NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

BS6207 ASSIGNMENT 4

LONG JINGYU
13/1/2022

WEAKLY SUPERVISED CLUSTERING BY EXPLOITING UNIQUE CLASS COUNT

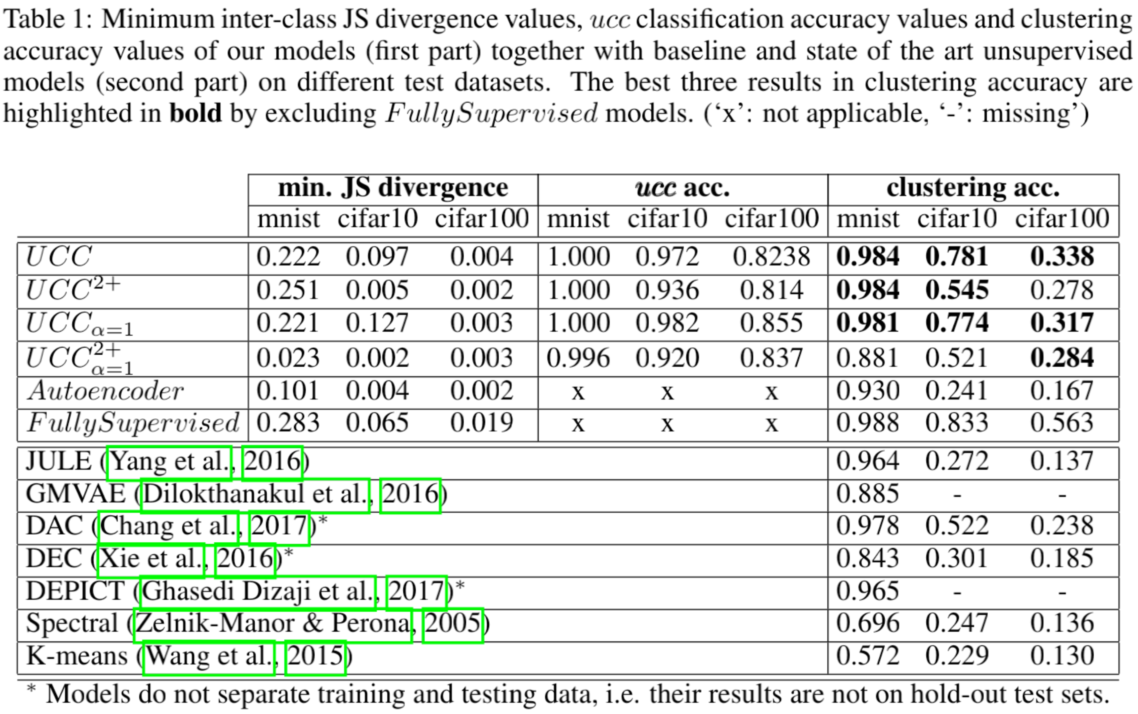
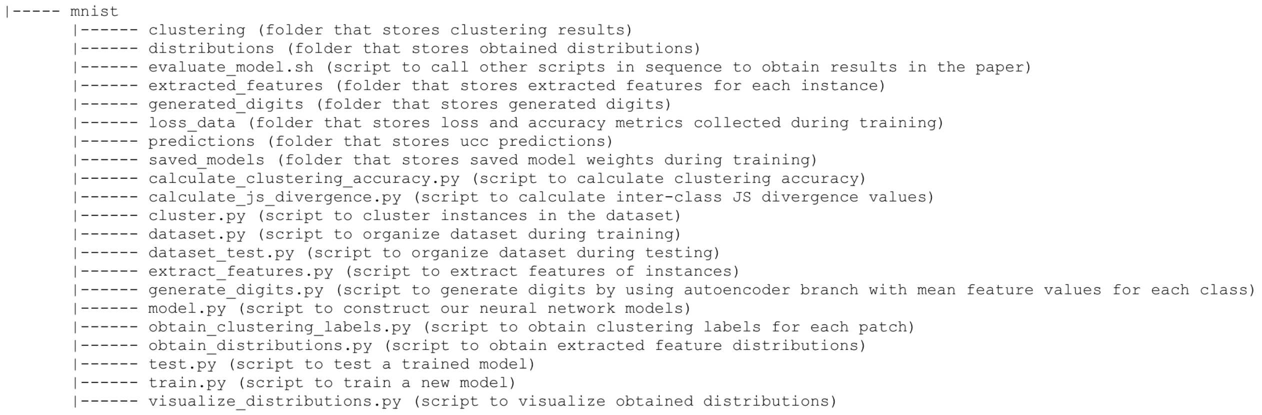


