AI-based Cloudmask Processor for Sentinel-2

Marharyta Domnich
• Established in 2015, Tartu Observatory (the Estonian hub of space science R&D) spin-off.
• 13 employees (4 PhDs) in Tartu and Kuressaare.
• Focus on developing agricultural services and providing value-added Sentinel-1 analysis ready data layers via API (B2B and B2G).

\[ k_z = \frac{2\pi \Delta \theta_i}{\lambda \sin \theta_i} \approx \frac{2B_n}{R \sin \theta_i} \]

(vertical wave number of InSAR)
Outline

- Cloudmask approach
- Data preparation
- Model fitting
- Evaluation
- Results
Project goal

Develop the most accurate cloud mask for Sentinel-2 for Northern European terrestrial summer season conditions.

Open-source approach
Cloudmask modules

- Data preprocessing
- Open source data integration
- Image labeling
- Labels postprocessing
- Model fitting
- Model prediction
- Comparison with Sen2Cor, Fmask, S2cloudless
- Active learning
Sentinel-2 open-source labelled datasets

- Sentinel-2 Cloud Mask Catalogue by Francis Alistair et al. November 2020
  (https://zenodo.org/record/4172871#.YH297ugzZPY)
  - 513 1022x1022-pixel cropped subscenes at 20 m resolution
- Sentinel-2 reference cloud masks generated by an active learning method
  - 32 full Sentinel-2 annotated scenes at 60 m resolution

Figure 2. Global distribution of selected Sentinel-2 scenes which are included in the database.
# KappaZeta labelled dataset

<table>
<thead>
<tr>
<th>Label</th>
<th>Class</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>INVALID</td>
<td>Missing or invalid pixels</td>
</tr>
<tr>
<td>1</td>
<td>CLEAR</td>
<td>Pixels without cloud or cloud shadow</td>
</tr>
<tr>
<td>2</td>
<td>CLOUD SHADOW</td>
<td>Cloud shadow’s pixels</td>
</tr>
<tr>
<td>3</td>
<td>SEMI TRANSPARENT CLOUD</td>
<td>Pixels of thin clouds through which the land is visible</td>
</tr>
<tr>
<td>4</td>
<td>CLOUD</td>
<td>Pixels of cloud</td>
</tr>
<tr>
<td>5</td>
<td>UNDEFINED</td>
<td>Labeler is not sure which class these pixels belongs to</td>
</tr>
</tbody>
</table>
S2 products selected for labelling

- STRATUS
- CUMULUS
- CIRRUS
Temporal distribution of S2 products

- For each month from April to October 30 Sentinel-2 products are selected
  - Represents all common cloud types: cirrus, cumulus, stratus

- 10 sub-tiles from each product are labelled

30 products * 7 months * 10 sub-tiles = 2100 sub-tiles
Sub-tiling procedure

Full S2 image 10980x10980 pixels

484 sub-tiles 512x512 pixels
# Full Sentinel-2 images labelled in Estonia

<table>
<thead>
<tr>
<th>Sentinel 2 product name</th>
<th>Preview</th>
<th>Description</th>
</tr>
</thead>
</table>
| S2B_MSIL2A_20200501T093029_N0214_R136_T35VMF_20200501T13501.SAFE | | Cloud cover percentage: 17%  
Cloud shadow percentage: 0.1%  
Vegetation percentage: 42%  
Thin cirrus percentage: 15.2%  
Water percentage: 4.6% |
| S2A_MSIL2A_20200509T094041_N0214_R036_T35VMF_20200509T11504.SAFE | | Cloud cover percentage: 28.7%  
Cloud shadow percentage: 10.9%  
Thin cirrus percentage: 2.9%  
Vegetation percentage: 39.8%  
Water percentage: 3.3% |
| S2A_MSIL2A_20200529T094041_N0214_R036_T35VLF_20200529T120441.SAFE | | Cloud cover percentage: 17.2%  
Cloud shadow percentage: 6.3%  
Thin cirrus percentage: 0.3%  
Vegetation percentage: 58.1%  
Water percentage: 12.4% |
| S2_A_MSIL2A_20200715T093041_N0214_R136_T35VMD_20200715T122147.SAFE | | Cloud cover percentage: 90.5%  
Cloud shadow percentage: 0.3%  
Thin cirrus percentage: 41.4%  
Vegetation percentage: 8.8%  
Water percentage: 0.05% |
| S2A_MSIL2A_20200728T094041_N0214_R036_T35VND_20200728T115051.SAFE | | Cloud cover percentage: 71.7%  
Cloud shadow percentage: 1.1%  
Thin cirrus percentage: 6.2%  
Vegetation percentage: 16.4%  
Water percentage: 0.2% |
| S2A_MSIL2A_20200824T093041_N0214_R136_T35VND_20200824T121941.SAFE | | Cloud cover percentage: 29.9%  
Cloud shadow percentage: 10.1%  
Thin cirrus percentage: 0.004%  
Vegetation percentage: 46.4%  
Water percentage: 0.4% |

