# ML Practicum: Image Classification

Preventing Overfitting

As with any machine learning model, a key concern when training a convolutional neural network is *overfitting*: a model so tuned to the specifics of the training data that it is unable to generalize to new examples. Two techniques to prevent overfitting when building a CNN are:

- **Data augmentation**: artificially boosting the diversity and number of training examples by performing random transformations to existing images to create a set of new variants (see Figure 7). Data augmentation is especially useful when the original training data set is relatively small.

- **Dropout regularization**: Randomly removing units from the neural network during a training gradient step.

For more on dropout regularization, see [Training Neural Nets](https://developers.google.com/machine-learning/crash-course/training-neural-networks/) (https://developers.google.com/machine-learning/crash-course/training-neural-networks/) in [Machine Learning Crash Course](https://developers.google.com/machine-learning/crash-course/) (https://developers.google.com/machine-learning/crash-course/).
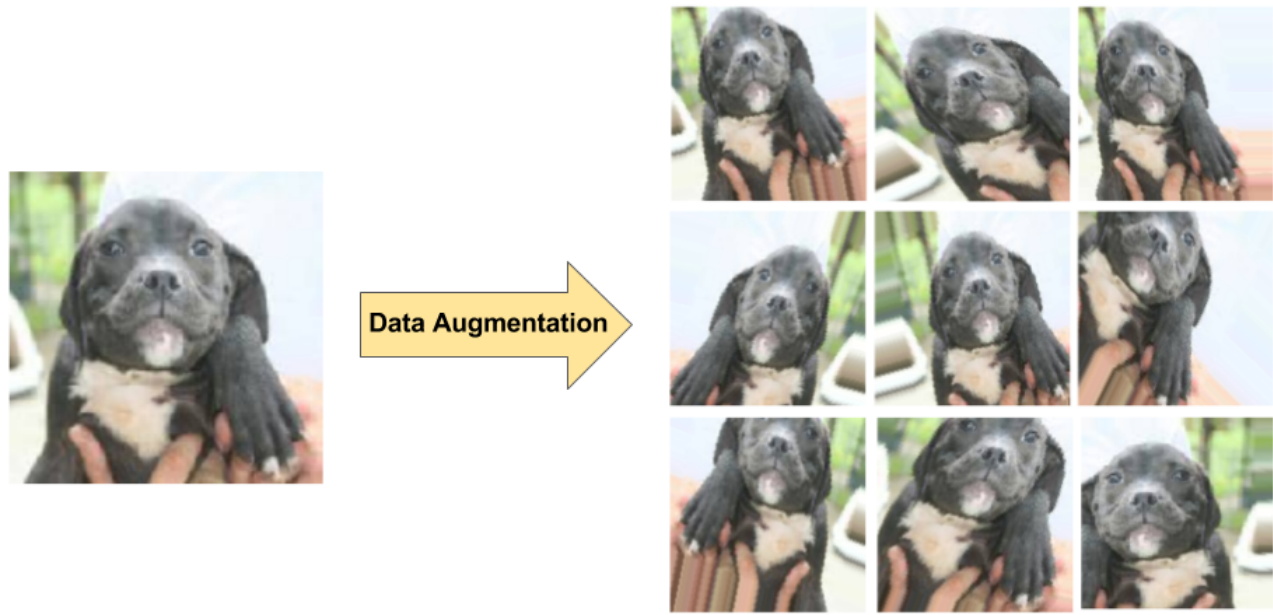
Figure 7. Data augmentation on a single dog image (excerpted from the "Dogs vs. Cats" dataset (https://www.kaggle.com/c/dogs-vs-cats/data) *available on Kaggle).* **Left**: *Original dog image from training set.* **Right**: *Nine new images generated from original image using random transformations.*

Overfitting is more of a concern when working with smaller training data sets. When working with big data sets (e.g., millions of images), applying dropout is unnecessary, and the value of data augmentation is also diminished.

Key Terms

- data augmentation

- overfitting
 (https://developers.google.com/machine-learning/glossary#overfitting)

- dropout regularization
 (https://developers.google.com/machine-learning/glossary#dropout_regularization)

Previous