SLICING AIDED HYPER INFERENCE AND FINE-TUNING FOR SMALL OBJECT DETECTION

Fatih Cagatay Akyon\textsuperscript{1,2}, Sinan Onur Altinuc\textsuperscript{1,2}, Alptekin Temizel\textsuperscript{2}

\textsuperscript{1}OBSS AI, Ankara, Turkey
\textsuperscript{2}Graduate School of Informatics, Middle East Technical University, Ankara, Turkey
Problem Definition

detecting small objects
5-10 px
in large images
1000+ px
Motivation

Pretraining with large objects on low-res images

COCO pretraining

model with general purpose weights

model with task-specific weights

Fine-tuning with small objects on high-res images

update architecture for small objects
Motivation

- Model with random weights
  - COCO pretraining

- Model with general purpose weights
  - Model with task-specific weights
  - Fine-tune with large input size
  - Small object fine-tuning
Motivation

- Requires large GPU memory
- Low GPU utilization
Slicing Aided Fine-tuning
Slicing Aided Hyper Inference (SAHI)

Full Inference (FI)

Slicing Aided Hyper Inference (SAHI)
Experiment Setup

Datasets:
- Visdrone:
  - 10 object categories
  - 6471 training images
- xView:
  - 60 object categories
  - 846 training images

Training Framework:
- Pytorch (v1.10.0)
- MMDetection (v2.21.0)
Experiment Setup

Object Detection Models:
- FCOS: Fully Convolutional One-Stage Object Detection
  - Anchor box free,Eliminates anchor-box related hyperparameters
  - Only requires NMS as post-processing
- VarifocalNet: An IoU-aware Dense Object Detector
  - Learns to predict the IoU-aware classification score which mixes the object presence confidence and localization accuracy together as the detection score for a bounding box.
- TOOD: Task-aligned One-stage Object Detection
  - Explicitly aligns the two tasks in a learning-based manner.

Notations:
- FI: Full-Image inference
- SAHI: Slicing aided inference
- PO: Patch Overlap
- SF: Slicing aided fine-tuning
Evaluation Results: Visdrone Dataset
### Evaluation Results: Visdrone Dataset

<table>
<thead>
<tr>
<th>Setup</th>
<th>AP&lt;sub&gt;50&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;50s&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;50m&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;50l&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCOS+FI</td>
<td>25.8</td>
<td>14.2</td>
<td>39.6</td>
<td>45.1</td>
</tr>
<tr>
<td>FCOS+SAHI+PO</td>
<td>29.0</td>
<td>18.9</td>
<td>41.5</td>
<td>46.4</td>
</tr>
<tr>
<td>FCOS+SAHI+FI+PO</td>
<td>31.0</td>
<td>19.8</td>
<td>44.6</td>
<td>49.0</td>
</tr>
<tr>
<td>FCOS+SF+SAHI+PO</td>
<td>38.1</td>
<td>25.7</td>
<td>54.8</td>
<td>56.9</td>
</tr>
<tr>
<td>FCOS+SF+SAHI+FI+PO</td>
<td>38.5</td>
<td>25.9</td>
<td>55.4</td>
<td>59.8</td>
</tr>
<tr>
<td>VFNet+FI</td>
<td>28.8</td>
<td>16.8</td>
<td>44.0</td>
<td>47.5</td>
</tr>
<tr>
<td>VFNet+SAHI+PO</td>
<td>32.0</td>
<td>21.4</td>
<td>45.8</td>
<td>45.5</td>
</tr>
<tr>
<td>VFNet+SAHI+FI+PO</td>
<td>33.9</td>
<td>22.4</td>
<td>49.1</td>
<td>49.4</td>
</tr>
<tr>
<td>VFNet+SF+SAHI+PO</td>
<td>41.9</td>
<td>29.7</td>
<td>58.8</td>
<td>60.6</td>
</tr>
<tr>
<td>VFNet+SF+SAHI+FI+PO</td>
<td>42.2</td>
<td>29.6</td>
<td>59.2</td>
<td>63.3</td>
</tr>
<tr>
<td>TOOD+FI</td>
<td>29.4</td>
<td>18.1</td>
<td>44.1</td>
<td>50.0</td>
</tr>
<tr>
<td>TOOD+SAHI</td>
<td>31.9</td>
<td>22.6</td>
<td>44.0</td>
<td>45.2</td>
</tr>
<tr>
<td>TOOD+SAHI+PO</td>
<td>32.5</td>
<td>22.8</td>
<td>45.2</td>
<td>43.6</td>
</tr>
<tr>
<td>TOOD+SAHI+FI</td>
<td>34.6</td>
<td>23.8</td>
<td>48.5</td>
<td>53.1</td>
</tr>
<tr>
<td>TOOD+SAHI+FI+PO</td>
<td>34.7</td>
<td>23.8</td>
<td>48.9</td>
<td>50.3</td>
</tr>
<tr>
<td>TOOD+SF+FI</td>
<td>36.8</td>
<td>24.4</td>
<td>53.8</td>
<td>66.4</td>
</tr>
<tr>
<td>TOOD+SF+SAHI</td>
<td>42.5</td>
<td>31.6</td>
<td>58.0</td>
<td>61.1</td>
</tr>
<tr>
<td>TOOD+SF+SAHI+PO</td>
<td>43.1</td>
<td>31.7</td>
<td>59.0</td>
<td>60.2</td>
</tr>
<tr>
<td>TOOD+SF+SAHI+FI</td>
<td>43.4</td>
<td>31.7</td>
<td>59.6</td>
<td>65.6</td>
</tr>
<tr>
<td>TOOD+SF+SAHI+FI+PO</td>
<td>43.5</td>
<td>31.7</td>
<td>59.8</td>
<td>65.4</td>
</tr>
</tbody>
</table>

