✨ Mushroom classifier ✨

Team members:

Glib Manaiev
Dmytro Fedorenko

Setting up

Business understanding

- Identifying your business goals
  - **Background**: As the automation of mushrooms harvesting is becoming more and more popular, we decided to develop a system for automatic mushroom classification, which in theory is supposed to help with mushroom harvesting automation.
  - **Business goals:** Bring a tool to the market which will find its use both for the general public (people curious about mushroom picking, to ensure the safeness of their findings) and large businesses (e.g., a food production company can install our system in their facilities to increase the safety of their products and reduce operating costs by using the automated system instead of manual mushroom checking by people).
  - **Business success criteria:** If our system proves itself to be compatible in the mushroom identification industry (e.g., will achieve high accuracy and will adequately handle non-professional pictures), then we will be able to launch the product and advertise it correctly to gain a sufficient number of customers to consider our product successful.
- Assessing your situation
  - **Inventory of resources:** to successfully complete our project, we may enroll experts mushroom pickers to

contribute valuable information on how to increase the prediction accuracy or which features should we mainly consider when identifying a mushroom; to improve our solution's sustainability, we will need to expand our mushroom library continuously and the size of the training data, for that we may use the help of the users or other mushroom industry players by using their data, also for convenient training of the neural network we will need high power hardware resources (computer with powerful GPU) to generate new possible models continuously.

- **Requirements:** take a picture of a mushroom and upload it to the system; Assumptions: the mushroom is clearly visible in the image, and it takes up most of the frame; Constrains: If the mushroom wasn't present in the training dataset, we wouldn't be able to identify it.

- **Risks:** a possibility to wrongly identify a mushroom which could result in the unfortunate outcome for the end-user (e.g., cause a death of a person to eat the wrongly identified mushroom); Contingencies: if a mushroom in a picture differs from its congeners usual look (unusual shape, color, size, proportions, etc.) we likely will have problems with the identification.

- **Terminology:** mushroom family - the outcome of the identification, confidence - how likely is it that predicted mushroom family matches the actual one, training dataset - the library of mushroom pictures with correct labels (mushroom family) used for the training of the neural network.

- **Costs and benefits:** our solution would enable users to use an automated tool for mushroom identification, meaning that they won't need to spend their time and money on learning the mushroom families or hiring an expert for that purpose. Also, it may significantly benefit the industry companies as they will be able to perform automated control over their products and won't need special workers for this, meaning that they can reduce operation and production costs while maintaining, or even highering (by eliminating the human error), the quality level of their production.

- **Defining your data-mining goals**
    - **Data-mining goals:** we aim to obtain a highly accurate and confident model that will identify mushrooms from the pictures taken by people correctly. To achieve this, we will be using the dataset containing pictures of mushrooms belonging to different families.
    - **Data-mining success criteria:** as mushroom identification in the real world is mainly done by people having extensive knowledge of mushrooms, our resulting model should be compatible with them, as this is the point of it: to eliminate the human factor from the process of mushroom classification. We need to get the model with high accuracy and confidence, sustainable to the possible variations to achieve this. And to verify the outcome, we will need special people with expertise in mushroom classification and datasets prepared by such people.

Data understanding

- **Gathering data**
    - **Outline data requirements**
        - We are looking at images of mushrooms with annotations containing a type of mushroom attached to them. For each mushroom type, at

        least 300-350 image annotation pairs are required.

    - **Verify data availability**
        - Data is available in excess. We were able to find one public and one private dataset, access to which could be requested in case some severe problems will occur with the public dataset we have found.

          The issue is that these datasets do not consist of many mushroom types, so the network will be able to distinguish mushroom types from relatively narrow spectra.

    - **Define selection criteria**
        - We are currently looking at single images or datasets of mushroom images with known mushroom types attached to them .

- **Describing data**
    - Most probably, the kaggle mushroom dataset will be used, which contains pictures with different types of mushrooms that are located in separate folders with the names of the folders corresponding to the mushroom family name.
    - The dataset contains 6714 mushroom images of 9 different types.
    - Except for converting annotations from the folders names to COCO JSON files, additional annotations are not needed. Usually, the mushroom takes up almost the entire area of the image, so the image border will be assumed to be the mushroom's bounding box.

- **Exploring data**

- As mentioned before, the only major drawback of the dataset we plan to use is that its annotations will have to be converted to COCO format.
- This could be achieved not only manually balancing the dataset, but also by using data augmetation during training.Also, it would likely need to be balanced, since, for some mushroom types, there are several times more pictures present than for the others.

- Verifying data quality
  - Images are small enough to provide relatively fast network training, but at the same time, the overall quality is good enough for effective object detection.
  - Also, each mushroom type has enough images in it for effective training.

Planning your project

- Collect a balanced image dataset with different kinds of mushroom pictures, divide it into train, eval and test datasets. (10h)
- Convert dataset annotations to the COCO format. (5h)
- Train network (Fast-RCNN) on the training dataset. (48h)
- Evaluate the resulting model on the evaluation dataset. (5h)
- Tune hyperparameters based on the evaluation results. Retrain if needed. (10h)
- Try to apply training data augmentation by randomly: resizing the image, cropping the image, changing brightness and other parameters. (5h)
- Apply the model to the testing images dataset. (1h)