuning of the high-level classifier module. In the case of (2), a linear probe is attached to the features encoded by the image encoder. In the case of (3), a CoOp approach is employed, as described in [12]. Finally, the Radar (4) approach applies a linear probe classifier to the image encodings of both RGB data and a false-color composite of the SAR Data (VV, VH, and mean of the two channels are encoded as 3 input channels). Furthermore, the learned approaches (2)-(4) are tested for (dataset/sensor) transferability. The (a) variants correspond to the training and testing data coming from the same sensor, while the (b) variants employ transfer. The text prompts for method (1) were arbitrarily selected as *"This is a satellite image with clouds"* and *"This is a satellite image with clear sky"* with no attempt to improve them.

The approach is tested on two benchmark datasets: (i) CloudSEN12 [13], containing Sentinel-2 and Sentinel-1 data and (ii) SPARCS [14], containing Landsat-8 imagery. Both datasets contain examples of cloudy images as well as images with no clouds present, representing the two labels. Testing on two datasets with Sentinel-2 and Landsat-8 data allows to measure the transferability of the proposed methods. Furthermore, the annotators of the SPARCS dataset, while labeling the images, have been shown false-color images with bands B6 (SWIR), B5 (NIR), and B4 (Red) assigned to RGB channels, respectively. While these images are artificial, they might still be interpreted by a CLIP model. Hence, two versions of the SPARCS dataset are tested here, one with the RGB bands and one with the false-color images observed by the annotators.

The results indicate that text-prompts can be used to achieve a non-trivial performance in a purely zero-shot manner, with very high accuracy for the cloudy images (0.929, 0.922, 0.900 on CloudSEN12, SPARCS/RGB, and SPARCS/False-Color, respectively) and lower accuracy for clear images (0.638, 0.737, 0.737 for the same dataset order). Further improvements can be achieved by an inexpensive fine-tuning stage (several minutes on a single consumer-grade GPU), yielding improved

This work was funded by EPSRC under Grant EP/R513349/1.

Table 1. Accuracy of cloud presence detection for the tested datasets and detection methods.

Test Dataset Modality	CloudSEN12 S2/RGB			SPARCS L8/RGB			L8/B6-B4		
	Cloudy	Clear	F1	Cloudy	Clear	F1	Cloudy	Clear	F1
1. Text Prompts	0.929	0.638	0.919	0.922	0.737	0.907	0.900	0.737	0.895
Trained on:	S2/RGB			L8/RGB			L8/B6-B4		
2a. Linear Probe	0.924	0.975	0.957	0.856	1.000	0.922	0.822	1.000	0.902
3a. CoOp	0.936	0.980	0.964	0.878	0.921	0.919	0.822	0.974	0.897
4a. Radar	0.930	0.960	0.959	N/A	N/A	N/A	N/A	N/A	N/A
Trained on:	L8/B6-B4			S2/RGB			S2/RGB		
2b. Linear Probe	0.961	0.759	0.950	0.811	1.000	0.896	0.811	1.000	0.896
3b. CoOp	0.988	0.578	0.943	0.789	1.000	0.882	0.844	0.974	0.910

performance, with high accuracy of around 0.9 and above for both cloudy and clear images, compared to text-based zero-shot variant. Finally, it is shown that the features learned via the linear probe method are most effective for generalizing across datasets and sensing modalities.

The results presented herein demonstrate the potential of harnessing a general vision-language model of CLIP for processing clouds in satellite imagery with minimal training requirements.

1. REFERENCES

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," 2021.

[2] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever, "Zero-shot text-to-image generation," in *Proceedings of the 38th International Conference on Machine Learning*, Marina Meila and Tong Zhang, Eds. 18–24 Jul 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831, PMLR.

[3] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," , no. Figure 3, 2022.

[4] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phuc Le Khac, Luke Melas, and Ritobrata Ghosh, "Dall·e mini," 7 2021.

[5] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," 2022.

[6] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu, "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation," 2022.

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10684–10695.

[8] Hu Xu, Gargi Ghosh, Po Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer, "VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding," *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 6787–6800, 2021.

[9] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman, "Make-A-Video: Text-to-Video Generation without Text-Video Data," pp. 1–13, 2022.

[10] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans, "Imagen Video: High Definition Video Generation with Diffusion Models," pp. 1–18, 2022.

[11] Savas Ozkan, Mehmet Efendioglu, and Caner Demirpolat, "Cloud detection from RGB color remote sensing images with deep pyramid networks," *International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 2018-July, pp. 6939–6942, 2018.

[12] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, "Learning to Prompt for Vision-Language Models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[13] Cesar Aybar, Luis Ysuhuaylas, Karen Gonzales, Jhomira Loja, Fernando Herrera, Lesly Bautista, Angie Flores, Roy Yali, Lissette Diaz, Nicole Cuenca, Fernando Prudencio, David Montero, Martin Sudmanns, Dirk Tiede, Gonzalo Mateo-garc, and G Luis, "CloudSEN12 - a global dataset for semantic understanding of cloud and cloud shadow in satellite imagery," *EarthArXiv*, 2022.

[14] M. Joseph Hughes and Robert Kennedy, "High-quality cloud masking of landsat 8 imagery using convolutional neural networks," *Remote Sensing*, vol. 11, no. 21, 2019.