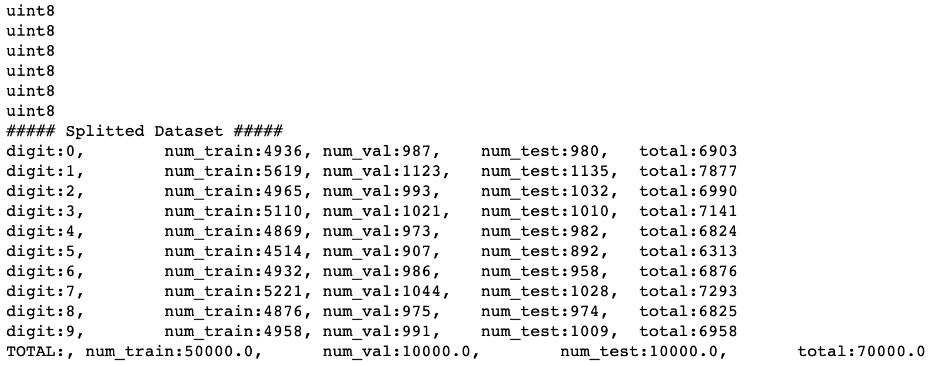
Table 1: Minimum inter-class JS divergence values, ucc classification accuracy values and clustering accuracy values of our models (first part) together with baseline and state of the art unsupervised models (second part) on different test datasets. The best three results in clustering accuracy are highlighted in **bold** by excluding $FullySupervised$ models. ('x': not applicable, '-': missing)

	min. JS divergence			***ucc* acc.**			**clustering acc.**		
	mnist	cifar10	cifar100	mnist	cifar10	cifar100	mnist	cifar10	cifar100
UCC	0.222	0.097	0.004	1.000	0.972	0.8238	**0.984**	**0.781**	**0.338**
UCC^{2+}	0.251	0.005	0.002	1.000	0.936	0.814	**0.984**	**0.545**	0.278
$UCC_{\alpha=1}$	0.221	0.127	0.003	1.000	0.982	0.855	**0.981**	**0.774**	**0.317**
$UCC^{2+}_{\alpha=1}$	0.023	0.002	0.003	0.996	0.920	0.837	0.881	0.521	**0.284**
$Autoencoder$	0.101	0.004	0.002	x	x	x	0.930	0.241	0.167
$FullySupervised$	0.283	0.065	0.019	x	x	x	0.988	0.833	0.563
JULE (Yang et al., 2016)							0.964	0.272	0.137
GMVAE (Dilokthanakul et al., 2016)							0.885	-	-
DAC (Chang et al., 2017)*							0.978	0.522	0.238
DEC (Xie et al., 2016)*							0.843	0.301	0.185
DEPICT (Ghasedi Dizaji et al., 2017)*							0.965	-	-
Spectral (Zelnik-Manor & Perona, 2005)							0.696	0.247	0.136
K-means (Wang et al., 2015)							0.572	0.229	0.130

