Aufgabe6 pytorch

(b) Early Stopping (1 Punkt)

Definition:
Early stopping is a regularization technique used in training deep learning models to **prevent overfitting** by **stopping training when the validation loss stops improving**.

Use Case:

- Helps in preventing **overfitting** by stopping the training **before the model memorizes noise**.
- Saves **computation time** and **prevents excessive updates** that degrade generalization.

(c) Beam Search Prediction (2 Punkte)

Beam search is a **heuristic search algorithm** used in sequence generation models like Transformers to maintain the most probable sequences at each step.

Steps for Beam Search:

1. **Beam size = 1 (Greedy decoding)**
    - Select the **most probable token** at each step.
    - Faster but may **miss better sequences**.
2. **Beam size = 2**
    - Keeps track of the **top-2 most probable sequences** at each step.
    - Allows exploration of **alternate sentence structures** that might be better in the long run.

Task 1: RLHF (Reinforcement Learning from Human Feedback) – Explanation & Solution

This question tests **RLHF (Reinforcement Learning from Human Feedback)**, a key technique used to train models like **InstructGPT**, aligning them with human preferences to improve response quality.

(a) Three Models in InstructGPT Training

Question:
List the three models used in **InstructGPT** training and briefly explain:

1. **(i) What they are trained on**
2. **(ii) What they produce**

Solution

1. **Pre-trained Model**
    - **(i) Training Objective:** This model is trained on a **large-scale unsupervised text corpus** using **auto-regressive language modeling (LM objective)**, predicting the next word given the previous words.
    - **(ii) Output:** Generates **general language text**, but it lacks alignment with human preferences.
2. **Reward Model (RM)**
    - **(i) Training Objective:** This model is trained using **human preference data**, learning to assign scores that reflect human judgments of response quality.
    - **(ii) Output: A numerical score** that evaluates the quality of generated responses (higher scores indicate better alignment with human preferences).
3. **Fine-tuned Policy (RL-Tuned Model)**
    - **(i) Training Objective:** This model starts from the **pre-trained model** and is fine-tuned using **reinforcement learning (PPO - Proximal Policy Optimization)** to maximize the scores assigned by the reward model.
    - **(ii) Output:** Generates **responses that align with human expectations**, such as more polite, relevant, and coherent answers.

(b) Which Model is Trained with the Given Objective?

The question provides a specific objective function and asks which model it corresponds to.

Solution:

- If the objective **maximizes a reward function**, it corresponds to **the fine-tuned policy**, as reinforcement learning optimizes this.
- If the objective **fits human-labeled ranking data**, it corresponds to **the reward model**, which learns to predict preference scores.

Reasoning:

- **If the goal is to optimize scores assigned by RM**, it belongs to **Fine-tuned Policy (RL Model)**.
- **If the goal is to learn human preferences directly**, it belongs to **Reward Model**.

(c) Mapping Another Objective to a Model

This sub-question is similar to (b), requiring you to match another given objective function to a specific model.

Solution:

- **If the objective is supervised learning (e.g., cross-entropy loss)** → It is used for **the Pre-trained Model or the Reward Model**.
- **If the objective involves policy updates in reinforcement learning** → It is for **the Fine-tuned Policy**.

(d) Explanation of Three Terms (r θ r_{\theta}, β \beta, γ \gamma) in the Objective Function

This question requires explaining the three key terms in **PPO (Proximal Policy Optimization)**.

Solution:

1. **r_{θ} (Policy Ratio)**
   - **Role:** Measures how different the current policy π_{θ} is from the old policy π_{old}: $r_{\theta} = \frac{\pi_{\theta}(a|s)}{\pi_{\text{old}}(a|s)}$
   - **Purpose:** Prevents excessively large updates to the model's policy, ensuring stable learning.
2. **β (Entropy Bonus)**
   - **Role:** Encourages exploration by preventing the model from collapsing into a deterministic strategy: $L_{\text{entropy}} = -\beta \sum_a \pi_{\theta}(a|s) \log \pi_{\theta}(a|s)$
   - **Purpose:** Ensures the model does not become overconfident in specific responses too early.
3. **γ (Discount Factor)**
   - **Role:** Determines how much future rewards influence the current decision: $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k}$
   - **Purpose:** Controls the importance of **long-term rewards**, where a lower γ makes the model prioritize short-term gains.

(e) What is the Objective of the Third Model?

Question:
The exam provides objective functions for two models and asks for the third model's objective function.

Solution

- **Pre-trained Model's Objective:**
  - Standard **language modeling loss (LM objective)**: $L_{\text{LM}} = - \sum_{t} \log P_{\theta}(x_t \mid x_{<t})$

    - This objective does **not involve reinforcement learning** or reward modeling—it purely learns to predict the next token.
- **Reward Model's Objective:**
    - The **Reward Model R_{ϕ}** is trained with **Pairwise Ranking Loss**: $L_{\text{RM}} = - \sum_{(x, x^+)} \log \sigma(R_{\phi}(x^+) - R_{\phi}(x))$
    - **Purpose:** Ensures that responses preferred by humans receive higher scores.
- **Fine-tuned Policy (RL Model's Objective):**
    - **Optimized using PPO loss**: $L_{\text{PPO}} = \mathbb{E} \left[\min(r_{\theta} A, \text{clip}(r_{\theta}, 1-\epsilon, 1+\epsilon) A) \right]$
    - **Purpose:** Updates the model's response generation strategy to maximize human preference scores.

Summary

Topic	Key Concept
(a) Three Models	Pre-trained Model (LM objective), Reward Model (Human preference learning), Fine-tuned Policy (PPO training)
(b), (c) Objective Matching	Supervised learning (Reward Model) vs. Reinforcement learning (Fine-tuned Policy)
(d) PPO Objective Components	Policy ratio r_{θ}, entropy bonus β, discount factor γ
(e) Different Objective Functions	Language modeling loss L_{LM}, reward model loss L_{RM}, PPO loss L_{PPO}

Exam Focus Areas

1. **Understanding the Three-Stage RLHF Process**
    - Pre-training → Reward Model Training → Reinforcement Learning (PPO)

2. **Recognizing Different Objective Functions**
   - **Cross-entropy loss** (Language Modeling)
   - **Pairwise ranking loss** (Reward Model)
   - **PPO loss** (Fine-tuned Policy)
3. **Why RLHF is More Effective than Supervised Learning**
   - **Combines human feedback with reinforcement learning** to improve response quality.

This section covers **one of the most important reinforcement learning techniques in NLP**. If you have any questions or need further clarifications, feel free to ask! 🚀