1 S2 product (10980x10980) = 484 sub-tiles 512x512 pixels

Six S2 products = 2904 labelled subscenes

Final: 5004 labelled sub-tiles
Labelling tools

**CVAT**
- Pixels smoothing
- Vector output (XML)
- Good API
- Max sub-tile size 512x512

**Segments.ai**
- No interpolation
- Output raster (TIFF)
- Good API
- Allows upload predictions
Development pipeline
## Input NetCDF files

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Ground resolution</th>
<th>Bit-depth</th>
<th>Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label</td>
<td>10 m</td>
<td>8</td>
<td>Mask</td>
<td>Segmentation mask from one of the supported labeling tools</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>TCI</td>
<td>10 m</td>
<td>8</td>
<td>RGB</td>
<td>True Color Image</td>
</tr>
<tr>
<td>2</td>
<td>SCL</td>
<td>20 m</td>
<td>8</td>
<td>Mask</td>
<td>Sen2Cor Scene Classification Image</td>
</tr>
<tr>
<td>3</td>
<td>AOT</td>
<td>10 m</td>
<td>16</td>
<td>Grayscale</td>
<td>Aerosol Optical Thickness map</td>
</tr>
<tr>
<td>4</td>
<td>B01</td>
<td>60 m</td>
<td>16</td>
<td>Grayscale</td>
<td>433 - 453 nm</td>
</tr>
<tr>
<td>5</td>
<td>B02</td>
<td>10 m</td>
<td>16</td>
<td>Grayscale</td>
<td>457.5 - 522.5 nm</td>
</tr>
<tr>
<td>6</td>
<td>B03</td>
<td>10 m</td>
<td>16</td>
<td>Grayscale</td>
<td>542.5 - 577.5 nm</td>
</tr>
<tr>
<td>7</td>
<td>B04</td>
<td>10 m</td>
<td>16</td>
<td>Grayscale</td>
<td>650 - 680 nm</td>
</tr>
<tr>
<td>8</td>
<td>B05</td>
<td>20 m</td>
<td>16</td>
<td>Grayscale</td>
<td>697.5 - 712.5 nm</td>
</tr>
<tr>
<td>9</td>
<td>B06</td>
<td>20 m</td>
<td>16</td>
<td>Grayscale</td>
<td>732.5 - 747.5 nm</td>
</tr>
<tr>
<td>10</td>
<td>B07</td>
<td>20 m</td>
<td>16</td>
<td>Grayscale</td>
<td>773 - 793 nm</td>
</tr>
<tr>
<td>11</td>
<td>B08</td>
<td>10 m</td>
<td>16</td>
<td>Grayscale</td>
<td>784.5 - 899.5 nm</td>
</tr>
<tr>
<td>12</td>
<td>B09</td>
<td>60 m</td>
<td>16</td>
<td>Grayscale</td>
<td>855 - 875 nm</td>
</tr>
<tr>
<td>13</td>
<td>B10</td>
<td>60 m</td>
<td>16</td>
<td>Grayscale</td>
<td>935 - 955 nm</td>
</tr>
<tr>
<td>14</td>
<td>B11</td>
<td>60 m</td>
<td>16</td>
<td>Grayscale</td>
<td>1358 - 1389 nm</td>
</tr>
<tr>
<td>15</td>
<td>B12</td>
<td>20 m</td>
<td>16</td>
<td>Grayscale</td>
<td>1565 - 1655 nm</td>
</tr>
<tr>
<td>16</td>
<td>WVP</td>
<td>10 m</td>
<td>16</td>
<td>Grayscale</td>
<td>Water Vapour map</td>
</tr>
<tr>
<td>18</td>
<td>GML</td>
<td>Vector</td>
<td></td>
<td>GML vector mask</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>S2CC</td>
<td>20 m</td>
<td>8</td>
<td>Mask</td>
<td>Sen2Cor cloud probabilities</td>
</tr>
<tr>
<td>20</td>
<td>S2CS</td>
<td>20 m</td>
<td>8</td>
<td>Mask</td>
<td>Sen2Cor snow probabilities</td>
</tr>
<tr>
<td>21</td>
<td>FMSC</td>
<td>20 m</td>
<td>8</td>
<td>Mask</td>
<td>Fmask classification map</td>
</tr>
<tr>
<td>22</td>
<td>SS2C</td>
<td>60 m</td>
<td>8</td>
<td>Mask</td>
<td>Sinergise S2Cloudless classification map</td>
</tr>
<tr>
<td>23</td>
<td>SS2CC</td>
<td>60 m</td>
<td>8</td>
<td>Mask</td>
<td>Sinergise S2Cloudless cloud probabilities</td>
</tr>
<tr>
<td>24</td>
<td>BHC</td>
<td>60 m</td>
<td>8</td>
<td>Mask</td>
<td>Baetens &amp; Hagolle classification map</td>
</tr>
<tr>
<td>25</td>
<td>FMSC</td>
<td>20 m</td>
<td>8</td>
<td>Mask</td>
<td>Francis, Mrziglod, and Sidiropoulos classification map</td>
</tr>
</tbody>
</table>
Model fitting setup

