Rise of Artificial Intelligence: A Deep Dive into Data Selection and Training

Hay ! I'm **Aryan Bhad**, a second-year CSE student certified in AI by **Oracle** and **Microsoft**. I've successfully **built multiple AI models**, including one that can differentiate between facial expressions. In this article, I'll share some **key insights** into the world of AI, focusing on the crucial aspects of **dataset selection and training**. My goal is to give you a solid understanding of the concepts and guide you towards filtering out the best data for your AI projects.

IN THIS ISSUE

Now whats an AI ?

An **Artificial intelligence** (AI) is a broad field aiming to create machines capable of human-like intelligence, such as problem-solving and decision-making. Machine learning (ML) is a subset of AI where machines learn from data without explicit programming. ML algorithms identify patterns, make predictions, and improve their accuracy over time. Essentially, AI is the overarching goal, while ML is a technique to achieve it.

There are three main methods of machine learning:

1.Supervised Learning: This involves training a model on labeled data, where each data point has an input and a corresponding output. The model learns to map inputs to outputs, allowing it to make predictions on new, unseen data. Examples include image classification (identifying objects in images) and spam detection (classifying emails as spam or not spam).

2.Unsupervised Learning: This involves training a model on unlabeled data, where the model must discover patterns and structures within the data on its own. Examples include clustering (grouping similar data points together) and dimensionality reduction (reducing the number of features in a dataset1 while preserving important information).

3.Reinforcement Learning: This involves training an agent to interact with an environment and learn to take actions that maximize rewards. The agent learns through trial and error, receiving feedback in the form of rewards or penalties. Examples include game playing (teaching a computer to play a game) and robotics (teaching a robot to perform a task).

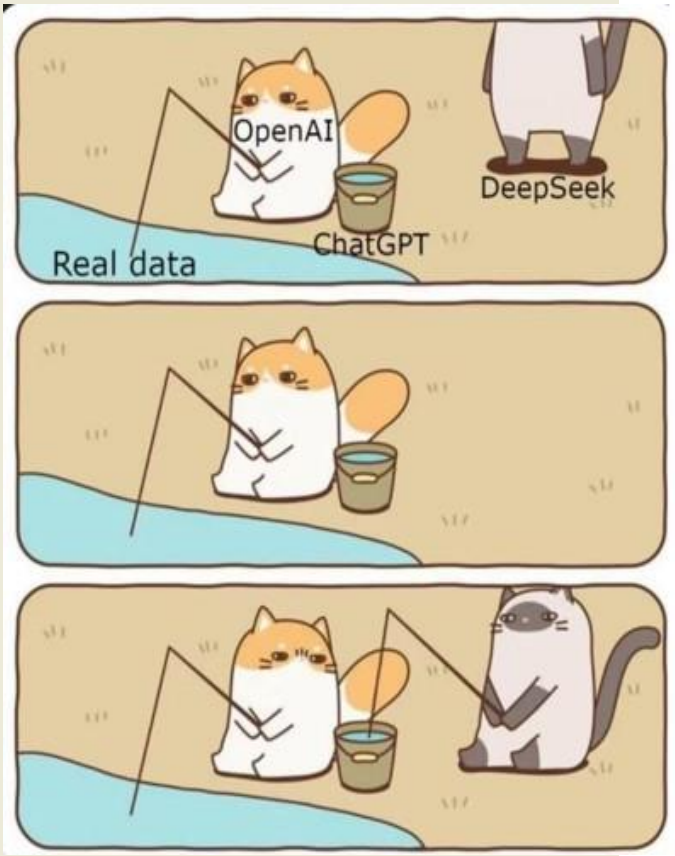
The Crucial Role of Data in AI/ML:

Data is the **lifeblood** of AI. Without it, AI algorithms cannot learn and make predictions on its own . The process of building an AI model involves feeding it vast amounts of data, known as a dataset, and training it to recognize patterns and make decisions based on that data.

DeepSeek's success demonstrates that strategic learning and focused development can rival sheer resources. The image illustrates how they've achieved such remarkable results.

DeepSeek benefits from a unique vantage point in its development: access to the vast dataset that includes OpenAI's generated information. This allows it to learn not only from its own training but also from the patterns and potential challenges inherent in OpenAI's data. Essentially, DeepSeek has the opportunity to build upon the foundation laid by others, much like a student learning from the successes and errors of a previous generation.

So that's the power of right dataset! the accuracy of an ai highly depends upon the dataset of which its trained on !



