Challenge 3: Synthesis Prediction

Giovanni Filomeno

Artificial Intelligence in Life Science

JOHANNES KEPLER UNIVERSITY LINZ

Problem overview

The goal of this task is to predict the reactant SMILES strings from given product SMILES using a sequence-to-sequence model. The model is trained on a dataset of chemical reactions formatted as "[Reactants] >> [Products]", aiming to learn the reverse transformation.

Model Architecture

- **Architecture**: Sequence-to-Sequence model based on the T5 transformer (ReactionT5v2).
- **Pretrained Model**: Initialized with sagawa/ReactionT5v2-retrosynthesis, specialized for retrosynthesis tasks.
- **Tokenizer**: Uses a SMILES-aware tokenizer from the HuggingFace Transformers library.
- **Training Objective**: Given a product SMILES, predict the corresponding reactant SMILES.
- **Loss Function**: Cross-entropy with label smoothing (0.1) to improve generalization.
- **Evaluation Metric**: Top-1 accuracy, based on exact matching of canonicalized SMILES sets.
- **Inference Strategy**: Beam search (num_beams=10) for robust generation of SMILES predictions.

Training Strategy & Tuning

- **Fine-Tuning**: The model is fine-tuned on a custom dataset of [Reactants] >> [Products] reactions.
- **Train/Test Split**: 90% training, 10% validation split from tokenized dataset.
- **Batching**: Effective batch size of 16 via gradient accumulation (batch size 4 x 4 steps).
- **Learning Rate**: Set to 5e-4 with a cosine scheduler and weight decay of 0.01.
- **Early Stopping**: Training stops early if no improvement in top-1 accuracy for 3 consecutive evaluations.
- **Label Smoothing**: 0.1 factor used to mitigate overfitting and encourage confident predictions.

Results

#	Top-1 Accuracy	Description	Upload Date
1	0.233	Try_before_christmas_4	June 22, 2025, 9:34 p.m.

> 0.20