- SAHI increases object detection AP by up to 6.8%.
- With SF, object detection AP increases up to 14.5% AP.
- Applying 25% overlap between slices during inference, increases small/medium object AP and overall AP.
### Evaluation Results: xView Dataset

<table>
<thead>
<tr>
<th>Setup</th>
<th>AP$_{50}$</th>
<th>AP$_{50s}$</th>
<th>AP$_{50m}$</th>
<th>AP$_{50l}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCOS+FI</td>
<td>2.20</td>
<td>0.10</td>
<td>1.80</td>
<td>7.30</td>
</tr>
<tr>
<td>FCOS+SF+SAHI</td>
<td>15.8</td>
<td>11.9</td>
<td>18.4</td>
<td>11.0</td>
</tr>
<tr>
<td>FCOS+SF+SAHI+PO</td>
<td><strong>17.1</strong></td>
<td><strong>12.2</strong></td>
<td><strong>20.2</strong></td>
<td><strong>12.8</strong></td>
</tr>
<tr>
<td>FCOS+SF+SAHI+FI</td>
<td>15.7</td>
<td>11.9</td>
<td>18.4</td>
<td>14.3</td>
</tr>
<tr>
<td>FCOS+SF+SAHI+FI+PO</td>
<td><strong>17.0</strong></td>
<td><strong>12.2</strong></td>
<td><strong>20.2</strong></td>
<td><strong>15.8</strong></td>
</tr>
<tr>
<td>VFNet+FI</td>
<td>2.10</td>
<td>0.50</td>
<td>1.80</td>
<td>6.80</td>
</tr>
<tr>
<td>VFNet+SF+SAHI</td>
<td>16.0</td>
<td>11.9</td>
<td>17.6</td>
<td>13.1</td>
</tr>
<tr>
<td>VFNet+SF+SAHI+PO</td>
<td><strong>17.7</strong></td>
<td><strong>13.7</strong></td>
<td><strong>19.7</strong></td>
<td><strong>15.4</strong></td>
</tr>
<tr>
<td>VFNet+SF+SAHI+FI</td>
<td>15.8</td>
<td>11.9</td>
<td>17.5</td>
<td>15.2</td>
</tr>
<tr>
<td>VFNet+SF+SAHI+FI+PO</td>
<td><strong>17.5</strong></td>
<td><strong>13.7</strong></td>
<td><strong>19.6</strong></td>
<td><strong>17.6</strong></td>
</tr>
<tr>
<td>TOOD+FI</td>
<td>2.10</td>
<td>0.10</td>
<td>2.00</td>
<td>5.20</td>
</tr>
<tr>
<td>TOOD+SF+SAHI</td>
<td>19.4</td>
<td>14.6</td>
<td>22.5</td>
<td>14.2</td>
</tr>
<tr>
<td>TOOD+SF+SAHI+PO</td>
<td><strong>20.6</strong></td>
<td><strong>14.9</strong></td>
<td><strong>23.6</strong></td>
<td><strong>17.0</strong></td>
</tr>
<tr>
<td>TOOD+SF+SAHI+FI</td>
<td>19.2</td>
<td>14.6</td>
<td>22.3</td>
<td>14.7</td>
</tr>
<tr>
<td>TOOD+SF+SAHI+FI+PO</td>
<td><strong>20.4</strong></td>
<td><strong>14.9</strong></td>
<td><strong>23.5</strong></td>
<td><strong>17.6</strong></td>
</tr>
</tbody>
</table>

