Towards Automatic Transformer-based Cloud Classification and Segmentation

Roshan Roy[*]
BITS Pilani
rroshanroy@gmail.com

Ahan M R[*]
BITS Pilani Goa Campus
f20160487@goa.bits-pilani.ac.in

Vaibhav Soni
MANIT Bhopal
vaibsoni@gmail.com

Ashish Chittora
BITS Pilani Goa Campus
ashishc@goa.bits-pilani.ac.in

Abstract

Clouds have been demonstrated to have a huge impact on the energy balance, temperature, and weather of the Earth. Classification and segmentation of clouds and coverage factors is crucial for climate modelling, meteorological studies, solar energy industry, and satellite communication. For example, clouds have a tremendous impact on short-term predictions or 'nowcasts' of solar irradiance and can be used to optimize solar power plants and effectively exploit solar energy. However even today, cloud observation requires the intervention of highly-trained professionals to document their findings, which introduces bias. To overcome these issues and contribute to climate change technology, we propose, to the best of our knowledge, the first two transformer-based models applied to cloud data tasks. We use the CCSD Cloud classification dataset and achieve 90.06% accuracy, outperforming all other methods. To demonstrate the robustness of transformers in this domain, we perform Cloud segmentation on SWIMSWG dataset and achieve 83.2% IoU, also outperforming other methods. With this, we signal a potential shift away from pure CNN networks.

1 Introduction

Nearly 60% of the Earth surface is covered by clouds (8) and thus, are crucial in regulating the Earth's radiation budget, weather forecasting, climate studies, water cycle and many other natural phenomena. (5) shows that cloud type variations induce radiative flux changes, which in turn influence changes in climate. Technological intervention into understanding clouds has a tremendous variety of applications to ameliorate climate change.

Accurate cloud type detection and segmentation can be crucial in building optimized solar energy power plants.(11) Hourly temporal fluctuation in shortwave solar irradiance is primarily caused by cloud cover, which is difficult to predict because of its variability (9). Thus, to create optimal solar grids, accurate 'nowcasts' that predict short-term variations is essential. Additionally, this would help reduce the burden on the plant's energy storage capacity as batteries could be used more flexibly (9). Thus, this technology plays a vital part in maintaining the shift towards renewable solar energy. Cloud classification techniques are also key to identification of dangerous weather hazards. Weather prediction systems, convective current models, and rainfall estimation systems all benefit from accurate cloud identification (18). With more than 11,000 natural disasters causing over two million deaths and USD 3.64 trillion in economic damage — all between 1970 and 2019 (nearly a 5x

[*]Equal Contribution

Tackling Climate Change with Machine Learning workshop at NeurIPS 2021.

increase) (10), this can be a crucial technological intervention. Cloud segmentation is important to understand the impact of cloud cover; estimation of local temperature and photo-voltaic power output in the area would be possible.

Ground-based cloud tasks on images taken from Earth's ground-level have been explored extensively in recent literature (13). However, in most practical systems, it still requires heavy manual intervention of domain experts. This subjectivity in evaluation process impacts large-scale projects with the because of inherent biases and uncertainties. To reduce the burden on domain experts and produce reliable technological intervention, classical ML and DL-based classification and segmentation algorithms have emerged.

The last few years have witnessed a tremendous rise in CNN-based image classification and segmentation research (17). However, CNNs have been known to suffer from large inductive biases. In particular, locality and translation equivariance are crucial components of a CNN. These assumptions hurt the interpretability of CNN algorithms. Of late, visual transformer-based approaches have gained massive traction for image classification, segmentation and object detection tasks (1; 15; 20; 2). They have been known to challenge some of these assumptions while maintaining comparable performance.

In this work, we propose a pure visual transformer: CloudViT and a hybrid CNN-Transformer: CloudUT to tackle cloud classification on the Cirrus Cumulus Stratus Nimbus Dataset (CCSN) (18) and cloud segmentation on the SWIMSWG Dataset (6) respectively. We achieve state-of-the-art results in both tasks and therefore note that the discriminative power of the transformer-driven approach for the first time (to the best of our knowledge) in this domain. However keeping quantiative results aside, this approach maintains the neighbourhood structure during the generation of patch embeddings, and thus does not induce spatial biases since all inter-patch dependencies are learnt from scratch. This means that the aforementioned inductive biases do not occur throughout the network, unlike for CNNs. Crucially, this improves the robustness of our approach.