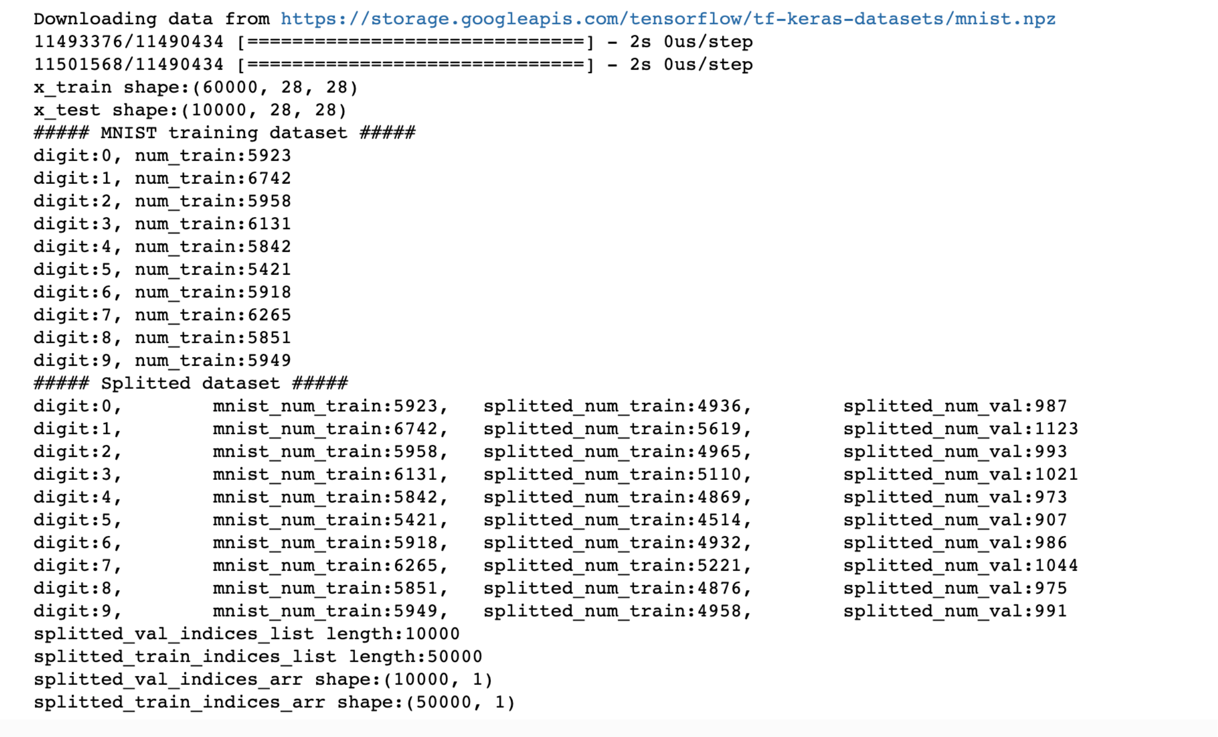
* Models do not separate training and testing data, i.e. their results are not on hold-out test sets.