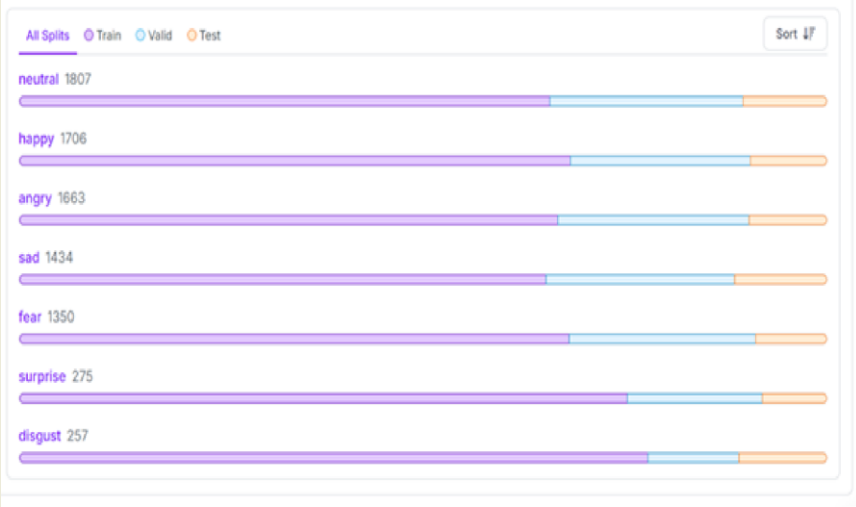
Now where do we start ?

Selecting the Right Dataset : Choosing the right dataset is crucial for the success of any AI project. Here are some key factors to consider:

- **Relevance:** The dataset should be relevant to the problem you're trying to solve.
- **Size:** The dataset should be large enough to capture the complexity of the problem.
- **Quality:** The data should be accurate, consistent, and free of errors.
- **Diversity:** The dataset should represent the real-world scenarios the AI will encounter.

Got the data ! Now what !?

Categorization



Categorize your Data into **3 main sets !!**

1. Training Set:

- **Purpose:** This is the largest portion of the data and is used to *train* the model. The model learns the patterns and relationships within this data by adjusting its internal parameters (weights and biases). Think of it as the textbook and practice problems the model studies.
- **Characteristics:** It should be representative of the overall data distribution. The larger the training set (within reason), the better the model can usually learn.

2. Validation Set:

- **Purpose:** This set is used to *evaluate* the model's performance *during* training. It helps to tune the model's hyperparameters (parameters that are set before training, like learning rate or network architecture). Think of it as a practice exam. You use the results to make adjustments *before* the final test.
- **Characteristics:** It should also be representative of the overall data and kept separate from the training set. It helps to detect overfitting. If the model performs very well on the training set but poorly on the validation set, it's a sign of overfitting.

3. Test Set:

- **Purpose:** This set is used to *evaluate* the *final* performance of the trained model. It's a completely unseen set of data that the model has never encountered during training or validation. Think of it as the final exam. It gives an unbiased estimate of how well the model will perform on real-world data.
- **Characteristics:** It should be representative of the overall data and kept completely separate from the training and validation sets. It's crucial for assessing the model's generalization ability.

NOW that we know the **basics of data** for AI what are the consecration you should take to **achive the accuracy as high as 80 to 85% like me** ?

• *Data Splitting Ratios*: Common ratios for splitting the data are 70% for training, 15% for validation, and 15% for testing, but these can vary depending on the size of the dataset and the specific problem.

• *Stratified Sampling*: When dealing with imbalanced datasets (where some classes have many more examples than others just like the inside shown above), stratified sampling is important. This ensures that each subset (training, validation, and test) has a similar proportion of examples from each class.

• *Cross-Validation*: In cases where the dataset is small, k-fold cross-validation can be used. The training set is divided into k parts (folds). The model is trained k times, each time using k-1 folds for training and the remaining fold for validation. The performance is averaged across all k folds. This helps to get a more robust estimate of the model's performance.

How Accuracy is Calculated:

Accuracy is a common metric used to evaluate the performance of classification models (models that predict categories). It's calculated as:

- `Accuracy = (Number of Correct Predictions) / (Total Number of Predictions)`

For example, if a model predicts the correct class for 80 out of 100 examples in the test set, its accuracy is 80%.

In Summary

Artificial intelligence (AI) is a broad field aiming to create machines with human-like intelligence. Machine learning (ML) is a subset of AI where machines learn from data without explicit programming. There are three main types of machine learning: supervised, unsupervised, and reinforcement learning. Data is crucial for AI/ML, and selecting the right dataset is essential. Datasets should be relevant, large, and diverse. Data is categorized into training, validation, and test sets. Accuracy is a common metric used to evaluate the performance of classification models.

You can connect here !

I am a highly motivated and skilled Tech enthusiast with a strong academic background and a proven track record of success. I have a deep understanding of the principles of AI and ML, and I am eager to apply my knowledge and skills to real-world problems. I am also a team player and I am confident that I can make a significant contribution to any organization. If you'd like to connect or explore more of my work, including my AI certifications from Oracle and Microsoft, and to verify my projects , please visit my website at [https://aryan4044.github.io/portfolio/].