2 Methodology

We approach the cloud classification task as a multi-class classification problem. Given an input cloud image, our pure transformer model CloudViT outputs a label prediction. We approach the segmentation task to achieve classification at a granular level. Our hybrid model CloudUT outputs a dense label prediction for every pixel, dividing them into distinct regions.

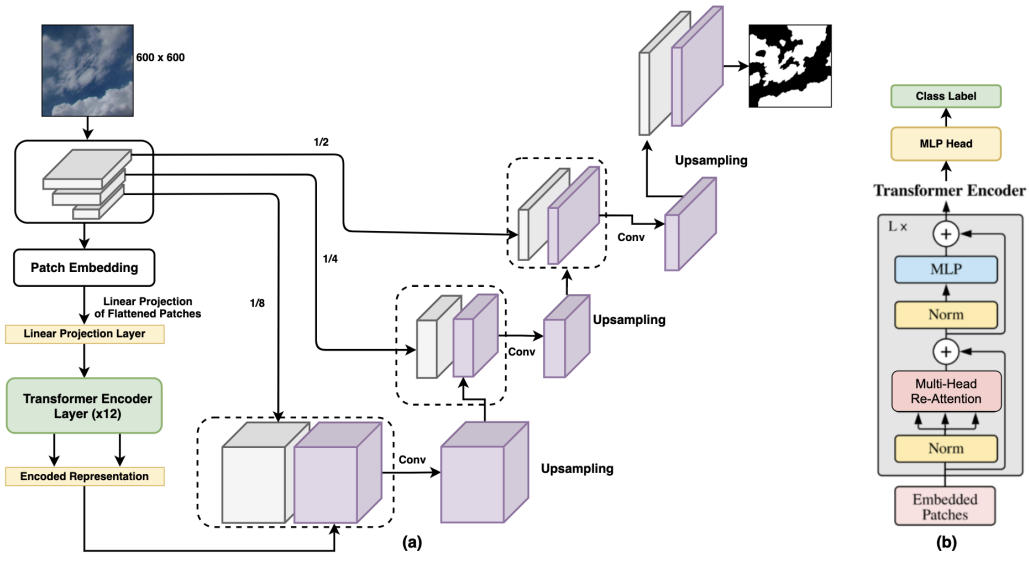

Figure 1: (a) Cloud U-Net Transformer (CloudUT) (b) Cloud Vision Transformer (CloudViT)

2.1 Cloud Vision Transformer (CloudViT):

We implement CloudViT, a modified visual Transformer model as proposed by (19), adapted from the original Vision Transformer(7). The transformer architecture consists of three parts — patch embeddings, encoder, and classification head as shown in the illustration in Fig. 1(b).

Patch Embeddings: Analogous to NLP, in CV an image is tokenized into a sequence of patches and then converted into patch embeddings using a linear layer. Concretely, the image I is cropped into N patches of resolution $P \times P$ and $N = HW/P^2$. The patch embeddings (14) retain spatial information in the transformer. To augment positional information, a learnable position embedding is added to the patch embeddings.

$$\mathbf{z}_0 = [\mathbf{x}_{cls}; \mathbf{E}_{patch}] + \mathbf{E}_{pos} \tag{1}$$

Encoder: Self-attention is a key concept in transformers. The attention weight A_{ij} is computed by using the similarity between query $\mathbf{q^i}$ and key $\mathbf{k^j}$. The output representation is generated by projecting the attention-weighted sum across all values in the sequence using a linear layer. In CloudViT, the Multi-Head Self-attention (MSA) layer with is replaced with a Multi-Head Re-attention (MRA) layer, which resolves "attention collapse", improves performance and helps train deeper ViT models. The learn-able transformation matrix $\Theta \in \mathbb{R}^{H \times H}$ is used to regenerate the multi-head attention maps into newer ones. Normalization is applied to reduce the layer-wise variance which is very useful in learning cloud representations of larger dimensions more accurately without “attention collapse”.

