PROMPTS

Dataset: https://github.com/joojs/fairface

1. Fine-Tuning Prompts

These prompts are designed to improve model generalization and fine-tuning strategies when working with the **FairFace dataset**.

Zero-Shot Prompts

Hey, I've been working with the FairFace dataset for age classification. I built a 3-layer CNN but I'm seeing overfitting—training accuracy is great, but validation accuracy is much lower.
Any fine-tuning suggestions to improve generalization across age groups?

Few-Shot Prompts

Example 1: I built a CNN for emotion detection using the FER2013 dataset. My initial model struggled with "disgust" and "fear" classes due to data imbalance. After switching to ResNet18 and applying class weights, I saw an improvement in recall for minority classes.
Example 2: I trained a model for age classification using a small dataset. My baseline CNN overfit to the majority age group. After increasing dropout and applying aggressive data augmentation, my model generalized better.
My Case: I built a CNN for age classification with FairFace. Started with 30.1% baseline accuracy but struggled with older age groups. After switching to MobileNetV2 and increasing dropout, my validation accuracy improved.
Now, should I adjust class-weighted loss or try focal loss to improve classification for the minority classes?

Chain-of-Thought Prompts (Show Chain of Thought Step-by-Step)

I'm struggling with my CNN model for age classification. Validation accuracy is much lower than training accuracy, suggesting overfitting.
Step-by-Step Debugging:

1. First, I analyzed the loss curves:
   - Training loss drops steeply, but validation loss is erratic.
2. Potential Fixes I tried:
   - Reduced learning rate from 0.005 to 0.001
   - Increased dropout (0.3 - 0.5) to prevent memorization
   - Applied more aggressive data augmentation

These helped somewhat, but I still see overfitting. What other techniques should I explore to generalize across age groups?

Role-Based Prompts

You are a deep learning expert specializing in CNN models for demographic classification.
Task:

- Diagnose potential overfitting issues in my CNN model.
- Recommend specific fine-tuning techniques to improve generalization.

Dataset: FairFace **Baseline Accuracy:** Low
Your response should include:

- Hyperparameter tuning strategies
- Regularization methods
- Alternative model architectures that could improve generalization

2. Data Cleaning Prompts

These prompts focus on **handling missing values, inconsistent labels, and duplicate images** in the FairFace dataset.

Zero-Shot Prompts

I just started working with the FairFace dataset for a facial recognition project, and I'm noticing some data quality issues. There are missing age values in quite a few records, and the gender labels are inconsistent throughout the dataset.

What's the best way to handle these missing values and standardize these categorical labels? I want to make sure I'm not introducing bias with my cleaning approach.

Few-Shot Prompts

Example 1: I worked with a medical dataset where gender labels varied as "M", "Male", "F", "Female". Standardizing them to "Male" and "Female" fixed inconsistencies.

Example 2: For a customer dataset, I had missing income values. Instead of dropping, I used median imputation to preserve data integrity.

My Case: In FairFace, some age values are missing, and gender labels vary as "M", "Male", "F", and "Female". I used mean imputation for age and standardized gender labels.

What's the best approach for ensuring consistency while maintaining dataset integrity?

Chain-of-Thought Prompts (Show Chain of Thought Step-by-Step)

I'm preparing the FairFace dataset for CNN training and have encountered several data quality issues.

Step-by-Step Analysis:

1. **Missing Values:**
   - About 10% of samples have missing age values.
   - Should I use mean imputation or drop these records entirely?
2. **Label Inconsistencies:**
   - Gender is labeled as "M", "Male", "F", and "Female".
   - Standardizing these to "Male" and "Female" seems logical, but should I consider other factors?
3. **Duplicate Images:**
   - Some images appear multiple times in the dataset.
   - What's the best method to automatically detect and remove these without biasing the dataset?

How should I approach these data cleaning challenges effectively?

Role-Based Prompts

You are a data scientist specializing in facial recognition datasets.

Task:

- Identify potential inconsistencies in the FairFace dataset.
- Recommend strategies for handling missing values, duplicate images, and imbalanced racial groups.

Dataset Details:

- Missing age values: 10%
- Duplicate images detected: 5%
- Racial class imbalance: Some groups have <5% representation.

Provide a structured data cleaning workflow to ensure the dataset is properly prepared while preserving its diversity.

3. Exploratory Data Analysis (EDA) Prompts

These prompts focus on **visualizing and understanding dataset biases before training**.

Zero-Shot Prompts

I'm about to start analysing the FairFace dataset before building a CNN model for age classification. What EDA techniques would you recommend to visualize class distributions and identify any imbalances? I want to make sure I understand the data well before jumping into modeling.

