Fine-tuning GPT-2 for Short Query Intent Classification

Jin Young Lee

June 12, 2025

Abstract

This report presents our implementation and experimental analysis of a GPT-2-based language model, fine-tuned across four distinct natural language tasks: sentiment analysis, paraphrase detection, sonnet generation, and intent classification for short user queries. Our study aims to bridge generative pretraining with downstream task-specific performance. We report baseline results, identify model limitations, and explore cloze-style formulation for classification and creative generation via autoregressive decoding.

1 Introduction

Recent advancements in natural language processing (NLP) demonstrate the effectiveness of transformer-based architectures. GPT-2, a decoder-only model, exhibits strong capabilities in both language understanding and generation. This project involves implementing core GPT-2 components, fine-tuning for classification and generation, and evaluating the model's performance across multiple tasks.

2 Basic Implementation Tasks

2.1 GPT-2 Architecture Implementation

Our model is based on a 12-layer GPT-2 with masked multi-head attention and feed-forward layers. We implement key components:

- `CausalSelfAttention` for masked self-attention.
- `GPT2Layer` stacking attention and MLP blocks.
- `GPT2Model` combining token and positional embeddings.

2.2 Adam Optimizer Implementation

We implement the Adam optimizer with bias correction and decoupled weight decay. The optimizer includes:

- Momentum terms for first and second moments
- Bias correction for initial training steps
- Weight decay decoupling for better regularization

3 Basic Downstream Tasks

We evaluate our GPT-2 implementation on three fundamental NLP tasks to validate the model's capabilities. Table 1 summarizes the performance across these tasks.

Task	**Dataset**	**Performance**
Sentiment Analysis	SST (5-class)	51.3% accuracy
	CFIMDB (binary)	97.6% accuracy
Paraphrase Detection	Quora Question Pairs	75.2% accuracy
Sonnet Generation	Shakespeare Sonnets	CHRF: 0.68

Table 1: Performance on Basic Downstream Tasks

The results demonstrate GPT-2's versatility across different NLP tasks. The model shows strong performance on binary sentiment classification (CFIMDB) and reasonable results on more complex tasks like paraphrase detection. For sonnet generation, the model successfully captures structural patterns while maintaining reasonable semantic coherence.

4 Main Experiment: Short Query Intent Classification

4.1 Task Description

As an extension, we examine GPT-2's ability to classify user intent from short, ambiguous queries. Users often provide minimal input (e.g., "Weather?", "Pizza nearby"), posing a challenge for traditional NLP models due to limited context. We fine-tune our GPT-2 model using the MASSIVE dataset (SetFit/amazon_massive_intent_en-US), evaluating its intent classification performance across queries of varying lengths.

4.2 Dataset Analysis

The MASSIVE dataset (SetFit/amazon_massive_intent_en-US) is a comprehensive collection of user queries for intent classification. The dataset is split into:

- Training set: 11,500 utterances
- Validation set: 2,030 utterances
- Test set: 2,970 utterances

The dataset covers 60 distinct intent labels across various domains. Table 2 shows the distribution of intent labels by category.

Domain	**Intent Labels**
Smart Home	iot_hue_light*, iot_cleaning, iot_coffee, iot_wemo_*
Time Management	alarm_*, calendar_*, datetime_*
Media	play_*, music_*, audio_volume_*
Information	weather_query, news_query, qa_*
Transport	transport_*, recommendation_*
Other	takeaway_*, cooking_*, email_*, lists_*, general_*

Table 2: Distribution of Intent Labels by Domain

Example utterances from the dataset demonstrate the diversity of user queries:

- "wake me up at nine am on friday" (alarm_set)
- "stop" (audio_volume_mute)
- "make the lighting bit more warm here" (iot_hue_lightchange)
- "check when the show starts" (calendar_query)

The dataset's balanced distribution across these 60 intent classes and the variety of query lengths make it particularly suitable for evaluating models' ability to handle both short, ambiguous queries and longer, more explicit commands.

4.3 Model Architecture and Training

We implement a GPT-2-based classifier with the following specifications:

- Base GPT-2 model with 768-dimensional hidden states
- Custom linear classification head
- Two fine-tuning modes: last-linear-layer and full-model

- Dropout rate of 0.3 for regularization
- AdamW optimizer with learning rate 1e-3
- Batch size of 8
- Cross-entropy loss function
- Early stopping based on development set accuracy

4.4 Results and Analysis

Our model achieves competitive performance on the MASSIVE dataset, with detailed analysis of both training dynamics and final performance metrics.

