Paper: Siamese Neural Networks for One-shot Image Recognition

Summary

In this paper, the authors explore a method for learning Siamese neural networks that employ a unique structure to naturally rank similarities between inputs. This is a one-shot learning setting in which the model learns to correctly make predictions given only a single example of each new class. In general, the system which incorporates one-shot learning by developing domain-specific features tend to excel at similar instances but fail to offer robust solutions that may be applied to other types of problems. The authors in this paper aim to deal with that issue by presenting a novel approach that limits assumptions on the structure of the inputs while automatically acquiring features that enable the model to generalize successfully from few examples. To deal with the invariance in transformation in the input space, they use many layers of non-linearities.

The authors use a supervised metric-based approach with Siamese neural network to learn the representation. After that, they reuse the network's feature for one-shot learning without retraining. They employ large Siamese CNN which are capable of learning generic image features and are easily trained using standard optimization techniques on pairs sampled from the source data. Their network also provides a competitive approach that does not rely upon domain-specific knowledge by instead exploiting deep learning techniques. To develop a model for on-shot image classification, they first developed a neural network that can discriminate between the class-identity of image pairs, which is called the verification task for image recognition. The verification model learns to identify input pairs according to the probability that they belong to the same class or different classes. The justification is, if the features learned by the verification model are sufficient to confirm or deny the identity of characters from one set of alphabets, they ought to be sufficient for other alphabets, provided that the model has been exposed to a variety of alphabets to encourage variance amongst the learned features.

The authors' standard model is a Siamese CNN which consists of twin networks and accepts distinct inputs but is joined by an energy function at the top. The parameters between the twin networks are tied. The authors in this paper use the weighted L1 distance between the twin feature vector combined with sigmoid activation. They use the backpropagation algorithm where the gradient is additive across the twin networks due to the tied weights. They also augmented the training set with small affine distortions.

They trained their model on a subset of the Omniglot dataset. They split the dataset into 40 alphabet background set and 10 alphabet evaluation set. The training, validation, and test sets can be generated from the background set to tune models for verification. The background set is used for developing a model by learning hyperparameters and feature mappings. Conversely, the evaluation set is used only to measure the one-shot classification performance.

To train their verification model, they put together 3 different data set sizes with 30,000, 90,000, and 150,000. With affine distortions, they also produced an additional copy of the data set. After optimizing a Siamese network to master the verification task, they demonstrated the potential of their model at one-shot learning. They achieved a score of 92% which is stronger than any other model except HBPL, which is a state of art in the field. However, the authors claim that their model did not include any extra prior knowledge about characters or strokes such as generative information about the drawing process.

They also tested how well the model trained on the Omniglot dataset can generalize to the MNIST dataset. Their model did not do an outstanding job, however, they claim that they were still able to achieve reasonable generalization from the features learned on Omniglot without training at all on MNIST.

Strengths

- I liked that they had a whole section dedicated to the Deep Siamese Network. That helps the reader with a little less background in the field understand the paper well.
- They have done a good job of explaining their model architecture and parameters.

Weaknesses

- Some of their figures were confusing, for example, Figure 7. They did not have a good enough description of this figure in the main text as well.

Confusions

- What do they mean by the "supervised metric-based approach"?
- How does choosing the number of Convolutional filters as multiple of 16 help optimize performance?
- What do they mean when they say, "gradient is additive across the twin networks due to the tied weights"?

Discussions

- How has the Siamese model used in industry today?
- How does this model handle the variability in the input space? For example, if the two images of the same person are very different in intensity, scale, angle, etc.
- Internally, how does the "verification model" work? I want to understand more about their justification that "if the features learned by the verification model are sufficient to confirm or deny the identity of characters from one set of alphabets, they ought to be sufficient for other alphabets".