Few-Shot Prompts

Example 1: I analysed a dataset for speech emotion recognition and initially assumed the classes were balanced. However, after visualization, I found the dataset heavily favoured neutral emotions.
Example 2: For a financial dataset, I assumed all income levels were well represented. But upon plotting distributions, I discovered that higher income brackets had significantly fewer samples.
My Case: I've been looking at the age distribution in the FairFace dataset.
Initially, I assumed the age groups would be fairly balanced, but after some basic visualization, I discovered the dataset is heavily skewed toward younger individuals, which might explain some of the model performance issues I've been seeing.
Do you think I should apply stratified sampling to balance these age groups? Or would that introduce other problems I should be aware of?

Chain-of-Thought Prompts (Show Chain of Thought Step-by-Step)

I'm exploring the FairFace dataset and trying to figure out how to handle the imbalanced racial categories.
Step-by-Step Analysis:

1. **First, I'm plotting histograms** of the different racial groups to visualize the distribution. This helps identify which categories are underrepresented.
2. **Next, I'm computing correlations** between features like age and gender labels to check if there are any biases in feature representation.

I'm not entirely sure what to do once I've identified these imbalances. Should I use weighting, oversampling, or some other technique?
What's the best practice when dealing with demographic imbalances in facial recognition datasets?

Role-Based Prompts

You are a data scientist analysing demographic data.
I'm working with the FairFace dataset before building a CNN model and need to:

- Identify potential biases in racial and gender distribution
- Find effective visualization techniques to highlight these biases
- Determine appropriate preprocessing steps to address class imbalance

Could you outline a structured approach for performing this exploratory data analysis? I want to make sure I'm thorough in understanding the dataset before modeling.

4. Debugging (Warnings, Errors) Prompts

These prompts focus on diagnosing and resolving errors, performance issues, and compatibility problems in machine learning workflows.

Zero-Shot Prompts

I've been training a CNN model on the FairFace dataset for age classification, but my CPU usage is constantly hitting 100%, and the training is painfully slow. I have a GPU available but haven't set things up for it yet. Should I shift the workload to the GPU? And if so, what specific steps would I need to take to migrate from CPU to GPU in TensorFlow (or PyTorch if that's easier)?

I keep running into version mismatch warnings while working on my machine learning pipeline. TensorFlow is installed, but I'm getting CUDA incompatibility errors that prevent GPU acceleration. What's the best approach to resolve these conflicts and ensure proper GPU utilization? The errors are cryptic.

Few-Shot Prompts

My CNN model training is taking forever because of high CPU usage.

Example 1:

- **Before Debugging:** Training solely on CPU, hitting 100% utilization.
- **After Debugging:** Switched to GPU acceleration in TensorFlow, reducing training time significantly.

Now I'm wondering if I should enable mixed precision training to further optimize GPU usage. Would that give me another significant speed boost, or is it more trouble than it's worth at this point?

My TensorFlow code keeps throwing FutureWarning messages about deprecated APIs.

Example 2:

- **Before Debugging:** Used an older version of TensorFlow, getting warnings about tf.placeholder being deprecated.
- **After Debugging:** Updated TensorFlow and replaced deprecated functions, resolving the warnings.

Is there a systematic way to track and fix all these deprecation warnings? I feel like I'm playing whack-a-mole with them, and I'm worried some might cause actual issues down the line.

Chain-of-Thought Prompts (Show Your Thinking Step-by-Step)

I'm training a CNN model, but it's running significantly slower than expected. I'm trying to debug this methodically:

1. **Checked system resource usage:**
   - CPU is at 100% constantly, but GPU usage is barely registering.
   - Should I enable TensorFlow's device placement logging to confirm if the GPU is being used?
2. **Potential fixes:**
   - Explicitly set operations to run on GPU with tf.device('/GPU:0').
   - Reduce batch size to prevent memory overflow issues.
3. **Further optimization:**
   - Should I enable mixed precision training to speed up computation?

What other debugging steps would you recommend? I feel like I'm missing something obvious about why the GPU isn't being utilized properly.

Role-Based Prompts

You are a deep learning engineer specializing in optimizing training performance.

I'm running into performance issues with my CNN model training—specifically high CPU usage. Could you:

- Analyse potential reasons why my model is experiencing high CPU usage.
- Recommend strategies for migrating to GPU acceleration efficiently.
- Suggest debugging steps to identify potential hardware bottlenecks.

I'm especially interested in understanding the trade-offs between different approaches so I can make an informed decision for my specific setup.