$$Re{-}Attention(Q, K, V) = Norm\left(\Theta^\top \left(Softmax\left(\frac{QK^\top}{\sqrt{d}}\right)\right)\right) V \tag{2}$$

Classification Head: The encoded representation for the prepended positional embedding is used for classification. A linear layer is applied to generate the classification predictions $\mathbf{y}$, given by Eq. 3.

$$\mathbf{y} = MLP(LN(\mathbf{z}_L^0)) \tag{3}$$

2.2 Cloud U-Net Transformer (CloudUT):

Segmenting clouds from sky-images is challenging due to the absence of sharp boundaries in the 'twilight zone' — a belt of evaporating fragments (9). We implement a hybrid CNN-Transformer model CloudUT with the familar encoder-decoder pattern, as illustrated in Figure 1(a).

Encoder: Given an input sky image, $\mathbf{x} \in \mathbb{R}^{L \times W \times C}$ we predict pixel-wise labels of size L x W by training the U-Net architecture (12) (CNN) to encode the input image into high-level feature representations. Cloud images possess several patterns and textures, and these intricate patterns are used in describing the class of clouds. We leverage self-attention with a hybrid encoder, by passing encoded U-Net feature representations into our Transformer model.

$$\mathbf{z}'_\ell = MRA\left(LN\left(\mathbf{z}_{\ell-1}\right)\right) + \mathbf{z}_{\ell-1}, \mathbf{z}_\ell = MLP\left(LN\left(\mathbf{z}'_\ell\right)\right) + \mathbf{z}'_\ell \tag{4}$$

Decoder: These encoded representations from the Transformer are decoded back into the segmented maps of full resolution, in a U-Net++ Decoder. The expansive path or the upsampling part of the U-Net has a large number of feature channels, which allow the network to propagate context information to higher resolution layers. Skip-connections between the encoder and decoder helps in easier transfer of gradients for calculations and concatenation of previous feature maps generated by encoder to ensure the preservation of spatial structure.

3 Experiments

3.1 Dataset Descriptions

Cirrus Cumulus Stratus Nimbus Dataset (CCSN): (18) Consists of 2543 ground-based cloud images split into 11 unique classes as shown in Fig. 2. All images are of 256x256 pixels resolution

and have large variations in illumination and heavy intra-class semantic variations. To ensure that a high variety of atmospheric conditions is represented, the dataset was examined by filtering via various parameters.

Figure 2: Cirrus Cumulus Stratus Nimbus Dataset(CCSN) Dataset

Singapore Whole Sky Imaging Segmentation Database (SWIMSWG): (6) Segmentation task that consists of 1013 images of resolution 600x600 pixels, that were shot on a custom sky image device at different camera elevation angles. Fig 3 shows a few input images. There is wide variation in illumination and color (which are only partially corrected) due to differences in time of the day, cloud coverage, and distance from sun.

3.2 Results

Cloud Classification As per (18), we split the CCSN dataset into the same training and test set. We evaluate the performance of our CloudViT model on the CCSN dataset based on class-wise accuracy, Precision, Recall and F1 score, as shown in Table 2. We also compare these performances with other benchmark and custom models. As shown in Table 1(a), CloudViT achieves state-of-the-art classification accuracy of 90.06%.

Cloud Segmentation Similar to (16), we split the SWIMSWG Dataset into 860 images in training set and 100 in test set. We use mIoU (Mean Intersection over Union) and mDice scores to evaluate our segmentation model as shown in Table 1(b). To show the robustness of our model, we also segment cloud images from CCSN Dataset and generate segmentation masks as output. Compared to the benchmark U-Net model (12), our CloudUT model achieves state-of-the-art mIoU of 0.832.

