

2. Convolutional Neural Networks

a. How large are the CIFAR10 images? What does the 10 refer to?

Each picture is 32×32 pixels with three color channels, and the 10 refers to the dataset's ten categories representing every single digit.

b. What is a MaxPooling2D layer? What's it do?

The MaxPooling2D layer slides a small window over each feature map and keeps the maximum value, downsampling while preserving strong features.

c. What's Adam?

Adam is a gradient-based optimizer that automatically adjusts learning rates.

d. What's the softmax function do?

Softmax turns a list of raw scores into probabilities that sum to one, letting the model pick the most likely class.

e. What is CategoricalCrossEntropy? What do we use it for?

CategoricalCrossEntropy is a loss function that measures how far the predicted probabilities are from the correct one-hot label and guides the model to improve.

f. What does the Flatten layer do?

Flatten flattens the 3-D output of the conv layers into one long 1-D vector

g. What does the Dense layer do?

A Dense layer multiplies the input vector by a weight matrix, adds a bias, and then applies an activation to create learned features.

h. Why does the height and width get smaller for each convolutional layer?

Each convolutional step aggregates data over small windows, so the spatial height and width decrease with each layer.

i. Try changing the number of epochs to 20. What happens to the error for the training set? What about for the validation set? Why is that?

Training loss keeps falling, but validation loss levels off and then rises after about eight epochs, indicating overfitting of the training data.

4. BERT Tutorial

a. What does this tutorial do?

The tutorial fine-tunes a pre-trained BERT-base model to detect spam in SMS messages.

b. BERT is a contextual model. What does this mean? How is it different from the bag-of-words approach?

A contextual model like BERT builds a word's meaning from the words around it, while bag-of-words just counts words and ignores order and context.

c. What are the three embeddings BERT uses?

BERT adds together token embeddings, segmentation embeddings, and position embeddings for each input token.

d. What is Masked Language Modeling?

In Masked Language Modeling, some words are hidden, and the model learns to predict the masked words from the rest of the sentence.

e. What is Next Sentence Prediction? = What is the final performance of BERT on this dataset?

In Next Sentence Prediction, the model learns to judge whether a candidate second sentence actually follows the first one.

f. Increase to 20 epochs. How does performance change?

Before the increase, the 10-epoch model reached about 84 percent accuracy, compared with about 92 percent accuracy after 20 epochs.