Datensatzanalyse mithilfe lernender Verfahren

Die Auswahl geeigneter Datensätze für das Training von Machine-Learning-Modellen stellt eine Herausforderung dar. Der Trainingsprozess verbraucht erhebliche Zeit- und Rechenressourcen, insbesondere bei komplexen Modellen, die für gute Leistungen benötigt werden. Die Qualität und Eignung der Trainingsdaten sind dabei entscheidend für die Leistung des Modells. Während Realdaten generell bevorzugt werden, haben sich auch synthetische Daten als nützliche Alternative erwiesen, die mithilfe von Scene Engineering und Domain Randomization generiert werden. Diese Techniken ermöglichen es, viele Variationen in Beleuchtung, Kamerapositionen und Texturen erzeugen. Dabei kann jede Variation die Leistung des Modells beeinflussen, selbst wenn die Bilddaten für das menschliche Auge sehr ähnlich erscheinen. Es besteht oft Unsicherheit darüber, welcher Datensatz die Anforderungen am besten erfüllt. Daher wäre wünschenswert, ein System zu haben, das die Vor-Auswahl aus einer Vielzahl von Datensätze erleichtert, ohne ein Modell auf alle einzeln trainieren und evaluieren zu müssen. Für die Analyse und Bewertung der Datensätze wird ein lernendes Verfahren zur Merkmalsextraktion und Dimensionsreduktion benötigt. Hierfür wurde ein Residual Convolutional Autoencoder (R-CAE) entwickelt, der auf ausgewählten Datensätzen trainiert wurde, die die Vielfalt der zu vergleichenden Datensätze abdecken. Der Encoder-Part wurde extrahiert, um Datensatz-Kodierungen zu erzeugen und die Latente Darstellung dieser zu untersuchen. Es wurden verschiedene R-CAE-Modelle evaluiert, um die optimale Dimensionsgröße zu ermitteln, die eine starke Kompression ermöglicht, aber dennoch die individuellen Charakteristiken der Datensätze bewahrt. Sowohl die Analyse des Latenten Raums als auch die Evaluierungsmethoden zeigten, dass die wesentlichen Merkmale nach der Dimensionsreduktion erhalten bleiben, trotz eines Verlusts in der Bildqualität. Die Datensatz-Kodierungen wurden mittels einer Aggregation zusammengefasst, um für jeden einen repräsentativen Feature-Vektor zu erzeugen, der den Datensatz kompakt in wenigen Werten repräsentiert. Konkret wurden 17 Datensätze der Dimensionalität 393.216 auf jeweils 3.072 Werte reduziert, was eine Reduktion von 99,22% darstellt. Die Repräsentanten wurden einem distanzbasierten Auswahlverfahren unterzogen. Dabei wurden verschiedene Bewertungskriterien implementiert, mit denen die Datensätze hinsichtlich ihrer analysierten Domänen-Eigenschaften, ihrer Gruppen-Zugehörigkeit und Real-Nähe in ein ordinales Ranking gebracht werden können, abhängig der individuellen Anwendung. Um die Leistung des Performanz-Schätzungssystems zu bewerten, standen Referenzwerte einer Objekterkennungsaufgabe zur Verfügung, die mit den Abschätzungen des Systems abgeglichen wurde. Dabei zeigt das Auswahlverfahren eine Übereinstimmung von 7 von 10 im ordinalen Ranking. Allerdings ergab die Validierung mittels anderer Encoder-Modelle, dass die resultierenden Datensatz-Kodierungen und Distanzmessungen keine aussagekräftigen Ergebnisse liefern, da die Daten trotz der starken Dimensionsreduktion noch sehr hochdimensional sind und daher dem Fluch der Dimensionalität unterliegen.

Dataset Analysis Using Machine Learning Techniques

Selecting appropriate datasets for training machine learning models poses a challenge. The training process consumes significant time and computational resources, especially for complex models aimed at achieving high performance. The quality and suitability of the training data are crucial for the model's performance. While real data is generally preferred, synthetic data generated through scene engineering and domain randomization have also proven to be useful alternatives. These techniques allow for the generation of variations in lighting, camera positions, and texture, each of which can influence the model's performance, even though the image data may seem visually indistinguishable to the human eye. There is often uncertainty about which dataset best suits the requirements. Hence, it would be advantageous to have a system that simplifies the choice of from range of available datasets, eliminating the need to train and evaluate a model on each one separately. For the analysis and evaluation of datasets, a learning method for feature extraction and dimension reduction is required. To this end, a Residual Convolutional Autoencoder (R-CAE) was developed and trained on selected datasets that encompass the diversity of the datasets to be compared. The encoder part was extracted to generate dataset encodings and to examine their latent representations. Various R-CAE models were evaluated to determine the optimal dimension size that allows for substantial compression while still preserving the individual characteristics of the datasets. Both the analysis of the latent space and the evaluation methods showed that the essential features remain intact after dimension reduction, despite a loss in image quality. The dataset encodings were aggregated to generate a representative feature vector for each dataset, concisely representing the dataset in a few values. Specifically, 17 datasets with an original dimensionality of 393,216 were reduced to 3,072 values each, which represents a reduction of 99.22%. The representatives were subjected to a distance-based selection process. Various evaluation criteria were implemented, allowing the datasets to be ranked ordinally based on their analysed domain properties, group affiliations, and proximity to real data, depending on the individual application. To evaluate the performance estimation system, reference values from an object recognition task were available, which were matched against the system's estimates. The selection process showed a match of 7 out of 10 in the ordinal ranking. However, validation using other encoder models indicated that the resulting dataset encodings and distance measurements do not provide meaningful results, as the data remains highly dimensional despite significant dimension reduction, thus still subject to the curse of dimensionality.