## 📋 Prerequisite: Your AWS and Kaggle Accounts

This process uses the Kaggle API, which is the fastest and most reliable way to get the ImageNet-1k (ILSVRC2012) dataset.

- **AWS Account:** You must have an AWS account.
- **Kaggle Account:**
    1. Go to Kaggle.com and create an account.
    2. Go to the "ImageNet Object Localization Challenge" page.
    3. Click the **"Join Competition"** button and **"Accept"** the rules. You cannot download the data until you do this.
    4. Go to your Kaggle account settings (click your profile picture -> "Account").
    5. Scroll to the "API" section and click **"Create New API Token"**.
    6. This will download a file named `kaggle.json`. Save this file. You will need it in Step 3.

## ⚙️ Step 1: Launch EC2 Instance & Create EBS Volume

First, you set up your "downloader" machine and the "virtual hard drive."

### Launch EC2 Instance

1. Go to the AWS EC2 Console and click "Launch instances".
2. **Name:** `ImageNet-Downloader`
3. **AMI:** Ubuntu (or Amazon Linux 2)
4. **Instance Type:** `t3.micro` (This is cheap and fine for downloading).
5. **Key Pair:** Select or create a key pair (e.g., `my-key.pem`) and download it. You will need this to log in.
6. **Network (Security Group):** Ensure it allows **SSH** from "My IP".

### Create EBS Volume

1. On the left menu, go to "Volumes" and click "Create volume".
2. **Volume Type:** `gp3`
3. **Size (GiB): 1000 GB.** (Do not use less, the uncompressed dataset is very large).
4. **Availability Zone:** Select the **exact same zone** as your EC2 instance (e.g., `us-east-1a`).
5. Click "Create volume".

### Attach Volume to Instance

1. Select your new 1000 GB volume (it will say "available").
2. Click "Actions" -> "Attach volume".
3. In the "Instance" field, choose your `ImageNet-Downloader` instance.
4. Click "Attach".

## 🔌 Step 2: Connect to EC2 and Mount the EBS Volume

Now you will log in to your machine and prepare the "hard drive."

### 1. Connect via SSH

Open a terminal on your computer.

```
# (Only needed once) Secure your key file
chmod 400 /path/to/your-key.pem

# Connect to the instance (get the Public IP from the EC2 console)
ssh -i /path/to/your-key.pem ubuntu@YOUR_INSTANCE_PUBLIC_IP
```

### 2. Format the Drive (First-Time-Only)

```
# Find your 1000GB drive. It's usually 'xvdf'.
lsblk

# Format the empty drive with the xfs filesystem
sudo mkfs -t xfs /dev/xvdf
```

### 3. Mount the Drive

```
# Create a folder to mount it to
sudo mkdir /data

# Mount the drive
sudo mount /dev/xvdf /data

# Give your 'ubuntu' user permission to write to it
sudo chown -R ubuntu:ubuntu /data
```

Your 1000 GB "hard drive" is now ready. Anything you put in the `/data` folder will be saved on it.

## 📥 Step 3: Download ImageNet using the Kaggle API

### 1. Install Kaggle API

```
sudo apt update
sudo apt install python3-pip
pip install kaggle
```

### 2. Upload Your Kaggle API Key

On your **local computer's terminal** (not the EC2 instance), use the `scp` command to upload the `kaggle.json` file you downloaded earlier.

```
# This command is run from your local machine
scp -i /path/to/your-key.pem /path/to/kaggle.json ubuntu@YOUR_INSTANCE_PUBLIC_IP:~/
```

**3. Configure the API Key (on your EC2 instance)**

Go back to your EC2 SSH terminal...

```
# Move the key to the correct folder
mkdir ~/.kaggle
mv ~/kaggle.json ~/.kaggle/

# Set the correct permissions (this is required)
chmod 600 ~/.kaggle/kaggle.json
```

**4. Start the Download**

1. Go to your data drive. This is important!

   ```
   cd /data
   ```

2. Use the `screen` command (so the download continues if you disconnect):

   ```
   screen
   ```

3. Start the download. This will take many hours.

   ```
   kaggle competitions download -c imagenet-object-localization-challenge
   ```

4. You can now safely disconnect by pressing **Ctrl+A**, then **D**. To check on it, log back in and type `screen -r`.

## 🗃️ Step 4: Extract All the Files

After the download finishes, you will have a file named `imagenet-object-localization-challenge.zip`.

**1. Install Unzip**

```
sudo apt install unzip
```

**2. Unzip the Main File**

```
# Make sure you are still in the /data directory
cd /data
unzip imagenet-object-localization-challenge.zip

# This will create several files, including the two big ones:
# ILSVRC2012_img_train.tar (~138 GB)
# ILSVRC2012_img_val.tar (~6.3 GB)
```

**3. Extract the Training Set (This is a 2-step process)**

First, extract the main training file (this creates 1000 `*.tar` files):

```
mkdir train
tar -xvf ILSVRC2012_img_train.tar -C ./train
```

Second, extract all 1000 of those smaller tar files. This script will do it for you. Run it from inside the `/data/train` directory:

```
cd /data/train

# This command finds every .tar file and extracts it into its own folder
find . -name "*.tar" | while read NAME ; do mkdir -p "${NAME%.tar}"; tar -xvf "${NAME}"

# Clean up the .tar files to save space
rm *.tar
cd /data
```

**4. Extract the Validation Set (Also a 2-step process)**

First, extract the files:

```
mkdir val
tar -xvf ILSVRC2012_img_val.tar -C ./val
```

Second, you must run a script to put them into subfolders (this is required by most training code).

```
# Download the official helper script
wget -qO- [https://raw.githubusercontent.com/soumith/imagenetloader.torch/master/valpre
```

(Note: If that `wget` command fails, you may need to go to that URL, copy the script, save it as `valprep.sh`, and run `bash valprep.sh`).

Your ImageNet data is now fully downloaded and extracted at `/data/train` and `/data/val`, ready for training.

## 🚀 Step 5: How to Use It Again (The Final Goal)

You are now finished with the `ImageNet-Downloader` instance.

1. **Shut Down:** Log out of the SSH terminal ( `exit` ).
2. **Stop Instance:** In the AWS Console, **Stop (do not terminate)** your `ImageNet-Downloader` instance.
3. **Detach Volume:**
   - Go to the "Volumes" section.
   - Select your 1000 GB volume.
   - Actions -> "Detach volume". Wait for its status to become "available".
4. **Launch Training Instance:**
   - Launch your powerful new GPU instance (e.g., `g4dn.xlarge` ).
   - **CRITICAL:** Make sure you launch it in the **same Availability Zone** as your EBS volume (e.g., `us-east-1a` ).
5. **Attach Volume:**
   - Once the GPU instance is running, go to "Volumes".
   - Select your 1000 GB ImageNet volume.
   - Actions -> "Attach volume".
   - Select your new GPU instance and attach it.
6. **Mount the Drive on the New Instance:**
   - SSH into your new GPU instance.
   - Run these commands (you don't format it this time):

```
# Create the mount folder
sudo mkdir /data

# Mount the drive (it already has an 'xfs' filesystem)
sudo mount /dev/xvdf /data

# Check your files
ls /data/train
```