Data Augmentation In The Cloud

Our proposal is to develop a web based service that performs the computationally expensive task of augmenting data sets. Users will upload an image or sets of images to the site, which will then be augmented by a background process. The output augmented set of images will then be presented back to the user. Parameters such as the type and number of augmentations performed will be configurable for the user. Our Minimal Viable Product (MVP) will be a feed-based site, similar to imgur.com, displaying a chronological tile-styled grid of uploaded images. Selecting an image will present the user with all its augmented versions, and an option to download them. The MVP will have no user accounts, and as such the home feed will display all uploaded images to any users. In the MVP, logistics such as how would the original uploader delete their images is yet to be decided.

Some stretch goals would be to include tags to image sets, to allow users to browse already available data. Adding user accounts would allow for private uploads, allowing users to only share their data with select others. Adding user account would require components such as secure password storage, session cookies, password reset options etc, hence accounts not being a part of the MVP. In addition to simple augmentations such as cropping, rotation and noise, another stretch goal will be to explore the use of Generative Adversarial Networks in data augmentation. This process has been shown to increase the accuracy of trained networks with small data sets by over 13% [1]. A further stretch goal would be tiers of user accounts. Splitting the service into two tiers, the free version would augment a user's uploads during off-peak times (for example overnight). A premium tier would give priority to the augmentation of a premium user's images, and would happen as soon as resources are available. The length of the period in which augmented images are retained could also be determined by a user's tier. A free user may have a week in which to access their images before deletion, whereas a premium user may have a month.

Using Amazon Web Services (AWS), we will use Elastic Cloud Compute (EC2) to host a web server and accompanying background processes that make up our product. EC2 will allow for elastic scaling of compute resources, for example provisioning additional capacity to perform batch augmentation upon a user uploading a large input set. Amazon's Simple Storage Service (S3) will be used to store images, as its object-based design is ideal. We plan to have a Node server handling requests, and a background Python service to perform the augmentations. React will be used for a responsive front-end. We aim to define an API contract so that development of the web UI can be entirely separate from development of the Node server.

To provide some context as to why we decided on this project, we will look at the background of data augmentation. Data augmentation is the process of artificially enlarging a data set to be used for training a neural network. Augmentation can be done in a number of ways, including cropping, translation, rotation and addition of noise. If used correctly, performing these types of transformation on a training set can greatly increase the accuracy of the trained network [2]. This is useful as the target system may be used in non-ideal settings, and augmenting the training data can account for varying real-world conditions.

There are two main types of data augmentation: online and offline. Randomly augmenting small batches of data prior to feeding them into a network is known as 'online' augmentation. This is a practical solution when augmenting an entire large data set would be computationally expensive and costly to store. Offline augmentation is the process of augmenting a data set pre-training. This is an effective method for smaller data sets, as the set size increases by a factor of the number of augmentations performed. Our service will offer an 'offline' augmentation service to users.

---

[1] - A Antoniou, A Storkey, H Edwards (2017) - Data augmentation generative adversarial networks

[2] - J Wang, L Perez (2017) - The Effectiveness of Data Augmentation in Image Classification using Deep Learning