2022

Unsupervised machine learning via transfer learning and k-means clustering to classify materials image data

SHWETA UMBARJE

Background

Machine learning is being used to solve materials science problems, but there are opportunities for doing much more unsupervised learning. This is especially helpful for extracting knowledge from unlabelled data sets and achieving the best possible machine learning performance. However, since unsupervised methods are not as commonly used in the background research, the best practices and pitfalls of using them are often overlooked.

Objective

This article- will outline how to build, use, and evaluate a high-performance unsupervised machine learning approach to a standard materials computer vision problem.

Methods

The analysis is split into 3 parts:

1. Pre-processing: Prepare the data to be read by the CNN.

2. Feature extraction (encoding): Use the CNN to generate a numerical representation of each image.

3. Clustering (decoding): Assign a label to the image, grouping images with similar features together.

Results

1. Standard Analysis

The standard analysis consists of the following steps in order:

1) Contrast limited adaptive histogram equalization is applied to each image.
2) Each image is resized to 224x224 pixels.
3) Each image is passed through VGG16 trained with ImageNet weights, and the outputs of the FC1 layer are saved as feature representations.
4) The features are transformed using PCA with 50 whitened components, which preserve 73.6% of the total variance.
5) Features are clustered via k-means clustering with 7 clusters, k-means++ initialization, and 500 different initialization steps.
6) The clustering with the lowest total inertia (not necessarily the greatest accuracy) is used as the final result.

1.1. Results with whitening

	Author's Result	Actual Result
Accuracy	99.6%	97.4%
Random initialization Accuracy	10 runs: avg: 0.9940 std: 0.001625, min: 0.9906	10 runs: avg: 0.9863 std: 0.007278, min: 0.9733

- *Classification report*

	precision		recall		f1-score	
Cr	0.993	0.990	1.000	1.000	0.997	0.995
In	0.987	0.890	0.993	0.973	0.990	0.995
Pa	1.000	1.000	0.993	0.993	0.997	0.997
PS	0.993	0.974	1.000	0.997	0.997	0.985
RS	1.000	1.000	1.000	1.000	1.000	1.000
Sc	1.000	1.000	0.987	0.880	0.993	0.936

1.2. Results without whitening

	Author's Result	Actual Result
Accuracy	99%	96.3%

- *Classification Report*

	precision		recall		f1-score	
Cr	0.961	0.955	0.997	0.997	0.979	0.976
In	0.926	0.942	0.913	0.917	0.919	0.929
Pa	0.989	0.993	0.930	0.927	0.959	0.959
PS	0.926	0.924	0.960	0.967	0.943	0.945
RS	0.971	0.974	0.997	1.000	0.984	0.987
Sc	0.993	0.997	0.967	0.973	0.980	0.985

2. Sensitivity Analysis- No Histogram Equalization

This evaluates the clustering performance of the standard analysis but replaces the features with the fc1 feature descriptors computed on images without histogram equalization

2.1. Clustering

	Author's Result	**Actual Result**
Accuracy	94.6%	90.2%

- *Classification Report*

	precision		**recall**	
Cr	0.997	0.987	0.997	1.000
In	0.824	0.844	0.890	0.957
Pa	0.990	0.990	0.997	0.993
PS	0.900	0.747	0.987	0.957
RS	0.993	0.919	1.000	0.720
Sc	1.000	1.000	0.810	0.787

3. Sensitivity study- using different layers than fc1

The accuracy of image classification depends on the selection of the CNN output layer selected as the feature descriptor, and it has been observed that different image types are best represented by different layers.

3.1. fc2 features

The fc2 fully-connected layer has the same size as the fc1 layer, generating 4096-dimensional features for each image.

	Author's Result	**Actual Result**
Accuracy	99.4%	99.3%

3.2. block5_pool features

	Author's Result	**Actual Result**
Accuracy	97.3%	97.3%

3.3. block5 conv3 features

	Author's Result	**Actual Result**
Accuracy	51%	60.9%

4. Sensitivity study: PCA

The standard analysis uses PCA with 50 components followed by whitening to compute the final feature representation of the images. To determine how classification performance changes with the number of components used, the analysis was conducted while varying the number of components, using both whitened and unwhitened PCA components.

No. of Components	**Author's Result**		**Actual Result**	
	accuracies (no whitening)	*accuracies (white ning)*	*accuracies (no whit ening)*	*accuracies (wh itening)*
1	0.460	0.460	0.457	0.456
5	0.861	0.878	0.888	0.855
10	0.954	0.972	0.954	0.970
20	0.956	0.983	0.957	0.983
50	0.961	0.996	0.963	0.976
100	0.961	0.852	0.964	0.839
250	0.961	0.547	0.964	0.684
500	0.961	0.297	0.963	0.224
1000	0.961	0.170	0.963	0.170
1800	0.961	0.191	0.963	0.196

5. Sensitivity study: K-means

This study mainly focused on 2 parameters,

- Number of initializations
- Number of cluster centres

The main points are as follows:

- Without whitening, there are a relatively small number of final cluster positions. Most of the time, even with a single initialization step, the model reaches 96% accuracy.
- With whitening, there are a lot of local minima that the centroids can become trapped in. On average the model archives 80% accuracy for a single run. However, accuracy tends to increase as the inertia decreases. Thus, running many initializations and taking the result with the lowest inertia allows us to consistently reach 99+% accuracy.

	Author’s Result		Actual Result	
	No Whitening	With Whitening	No Whitening	With Whitening
Min Accuracy	0.609	0.362	0.763	0.431
Max Accuracy	0.964	0.996	0.966	0.994

6. Predictive Model

This model is used to see how well the performance generalizes to new data.

6.1. K-fold cross-validation

The data is split into k-folds, 4 folds are used for training, and the remaining fold is held out for testing.

	Author’s Result		**Actual Result**	
PCA 35 components + whitening	*Train Accuracy*	*Validation Accuracy*	*Train Accuracy*	*Validation Accuracy*
	0.9882	0.9833	0.9938	0.9889
	0.9896	0.9972	0.9889	0.9972
	0.9917	0.9944	0.9882	0.9694
	0.9917	0.9944	0.9889	0.9944
	0.9931	0.9917	0.9868	0.9889

Conclusions

This approach outlined in this study achieved 99.4% which is the same we have got after the execution of each step. A sensitivity analysis was conducted to demonstrate the impact of each step in the analysis on the results. Histogram equalization improves classification performance by reducing differences between images with the same defects but different brightness profiles. Using the outputs of the fully connected layers in VGG16 maximize the signal-to-noise ratio of the feature descriptors, resulting in a useful and relatively compact feature description of each image. Clustering whitened PCA components results in the maximum classification performance but also increases the variance in performance of individual trials. Thus, to maximize the classification accuracy, k-means was run with many different initialization steps, and the model with the lowest total inertia was used.