- SAHI+FI yielded up to 3.3% increase in large object AP compared to only SAHI.
- 25% overlap between slices increase the detection AP by up to 1.7%. 
Future work

- Other postprocessing techniques
- Slicing aided small instance segmentation
- Comparison with more models
- Slicing aided video object detection
Supporting Most Trending Detectors:

- YOLOX + SAHI demo: [HF Spaces](https://huggingface.co/spaces) (RECOMMENDED)
- YOLOv5 + SAHI walkthrough: [Open in Colab](https://colab.research.google.com)
- MMDetection + SAHI walkthrough: [Open in Colab](https://colab.research.google.com)
- Detectron2 + SAHI walkthrough: [Open in Colab](https://colab.research.google.com)
- HuggingFace + SAHI walkthrough: [Open in Colab](https://colab.research.google.com)
- TorchVision + SAHI walkthrough: [Open in Colab](https://colab.research.google.com)

More Detector Support In-progress:

- add YOLOX model support ✓ enhancement
  #557 opened 24 days ago by kadirnar • Review required
- add Yolov7 model support ✗ enhancement
  #544 opened 4 Aug by kadirnar • Approved
- refactor demo notebooks by utilizing newly improved documentation enhancement
  #516 opened 5 Jul by ishswori • Review required
- add Tensorflow Hub detector support ✓ enhancement
  #501 opened 19 Jun by kadirnar • Changes requested

Active Learning Based Synthetic Sample Selection for Endoscopic Image Classification

Alperen İnci, Ümit Mert Çağlar, Gökrem Polat, Öğuz Hanoğlu, Alptekin Temizel
Graduate School of Informatics, Middle East Technical University, Ankara, Turkey
Motivation and Problem Definition

- Ulcerative Colitis is a chronic inflammatory bowel disease.
- Assessment of the severity of the disease is crucial for physicians to administer appropriate treatment for UC disease.

Mayo 0 - healthy
Mayo 1 - mild disease
Mayo 2 - moderate disease
Mayo 3 - severe disease

Real colonoscopy images
Data Labelling Process

UC Mayo Annotator

Progress: 76.3%  Total images to annotate: 1468

Current image: G000774803.bmp

Your annotation
- [ ] Değerlendirmeye uygun değil
- [ ] Mayo 0
- [ ] Mayo 1
- [ ] Mayo 2
- [ ] Mayo 3

Annotate

Show annotations

OA: Mayo 3

YOA: Etiketlenmemiş!
Data Labelling Process

572 Patients
1043 Colonoscopies
19537 Images

Annotation

Not suitable to make an assessment (due to debris, artifacts vs.)