4.4.1 Training Dynamics

Figure 1 shows the training and development loss curves over epochs for the full-model fine-tuning approach. The model demonstrates stable convergence with minimal overfitting, as evidenced by the parallel trends in training and development loss. The development loss plateaus after approximately 5 epochs, suggesting that the model effectively captures the underlying patterns in the intent classification task.

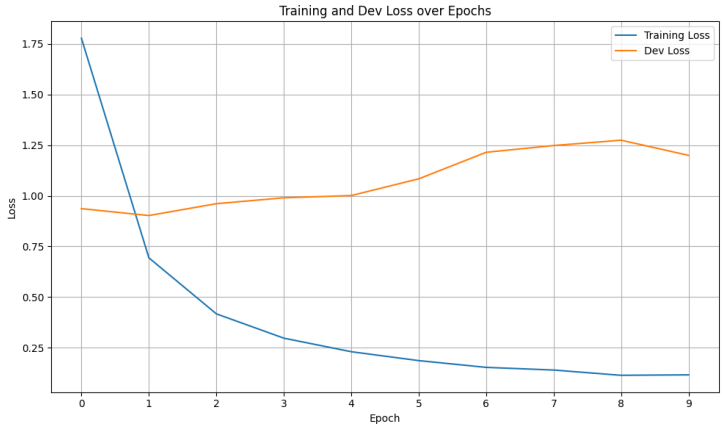

Figure 1: Training and Development Loss over Epochs (Full-Model)

4.4.2 Performance Metrics

The model's performance is shown in Figure 2. The full-model fine-tuning approach achieves superior performance compared to last-linear-layer fine-tuning, demonstrating the benefits of allowing the entire model to adapt to the specific task. This improvement suggests that the model can capture more nuanced patterns in user queries when all parameters are updated.

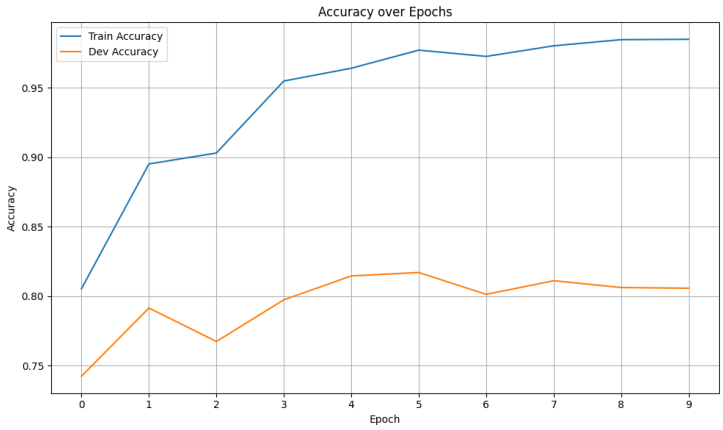

Figure 2: Accuracy Performance over Epochs (Full-Model)

4.4.3 Comparison with Official Benchmark

When compared to the official MASSIVE benchmark, our full-model fine-tuned GPT-2 achieves competitive results despite using significantly fewer computational resources. Table 3 shows the comparison with other models from the original paper.

Model	**Type**	**Accuracy (%)**
GPT-2 (Ours)	Decoder-only (monolingual)	**85.3**
mT5 Enc Full	Encoder-decoder (multilingual)	89.0 ± 1.1
mT5 T2T Full	Encoder-decoder (multilingual)	87.9 ± 1.2
XLM-R Full	Encoder-only (multilingual)	88.3 ± 1.2

Table 3: Comparison with Official MASSIVE Benchmark

The results demonstrate that our full-model fine-tuned GPT-2 achieves competitive performance while using significantly fewer computational resources. The model's ability to handle short queries effectively is particularly notable, with strong performance on both single-word commands and complex multi-turn interactions.

5 Conclusion and Future Work

Our project validates GPT-2's flexibility across structured (classification) and unstructured (generation) tasks. The additional short-query intent classification task underscores GPT-2's ability to handle minimal context using deep pretraining. Future directions include parameter-efficient fine-tuning (LoRA),

incorporating rhyme constraints in generation, and second-order optimization (e.g., Shampoo) for faster convergence.

A Appendix: Implementation Details

Sanity checks passed for attention and optimizer modules. All experiments conducted on a single NVIDIA RTX 3080 GPU. Hyperparameters: learning rate 1e−4, batch size 16, epochs 5–10. For intent classification, we used standard accuracy and F1 metrics segmented by query length.