CVPR Supplementary document

Paper ID: 1228

November 2020

Implementation and training details not covered in the main manuscript are listed below.

1 Implementation details

DSMIL aggregator.
Size of the query vector $\mathbf{q} : 128 \times 1$.
Size of the information vector $\mathbf{v}$: same as the input instance feature vector size.
Normalization term for attentions: $\mathbf{d}_q^{-1/2}$, $\mathbf{d}_q$ is the dimension of $\mathbf{q}$.
Weight decay: 0.005. Betas for Adam optimizer $(0.5, 0.9)$.

SimCLR.
Size of the embedding vector $\mathbf{z} : 256 \times 1$.
Temperature parameter for the contrastive objective: 0.5.
Wegiht decay: 0.00001.
Backbone: ResNet18 (size of output feature vector: 512×1).

2 Training details

Stop criterion. Best accuracy, same accuracy then best AUC.
Epochs of training. DSMIL: 40. SimCLR: 30.

3 Code

We provide the code for cross validation on the MIL benchmark dataset and the code for training and testing on the precomputed TCGA Lung Cancer features dataset. GPU support is required to run the code.

3.1 Installation

Install anaconda/miniconda. Required packages:

```
$ conda env create --name dsmil --file env.yml
$ conda activate dsmil
```

3.2 Dataset preparation

The MIL benchmark dataset can be downloaded via:

```
$ python download.py --dataset=mil
```

This dataset requires 30MB of free disk space.
Precomputed features for TCGA dataset can be downloaded via:

```
$ python download.py --dataset=tcga
```

This dataset requires 20GB of free disk space.

3.3 Training and validation

To train DSMIL on standard MIL benchmark dataset:

```
$ python train_mil.py
```

To switch between MIL benchmark dataset, use option:

```
[--datasets]        # musk1, musk2, elephant, fox, tiger
```

Other options are available for learning rate (0.0002), cross validation fold (10),
weight-decay (5e-3), and number of epochs (40). To train DSMIL on TCGA
Lung Cancer dataset:

```
$ python train_tcga.py
```