Institut für Photogrammetrie
und GeoInformation

Leibniz
Universität
Hannover

Institut für Photogrammetrie und GeoInformation,
Nienburger Straße 1, 30167 Hannover

Multi-temporal land cover classification with transformer models using satellite image time series

Proposal for a Master thesis topic (DE/EN)

Nienburger Straße 1, 30167 Hannover
Fakultät für Bauingenieurwesen und Geodäsie

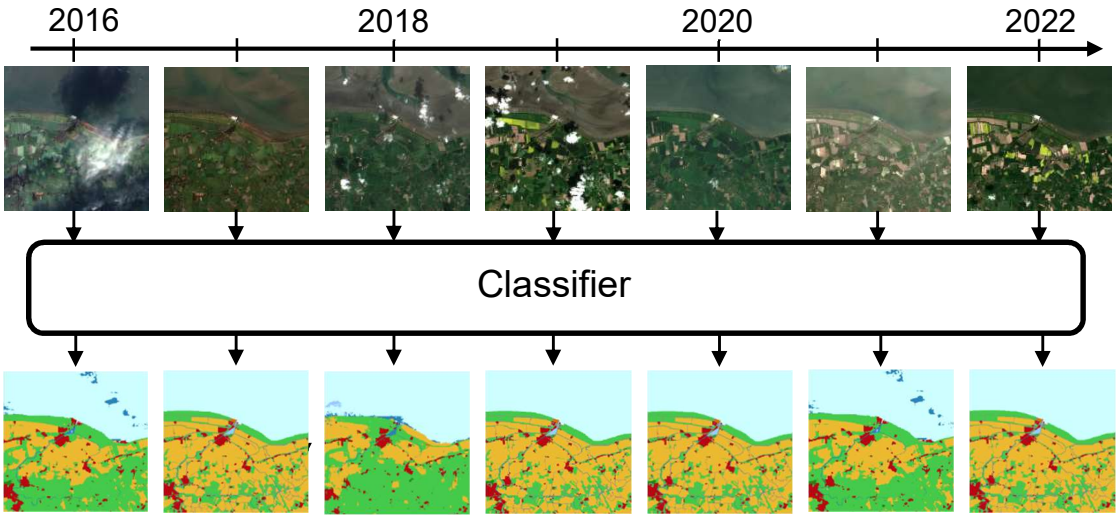
Institut für Photogrammetrie
und GeoInformation

Prof. Dr.-habil. Christian Heipke

Mirjana Voelsen

Tel.+49 511 762-4401
Fax +49 511 762-2483
E-Mail: voelsen
@ipi.uni-hannover.de

Classification of land cover is a standard task in remote sensing, in which each image pixel is assigned a class label indicating the physical material of the object surface (e.g. grass, building). This task is highly relevant for applications such as the detection of changes or rapid mapping. Recent work has focused on Deep Neural Networks (DNNs), delivering considerably better results than traditional classifiers. This is mainly due to the fact that, unlike traditional classifiers using hand-crafted features, DNNs provide a framework in which these features can be learned from training data. One of the state-of-the-art DNN-architectures are fully convolutional neural networks (FCNs) that compute these features based on the local neighborhood of a pixel and provide one output class per pixel (also called semantic segmentation). When multitemporal input images are used, besides FCNs, several other architectures achieve state-of-the-art results. Transformer networks, originally applied to natural language processing tasks, can handle input sequences of arbitrary length and are able to predict one output for each input. These models were also applied to image classification, e. g. by Strudel et al. (2021) or Liu et al. (2021), who split the images to fixed size patches and flatten them by a linear projection, because the computational complexity increases quadratically with the number of input tokens. This enables them to use a transformer architecture from the NLP field. Within a Transformer network, weights (so-called attentions) are computed between all tokens (e.g. words) of the input sequence. Of course, using patches instead of pixels comes at the cost of losing spatial resolution. There are approaches that mitigate this effect by computing the attentions only in local windows, that include several patches and thus a smaller patch size is possible. A lot of approaches combine convolutional and transformer layers to compute local and global features, respectively. When satellite image time



Besucheradresse:
Nienburger Straße 1
30167 Hannover
www.ipi.uni-hannover.de

series are used, not only spatial features need to be extracted but also temporal ones, raising the question which kind of model or model combination is beneficial for this kind of input data.

The main goal of this thesis is the extension of an existing Transformer model for the pixel-wise classification of land cover with multi-temporal satellite images. As a baseline, an existing model based on a Swin Transformer backbone with a FCN decoder is available that predicts a land cover map for each input timestep. This model shall be modified, for instance, an **extension to a varying number of input and output timesteps** would be interesting to investigate. This can also include the extension of the time period that is used to create one input timeseries, which is currently fixed to one year. Another interesting approach is a **preceding FCN to extract spatial features** from the images first and use them as the input to a transformer model. To compare different network variants and parameter settings, a series of experiments shall be conducted and evaluated. The obtained results will also be compared to the existing Transformer and FCN models. The student will be provided with datasets as well as initial classification architectures. Previous knowledge in the field of image analysis and programming is mandatory. The thesis can be written in English or in German.

This thesis will be supervised be Mirjana Voelsen M.Sc.

References

STRUDEL, R.; GARCIA, R.; LAPTEV, I.; SCHMID, C., 2021: Detecting Segmenter: Transformer for Semantic Segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), *pp. 7262 – 7272.*

LIU, Z.; LIN, Y.; CAO, Y.; HU, H.; WEI, Y.; ZHANG, Z.; LIN, S.; GUO, B., 2021: SwinTransformer: Hierarchical Vision Transformer using Shifted Window. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), *pp. 9992 – 10002.*