	Precision	Recall	F1	Accuracy
ResNet	0.84	0.82	0.81	83.30
CNet (18)	0.84	0.87	0.86	87.62
CloudViT	**0.91**	**0.92**	**0.89**	**90.06**

Model	mIoU	mDice
U-Net (12)	0.7626	0.8388
DeepLabV3 (3)	0.6281	0.7036
PLS (4)	0.6467	0.6919
CloudUT (Ours)	**0.832**	**0.8927**

Table 1: Model comparison on (a) CCSN dataset and (b) SWIMSWG dataset

	Ci	Ac	As	Cu	Cb	Ct	Sc	St	Mean
Accuracy (%)	91.30	94.50	69.29	96.73	91.70	100	68.92	88.14	90.06
F1 score	0.96	0.96	0.59	0.96	0.92	1	0.69	0.82	0.89
Precision	0.92	0.92	1	0.92	0.92	1	0.72	0.9	0.91
Recall	1	1	0.72	1	0.92	1	1	0.75	0.92

Table 2: Classification metrics on CCSN Dataset

4 Discussion and Conclusion

In this work, we propose two visual transformer-based approaches: CloudViT and CloudUT for cloud classification and segmentation respectively. We achieve state-of-the-art results in both tasks and empirically demonstrate the power and robustness of visual transformers for cloud image data. To the best of our knowledge, we are the first work to extend visual transformers to cloud data domain and demonstrate a deviation from popular CNN-based approaches. With these results, we hope to stimulate further research in pure transformer-based approaches for semantic segmentation, as well as in CV for cloud data in general.

We believe this technology can be deployed on edge devices and be used in the domains of weather forecasting, climate studies, water cycle studies, solar irradiance prediction and other phenomena.

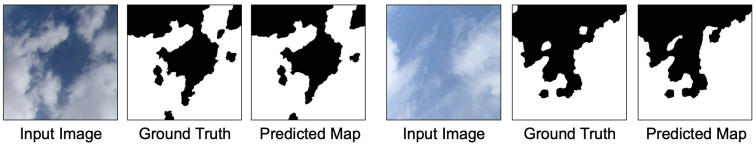

Figure 3: Segmentation Results from CloudUT Model

This automated DL method can also be used to save human expert hours, reduce inherent biases and consequently create larger-scale datasets to study global warming.

References

[1] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit. Understanding robustness of transformers for image classification. *ArXiv*, abs/2103.14586, 2021.

[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587, 2017.

[4] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587, 2017.

[5] T. Chen, W. B. Rossow, and Y. Zhang. Radiative effects of cloud-type variations. *Journal of Climate*, 13(1):264 – 286, 2000.

[6] S. Dev, Y. H. Lee, and S. Winkler. Color-based segmentation of sky/cloud images from ground-based cameras. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(1):231–242, 2017.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[8] D. Duda, P. Minnis, K. Khlopenkov, T. Chee, and R. Boeke. Estimation of 2006 northern hemisphere contrail coverage using modis data. *Geophysical Research Letters*, 40:612–617, 02 2013.

[9] Y. Fabel. *Cloud Segmentation and Classification from All-Sky Images Using Deep Learning*. PhD thesis, 07 2020.

[10] M. McGrath. Climate change: Big increase in weather disasters over the past five decades. 2021.

[11] Z. Peng, D. Yu, D. Huang, J. Heiser, S. Yoo, and P. Kalb. 3d cloud detection and tracking system for solar forecast using multiple sky imagers. *Solar Energy*, 118:496–519, 2015.

[12] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[13] G. Terrén-Serrano and M. Martínez-Ramón. Comparative analysis of methods for cloud segmentation in ground-based infrared images. *Renewable Energy*, 175:1025–1040, 2021.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.

[15] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. Álvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *ArXiv*, abs/2105.15203, 2021.

[16] W. Xie, D. Liu, M. Yang, S. Chen, B. Wang, Z. Wang, Y. Xia, Y. Liu, Y. Wang, and C. Zhang. Segcloud: a novel cloud image segmentation model using a deep convolutional neural network for ground-based all-sky-view camera observation. *Atmospheric Measurement Techniques*, 13(4):1953–1961, 2020.

[17] M. Yapıcı, A. Tekerek, and N. Topaloglu. Literature review of deep learning research areas. 5:188–215, 12 2019.

[18] J. Zhang, L. Pu, F. Zhang, and Q. Song. Cloudnet: Ground-based cloud classification with deep convolutional neural network. *Geophysical Research Letters*, 45, 08 2018.

[19] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng. Deepvit: Towards deeper vision transformer, 2021.

[20] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.