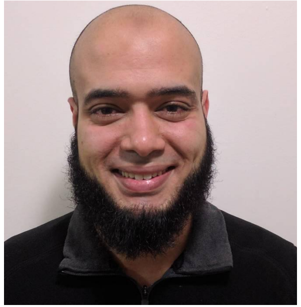
Machine Learning

Augmented Data

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)

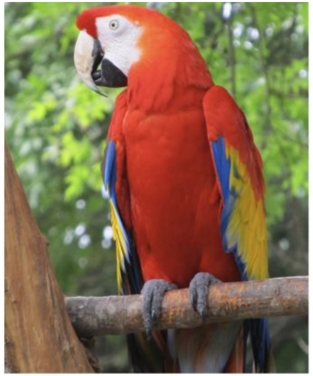


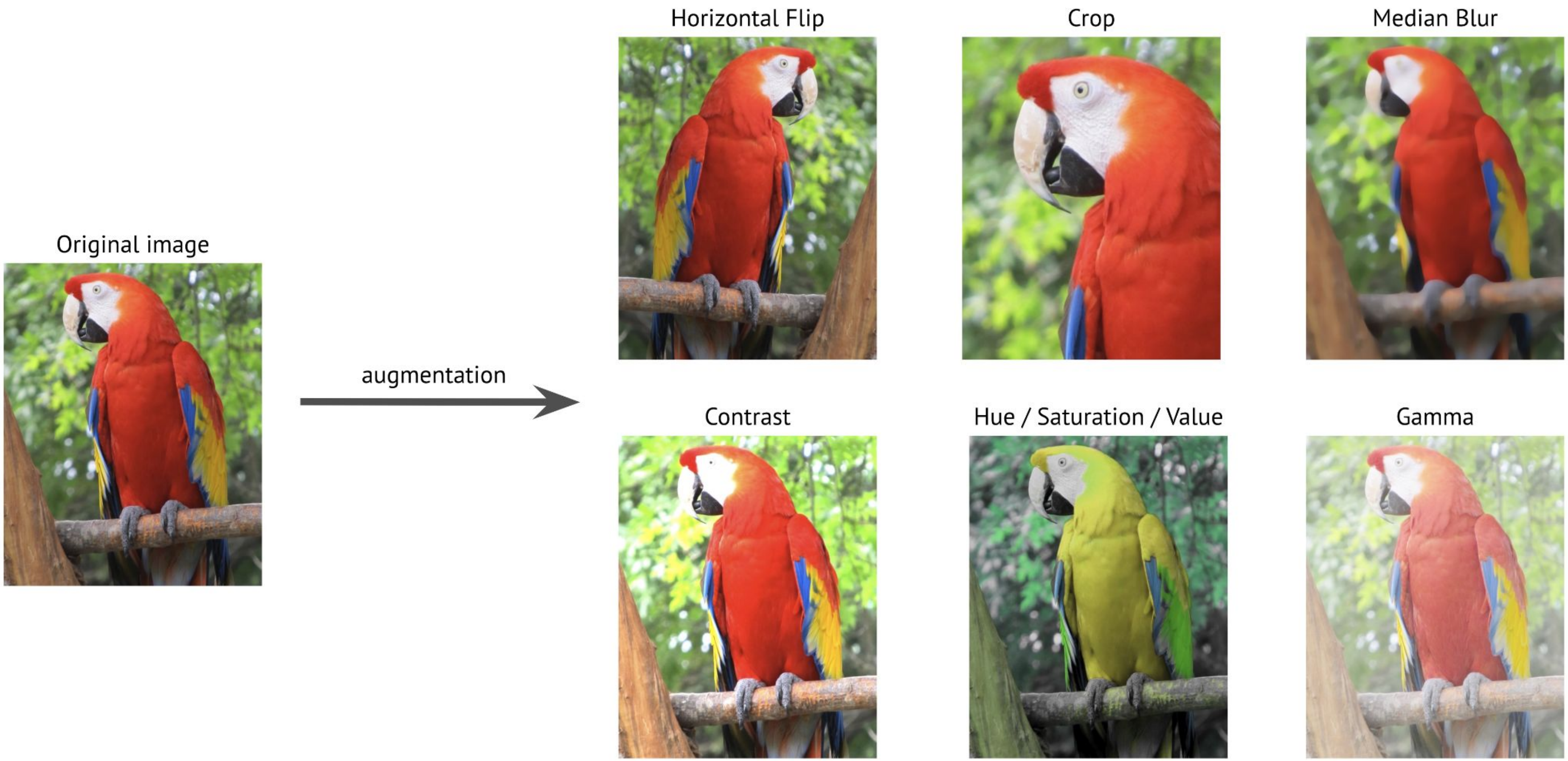
Augmented Data

- Data augmentation: Augmentation techniques can be used to **artificially increase the diversity** and quantity of the training data
- The idea behind data augmentation is to introduce **variations** in the training data that mimic real-world scenarios and increase the diversity of the examples seen by the model. By exposing the model to a broader range of data instances, it can learn more **robust and generalized** patterns.
- Data augmentations is created from **EXISTING** dataset already
- Basic **augmentations** techniques were used almost in **all papers** that describe the state-of-the-art models for image recognition

Question!

- Given an image, think all ways we can augment the dataset with it
- An example, **flip** the image!