8060 Images

Labeled according to EMS (0-3)

564 Patients
11276 Images

Differently labeled by all three reviewers

201 Images

8261 images were removed from the dataset

Model Development (~85%)
479 Patients
9590 Images

Test Set (~15%)
85 Patients
1686 Images

10-fold Cross-Validation

Training Set (~76.5%)
~431 Patients
~8631 Images

Validation Set (~8.5%)
~48 Patients
~959 Images

Trained DNN Model

Inference on Test Set

Calculate Performance Metrics
Data Labelling by Subject Matter Experts

<table>
<thead>
<tr>
<th></th>
<th>Reviewer-1</th>
<th>Reviewer-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total images to evaluate</td>
<td>19537</td>
<td></td>
</tr>
<tr>
<td>Not suitable to assign a Mayo score</td>
<td>7621</td>
<td>9207</td>
</tr>
<tr>
<td>Mayo score is assigned</td>
<td>11916</td>
<td>10330</td>
</tr>
<tr>
<td>Mayo-0</td>
<td>7398</td>
<td>4503</td>
</tr>
<tr>
<td>Mayo-1</td>
<td>2473</td>
<td>3796</td>
</tr>
<tr>
<td>Mayo-2</td>
<td>1190</td>
<td>1014</td>
</tr>
<tr>
<td>Mayo-3</td>
<td>855</td>
<td>1017</td>
</tr>
</tbody>
</table>
Data Labelling by Subject Matter Experts

<table>
<thead>
<tr>
<th></th>
<th>Reviewer-3</th>
<th>From Reviewer 1&amp;2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total images to evaluate</td>
<td>7652</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Not suitable to assign a Mayo score</td>
<td>1895</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mayo score is assigned</td>
<td>5757</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>All reviewers annotate differently</td>
<td>201</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>To join the final dataset</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(agreement by two reviewers)</td>
<td>5556</td>
<td>5720</td>
<td>11276</td>
</tr>
<tr>
<td>Mayo-0</td>
<td>2633</td>
<td>3472</td>
<td>6105</td>
</tr>
<tr>
<td>Mayo-1</td>
<td>1842</td>
<td>1210</td>
<td>3052</td>
</tr>
<tr>
<td>Mayo-2</td>
<td>784</td>
<td>470</td>
<td>1254</td>
</tr>
<tr>
<td>Mayo-3</td>
<td>297</td>
<td>568</td>
<td>865</td>
</tr>
</tbody>
</table>
Data Labelling by Subject Matter Experts

Histogram of number of images per patient after annotation

Number of images per Mayo subscore

- Mayo 0: 54.14%
- Mayo 1: 27.07%
- Mayo 2: 11.12%
- Mayo 3: 7.67%
LIMUC Resources

Publication date: March 14, 2022
DOI: 10.5281/zenodo.5827695

Keyword(s):
- Ulcerative Colitis
- Inflammatory Bowel Disease
- Computer-Aided Diagnosis
- Deep Learning
- Colonoscopy
- Endoscopic Mayo Score

Published in:
Inflammatory Bowel Diseases: 2022 (11).

Related identifiers:
Supplement to 10.1093/ibd/izac225 (Journal article)

License (for files):
Creative Commons Attribution 4.0 International

---

patient_based_classified_images: Images of each patient are separated according to Mayo classes. If a train-val-test splitting is to be made according to the ratios desired by the user, this folder should be used.

train_and_validation_sets: Train and validation sets used in the paper. Using the scripts in dataset's GitHub repository, same 10-fold can be generated for replicating the results.

test_set: Test set used for performance measurement in the research paper. For a fair performance comparisons, this should be used to report performances.

---

Research Questions

• When there are limited number of labelled images, can we improve model performance by generating and adding synthetic samples?
• How can we best select the synthetic samples that would be the most useful in training?

Example synthetic colonoscopy images

Mayo 0- healthy  Mayo 1-mild disease  Mayo 2-moderate disease  Mayo 3-severe disease
Method: GAN Model Training

**StyleGAN2-ADA-PyTorch**
- Resolution 256x256
- Training length 5M images (initially 25M)
- Best model save at 200k images
- r1 Gamma=2 (best FID among 1,2,4,8)
- All augmentations
- ADA target 0.6
- Class Conditional GANs
- Class Specific GANs
GAN Model Training

Class Conditional GANs
- Employs class information
- One GAN for all classes
- Better FID on original dataset (imbalanced)
- No transfer learning, trained from scratch