- Data normalization: z-score, min-max
- Class imbalance:
  - 40% data – clear,
  - 7% - cloud shadow,
  - 11% - semi-transparent,
  - 29% - cloud
- Features used for training:
  - B01, B02, B03, B04, B05, B06, B07, B08, B8A, B09, B11, B12, AOT, WVP
Unet and Unet++

- Input size: 512x512x13
- Output size: 512x512x1
- Final output is generated by taking argmax of the probabilities
- Trying different layer depth and filters
- Loss: Binary cross-entropy, dice loss, weighted dice
- Unet++ allows deep supervision
Evaluation

**KAPPAMASK**

Confusion matrix for KappaMask, dice score: 0.92

<table>
<thead>
<tr>
<th></th>
<th>CLEAR</th>
<th>CLOUD_SHADOW</th>
<th>SEMI_TRANSPARENT_CLOUD</th>
<th>CLOUD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEAR</td>
<td>0.95</td>
<td>0.07</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>CLOUD_SHADOW</td>
<td>0.07</td>
<td>0.86</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>SEMI_TRANSPARENT_CLOUD</td>
<td>0.04</td>
<td>0.02</td>
<td>0.89</td>
<td>0.07</td>
</tr>
<tr>
<td>CLOUD</td>
<td>0.02</td>
<td>0.03</td>
<td>0.94</td>
<td></td>
</tr>
</tbody>
</table>

**Sen2Cor**

Confusion matrix for sen2cor, dice score: 0.57

<table>
<thead>
<tr>
<th></th>
<th>CLEAR</th>
<th>CLOUD_SHADOW</th>
<th>SEMI_TRANSPARENT_CLOUD</th>
<th>CLOUD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEAR</td>
<td>0.71</td>
<td>0.01</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>CLOUD_SHADOW</td>
<td>0.44</td>
<td>0.42</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>SEMI_TRANSPARENT_CLOUD</td>
<td>0.17</td>
<td>0.03</td>
<td>0.23</td>
<td>0.56</td>
</tr>
<tr>
<td>CLOUD</td>
<td>0.16</td>
<td>0.04</td>
<td>0.08</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Validation set, 692 sub-tiles
### Sub-tile comparison with other masks

<table>
<thead>
<tr>
<th>Original image</th>
<th>Label</th>
<th>KappaMask Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Original Image" /></td>
<td><img src="image2.png" alt="Label" /></td>
<td><img src="image3.png" alt="KappaMask Prediction" /></td>
</tr>
<tr>
<td><img src="image4.png" alt="Sen2Cor" /></td>
<td><img src="image5.png" alt="Fmask" /></td>
<td><img src="image6.png" alt="S2cloudless" /></td>
</tr>
</tbody>
</table>

**Legend:**
- **Sen2Cor**
- **Fmask**
- **S2cloudless**
Original sub-tile

Comparison of KappaMask prediction and Sen2Cor

Legend:
- True cloud
- True clear
- True cloudshadow
- True undefined
- True semitransparent
- Both models false
- Pred. true, s2c false
- S2c true, pred. false
Slider comparison with rule-based masks

https://kappazeta.ee/cloudcomparison

https://kappazeta.ee/cloudcomparison2
Slider comparison with machine learning-based mask

- S2cloudless was trained with XGBoost and LightGBM
- claims to outperform Sen2Cor and Fmask
- generates only cloud vs non-cloud mask
Full Sentinel-2 mask
Whole Sentinel-2 tile mask

<table>
<thead>
<tr>
<th>Full S2 original image</th>
<th>KappaMask prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Full S2 original image" /></td>
<td><img src="image2.png" alt="KappaMask prediction" /></td>
</tr>
</tbody>
</table>
Summary of accomplishments

• ~3500 labelled Sentinel-2 sub-tiles on 10 m resolution
• Active learning pipeline for selecting the worst predicted sub-tiles for future labelling
• Interactive tool for analysing whole Sentinel-2 mask
• Validation framework for the comparison of different masks (Sen2Cor, Fmask, S2cloudless by Sinergise)
• Validation results:
dice coefficient for KappaMask - 92%, Sen2Cor - 57%
Thank you for your attention!