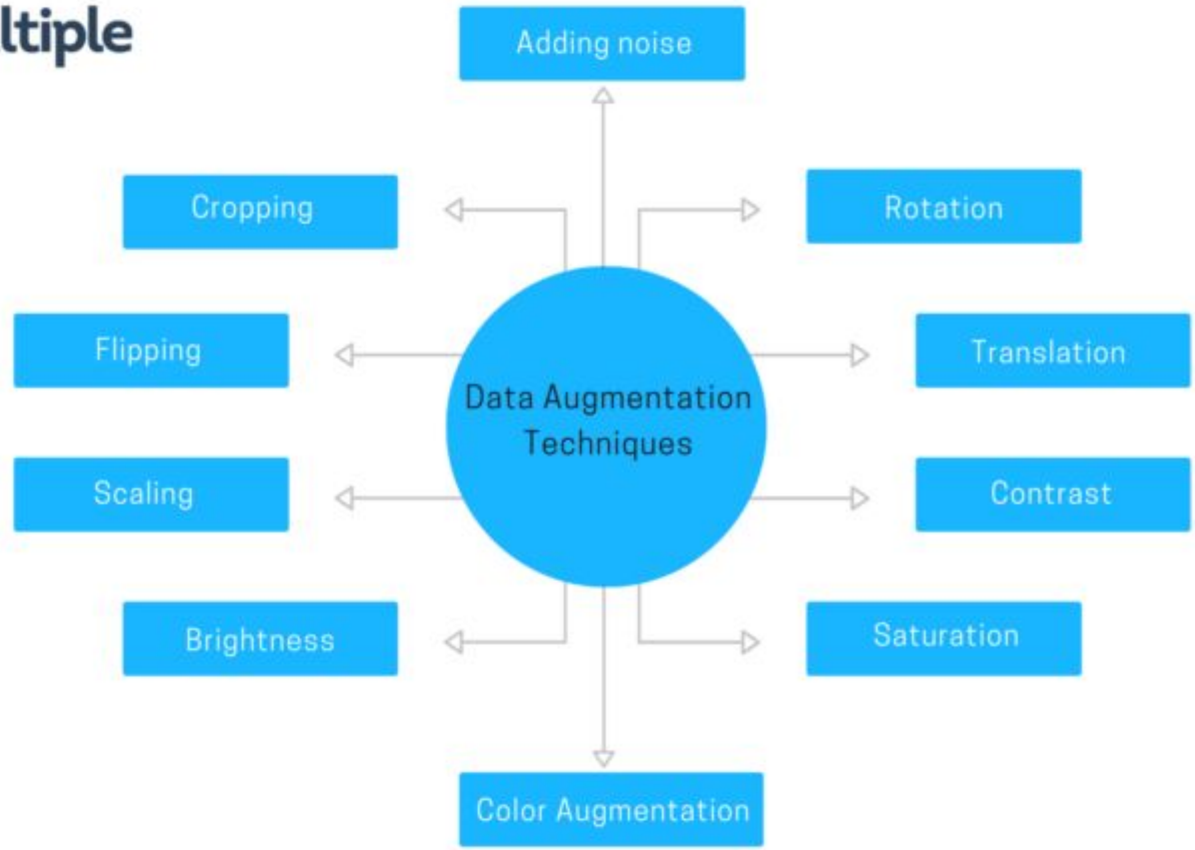



Original image
augmentation
Horizontal Flip
Crop
Median Blur
Contrast
Hue / Saturation / Value
Gamma

Img src

AIMultiple

- **Geometric transformations**
  - rotation, translation, scaling, and flipping
- Noise injection
- Cropping and resizing
- Color and contrast variations



Data Augmentation Techniques

Adding noise

Rotation

Translation

Contrast

Saturation

Color Augmentation

Brightness

Scaling

Flipping

Cropping

Invariant Models

- In deep learning, if you want your model to stay consistent regardless a specific property, just prepare wide diversity of the samples
- For example, assume you want a scale-invariant model for hand detection
  - Then augment the data with many samples of different **scale** of the same object
- For example, assume you want a rotation-invariant model for hand detection
  - Then augment the data with many samples of different **rotations** of the same object

NLP Augmentations

- Synonym Replacement
- Backtranslation (Start English, Translate to Germany, Translate to English)
- Random Insertion, Deletion, and Swap words or phrases
- Masking and Dropout words to simulate missing or unknown words
- Character-level Augmentation
- Sentence Reordering

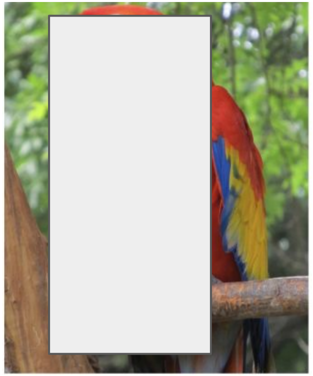
Audio Augmentations

- Time Stretching
- Pitch Shifting
- Background Noise Injection
- Reverberation
- Time and Amplitude Masking
- SpecAugment

Tabular Augmentation

- SMOTE extends the dataset with synthetic examples as follows:
  - Goal: interpolate between neighbour instances
- Select a random example (A) from data
- Find k-neighbours of A
- Pick a random neighbour (B) out of the k neighbours
- Pick a random point C on the line segment A==B
  - C ~= A + (B - A) * t [t is random value in range [0-1]
- Tip: oversampling must be done only on the train set to avoid data leakage
  - Split to train/val/split, then oversample the train

Misc



- Make sure your augmentations don't corrupt the data!
  - A lot of added noise or clear data mismatch
  - For example, you introduced big white crops in your image
- A lot of augmented examples implies more computational **cost**!
- Can augmentation hurt performance?
  - In deep learning, typically no.
    - In computer vision, augmentation is a typical step
  - In tabular data, there is a chance your augmentations alter the data distribution
    - Non-complex models can easily go in the wrong direction

Relevant Materials

- Paper: AutoAugment: Learning Augmentation Policies from Data
- Imgaug tool
- Image Data Augmentation for Deep Learning: A Survey
- A survey on Image Data Augmentation for Deep Learning

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."