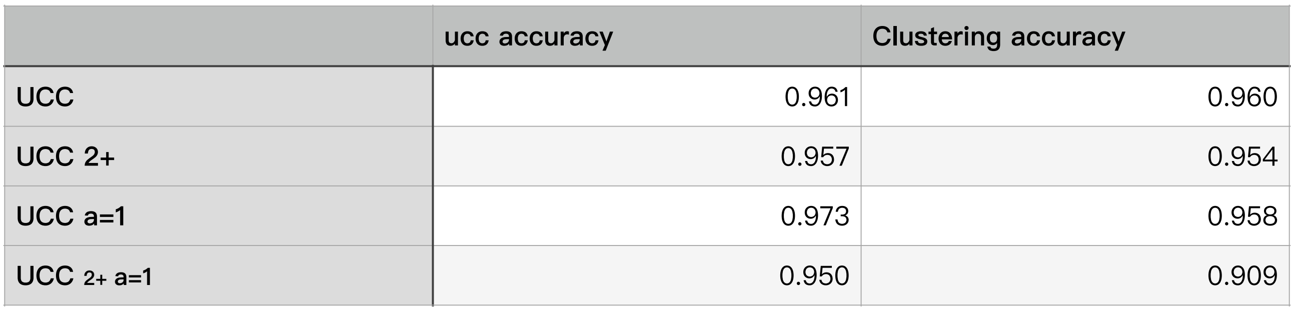
```
|----- mnist
       |------ clustering (folder that stores clustering results)
       |------ distributions (folder that stores obtained distributions)
       |------ evaluate_model.sh (script to call other scripts in sequence to obtain results in the paper)
       |------ extracted_features (folder that stores extracted features for each instance)
       |------ generated_digits (folder that stores generated digits)
       |------ loss_data (folder that stores loss and accuracy metrics collected during training)
       |------ predictions (folder that stores ucc predictions)
       |------ saved_models (folder that stores saved model weights during training)
       |------ calculate_clustering_accuracy.py (script to calculate clustering accuracy)
       |------ calculate_js_divergence.py (script to calculate inter-class JS divergence values)
       |------ cluster.py (script to cluster instances in the dataset)
       |------ dataset.py (script to organize dataset during training)
       |------ dataset_test.py (script to organize dataset during testing)
       |------ extract_features.py (script to extract features of instances)
       |------ generate_digits.py (script to generate digits by using autoencoder branch with mean feature values for each class)
       |------ model.py (script to construct our neural network models)
       |------ obtain_clustering_labels.py (script to obtain clustering labels for each patch)
       |------ obtain_distributions.py (script to obtain extracted feature distributions)
       |------ test.py (script to test a trained model)
       |------ train.py (script to train a new model)
       |------ visualize_distributions.py (script to visualize obtained distributions)
```



```
uint8
uint8
uint8
uint8
uint8
uint8
##### Splitted Dataset #####
digit:0,        num_train:4936, num_val:987,    num_test:980,   total:6903
digit:1,        num_train:5619, num_val:1123,   num_test:1135,  total:7877
digit:2,        num_train:4965, num_val:993,    num_test:1032,  total:6990
digit:3,        num_train:5110, num_val:1021,   num_test:1010,  total:7141
digit:4,        num_train:4869, num_val:973,    num_test:982,   total:6824
digit:5,        num_train:4514, num_val:907,    num_test:892,   total:6313
digit:6,        num_train:4932, num_val:986,    num_test:958,   total:6876
digit:7,        num_train:5221, num_val:1044,   num_test:1028,  total:7293
digit:8,        num_train:4876, num_val:975,    num_test:974,   total:6825
digit:9,        num_train:4958, num_val:991,    num_test:1009,  total:6958
TOTAL:, num_train:50000.0,      num_val:10000.0,        num_test:10000.0,       total:70000.0
```



```
Downloading data from https://storage.googleapis.com/tensorflow/tf-keras-datasets/mnist.npz
11493376/11490434 [==============================] - 2s 0us/step
11501568/11490434 [==============================] - 2s 0us/step
x_train shape:(60000, 28, 28)
x_test shape:(10000, 28, 28)
##### MNIST training dataset #####
digit:0, num_train:5923
digit:1, num_train:6742
digit:2, num_train:5958
digit:3, num_train:6131
digit:4, num_train:5842
digit:5, num_train:5421
digit:6, num_train:5918
digit:7, num_train:6265
digit:8, num_train:5851
digit:9, num_train:5949
##### Splitted dataset #####
digit:0,        mnist_num_train:5923,   splitted_num_train:4936,        splitted_num_val:987
digit:1,        mnist_num_train:6742,   splitted_num_train:5619,        splitted_num_val:1123
digit:2,        mnist_num_train:5958,   splitted_num_train:4965,        splitted_num_val:993
digit:3,        mnist_num_train:6131,   splitted_num_train:5110,        splitted_num_val:1021
digit:4,        mnist_num_train:5842,   splitted_num_train:4869,        splitted_num_val:973
digit:5,        mnist_num_train:5421,   splitted_num_train:4514,        splitted_num_val:907
digit:6,        mnist_num_train:5918,   splitted_num_train:4932,        splitted_num_val:986
digit:7,        mnist_num_train:6265,   splitted_num_train:5221,        splitted_num_val:1044
digit:8,        mnist_num_train:5851,   splitted_num_train:4876,        splitted_num_val:975
digit:9,        mnist_num_train:5949,   splitted_num_train:4958,        splitted_num_val:991
splitted_val_indices_list length:10000
splitted_train_indices_list length:50000
splitted_val_indices_arr shape:(10000, 1)
splitted_train_indices_arr shape:(50000, 1)
```



	ucc accuracy	Clustering accuracy
UCC	0.961	0.960
UCC 2+	0.957	0.954
UCC a=1	0.973	0.958
UCC 2+ a=1	0.950	0.909

We defined ucc as a bag level label in MIL setup and mathematically proved that a perfect ucc classifier can be used to perfectly cluster individual instances inside the bags. We designed a neural network based ucc classifer and experimentally showed that clustering performance of our framework with our ucc classifiers are better than the performance of unsupervised models and comparable to performance of fully supervised learning models.
In the future, we want to check the performance of our UCCsegment model with other medical image datasets and use it to discover new morphological patterns in cancer that had been overlooked in traditional pathology workflow.