Class Specific GANs
- A separate GAN for each class
- Worse FID on original dataset
- Can apply transfer learning (FFHQ)

<table>
<thead>
<tr>
<th>Training Method</th>
<th>Training set</th>
<th>Class-Conditional FID</th>
<th>Class-Specific GAN FID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mayo 0</td>
<td>Mayo 1</td>
</tr>
<tr>
<td>Subset 50</td>
<td>154.8</td>
<td>129.7</td>
<td>110.7</td>
</tr>
<tr>
<td>Subset 100</td>
<td>128.8</td>
<td>117</td>
<td>110.7</td>
</tr>
<tr>
<td>Subset 150</td>
<td>111.9</td>
<td>98.2</td>
<td>86.6</td>
</tr>
<tr>
<td>Subset 200</td>
<td>94.6</td>
<td>88.7</td>
<td>77.2</td>
</tr>
<tr>
<td>Subset 250</td>
<td>96.5</td>
<td>78</td>
<td>66.9</td>
</tr>
<tr>
<td>Subset 300</td>
<td>32.4</td>
<td>70.7</td>
<td>63.4</td>
</tr>
<tr>
<td>Subset 350</td>
<td>29.5</td>
<td>67.2</td>
<td>58.8</td>
</tr>
<tr>
<td>Subset 400</td>
<td>23.7</td>
<td>61.8</td>
<td>53.5</td>
</tr>
<tr>
<td>Subset 450</td>
<td>24.5</td>
<td>57.2</td>
<td>49.5</td>
</tr>
<tr>
<td>Subset 500</td>
<td>18.9</td>
<td>54.1</td>
<td>50.5</td>
</tr>
<tr>
<td>Original</td>
<td>8.5</td>
<td>15.1</td>
<td>21.5</td>
</tr>
</tbody>
</table>
Collective Dataset Creation

The truncation value controls the variance of generated samples.
- Truncation 0.5 -> samples are mostly around distribution center
- Truncation 2.0 -> samples are too diverse/unrealistic
- Truncation 1.2 -> trade-off between 2.0 and 0.5.
Results: Class-Specific GAN
System Architecture

- **Generation Style GAN2 - ADA**
- **Diversity Sampling (embedding space distance based)**
- **Uncertainty Sampling (Entropy, Margin)**
- **Active Learning Sampler**
- **Diversity Sampling (Coreset)**
- **Inference results on the large synthetic image set**

- **Synthetic images (180K samples)**
- **Original images**
- **Synthetic subset**

- **Neural Network Training and Inference**
System Architecture

- Original images
- Synthetic images (180K samples)
- Neural Network Training
- Generation
  - Style GAN2 - ADA
Diversity Sampling of Synthetic Images

Diversity Sampling (embedding space distance based)
Active Learning Based Sampling of Synthetic Images

- Entropy
- Coreset
- Margin
- Weighted Margin
Active Learning Based Sampling of Synthetic Images

- **Entropy**
  - Higher entropy indicates higher uncertainty - model is not confident about classification of the sample.

- **Coreset**
  - Aims to extract a diverse set of points with the maximum distance from others to represent the whole dataset.

- **Margin**

- **Weighted Margin**
Active Learning Based

- Entropy
- Coreset

Uncertainty-based active learning strategies frequently select similar samples since the trained model is likely to struggle to make decisions on almost identical samples. Therefore, uncertainty-based selection methods are prone to suffer from the overlapping problem.

- Margin
  ○ computes the difference between the top two class probabilities

- Weighted Margin
  ○ computes the uncertainty score by taking the power of Margin score with class distance
Results (50 Real Images Per Class)

Baseline QWK: 68.0
Results (50 Real Images Per Class)

Baseline F1: 54.3, Naïve Method F1: 55.8
Results (100 Real Images Per Class)

Baseline QWK: 74.6
Results (100 Real Images Per Class)

Baseline F1: 59.5, Naïve Method F1: 61.5
Conclusion

- Performance improvements can be achieved by using active learning methods.
- Comparative evaluations against random sample selection has to be done as it may outperform more sophisticated selection methods.
- Weighted Margin is the best approach according to the experimental results.