Data Augmentation

Data augmentation is a technique of artificially increasing the training set by creating modified copies of a dataset using existing data. It includes making minor changes to the dataset or using deep learning to generate new data points.

Note:

If we does not have required data, and we have to work on same data so we artificially generated data using data augmentation. Because if data is not larger in volume over-fitting error arise.

It's the strategy to increasing the training size of the data artificially.

Augmented vs. Synthetic data

Augmented data is driven from original data with some minor changes.

In the case of image augmentation, we make geometric and color space transformations (flipping, resizing, cropping, brightness, contrast) to increase the size and diversity of the training set.

Synthetic data is generated artificially without using the original dataset.

It often uses DNNs (Deep Neural Networks) and GANs (Generative Adversarial Networks) to generate synthetic data.

Note: Augmented data is made by using minor changes in data, while synthetic data create own data i.e. fake data from information.

The augmentation techniques are not limited to images. You can augment audio, video, text, and other types of data too.

When Should You Use Data Augmentation?

- To prevent models from overfitting.
- The initial training set is too small.
- To improve the model accuracy.
- To reduce the operational cost of labeling and cleaning the raw dataset.

Note: Reasons for overfitting

- If model A data is larger than model B data. E.g. for cat dog classifications, cat images are larger than dog images it create biasness that cause overfitting
- If data is very small in volume, overfiting cause. Because less data, less information.

Limitations of Data Augmentation

- The biases in the original dataset persist in the augmented data.
- Quality assurance for data augmentation is expensive.
- Research and development are required to build a system with advanced applications. For example, generating high-resolution images using GANs can be challenging.
- Finding an effective data augmentation approach can be challenging.

Data Augmentation Techniques

In this section, we will learn about audio, text, image, and advanced data augmentation techniques.

Audio Data Augmentation

Noise injection: add gaussian or random noise to the audio dataset to improve the model performance.

Shifting: shift audio left (fast forward) or right with random seconds.

Changing the speed: stretches times series by a fixed rate.

Changing the pitch: randomly change the pitch of the audio.

Text Data Augmentation

Word or sentence shuffling: randomly changing the position of a word or sentence.

Word replacement: replace words with synonyms.

Syntax-tree manipulation: paraphrase the sentence using the same word.

Random word insertion: inserts words at random.

Random word deletion: deletes words at random.

Image Augmentation

Geometric transformations: randomly flip, crop, rotate, stretch, and zoom images. You need to be careful about applying multiple transformations on the same images, as this can reduce model performance.

Color space transformations: randomly change RGB color channels, contrast, and brightness.

Kernel filters: randomly change the sharpness or blurring of the image.

Random erasing: delete some part of the initial image.

Mixing images: blending and mixing multiple images.

Advanced Techniques

Generative adversarial networks (GANs): used to generate new data points or images. It does not require existing data to generate synthetic data.

Neural Style Transfer: a series of convolutional layers trained to deconstruct images and separate context and style.

Data Augmentation Applications

Healthcare: Using geometric and other transformations can help you train robust and accurate machine-learning models.

Self-Driving Cars: It can help you train and test machine learning applications where data security is an issue.

Natural Language Processing: You can apply synonym augmentation, word embedding, character swap, and random insertion and deletion. These techniques are also valuable for low-resource languages.

Automatic Speech Recognition: It improves the model performance even on low-resource languages.