

Spis treści

Przedmowa 7

Podziękowania 8

O tej książce 9

O autorze 13

Czym są duże modele językowe? 15

- 1.1. Czym jest model LLM? 16
 - 1.2. Zastosowania modeli LLM 18
 - 1.3. Etapy tworzenia modeli LLM i korzystania z nich 20
 - 1.4. Wprowadzenie do architektury transformerów 22
 - 1.5. Wykorzystanie dużych zbiorów danych 24
 - 1.6. Szczegóły architektury modeli GPT 27
 - 1.7. Tworzenie dużego modelu językowego 29
- Podsumowanie 30

Praca z danymi tekstowymi 32

- 2.1. Czym są osadzenia słów? 33
 - 2.2. Tokenizacja tekstu 36
 - 2.3. Konwersja tokenów na identyfikatory 39
 - 2.4. Dodawanie specjalnych tokenów kontekstowych 43
 - 2.5. Kodowanie par bajtów 47
 - 2.6. Próbkowanie danych z oknem przesunym 50
 - 2.7. Tworzenie osadzeń tokenów 56
 - 2.8. Kodowanie pozycji słów 58
- Podsumowanie 62

Spis treści

Kodowanie mechanizmów uwagi 64

- 3.1. Problem z modelowaniem długich sekwencji 65
- 3.2. Przechwytywanie zależności między danymi za pomocą mechanizmów uwagi 67
- 3.3. Zwracanie uwagi na różne części danych wejściowych przez mechanizm samouwagi 69
 - 3.3.1. *Prosty mechanizm samouwagi bez trenowalnych wag* 70
 - 3.3.2. *Obliczanie wag uwagi dla wszystkich tokenów wejściowych* 75
- 3.4. Implementacja mechanizmu samouwagi z trenowalnymi wagami 77
 - 3.4.1. *Obliczanie wag uwagi krok po kroku* 78
 - 3.4.2. *Implementacja kompaktowej klasy samouwagi w Pythonie* 82
- 3.5. Ukrywanie przyszłych słów dzięki zastosowaniu uwagi przyczynowej 87
 - 3.5.1. *Wykorzystanie maski uwagi przyczynowej* 88
 - 3.5.2. *Maskowanie dodatkowych wag uwagi z myciem dropoutu* 91
 - 3.5.3. *Implementacja zwięzłej klasy przyczynowej uwagi* 93
- 3.6. Rozszerzenie uwagi jednogłowicowej na wielogłowicową 95
 - 3.6.1. *Utworzenie stosu wielu jednogłowicowych warstw uwagi* 95
 - 3.6.2. *Implementacja uwagi wielogłowicowej z podziałem wag* 98
- Podsumowanie 103

Implementacja od podstaw modelu GPT do generowania tekstu 105

- 4.1. Kodowanie architektury *LIM* 106
- 4.2. Normalizacja warstwowa aktywacji 112
- 4.3. Implementacja sieci ze sprzężeniem w przód z aktywacjami GELU 118
- 4.4. Dodawanie połączeń skrótowych 122
- 4.5. Łączenie warstw uwagi i warstw liniowych w bloku transformera 126

Spis treści

4.6.	Kodowanie modelu GPT	129
4.7.	Generowanie tekstu	134
	Podsumowanie	139
<i>Wstępne szkolenie na nieoznakowanych danych</i>		<i>140</i>
5.1.	Ocena generatywnych modeli tekstowych	141
5.1.1.	<i>Używanie modelu GPT do generowania tekstu</i>	<i>142</i>
5.1.2.	<i>Obliczanie strat związanych z generowaniem tekstu</i>	<i>144</i>
5.1.3.	<i>Obliczanie strat w zestawie szkoleniowym i walidacyjnym</i>	<i>152</i>
5.2.	Szkolenie modelu LLM	157
5.3.	Strategie dekodowania w celu zarządzania losowością	163
5.3.1.	<i>Skalowanie temperaturą</i>	<i>164</i>
5.3.2.	<i>Próbkowanie top-k</i>	<i>167</i>
5.3.3.	<i>Modyfikacja funkcji generowania tekstu</i>	<i>169</i>
5.4.	Wczytywanie i zapisywanie wag modeli z użyciem frameworka PyTorch	171
5.5.	Ładowanie wstępnie przeszkolonych wag z modelu OpenAI	173
	Podsumowanie	179
<i>Dostrajanie modelu LLM do zadań klasyfikacji</i>		<i>181</i>
6.1.	Różne kategorie dostrajania	182
6.2.	Przygotowanie zbioru danych	183
6.3.	Tworzenie mechanizmów ładujących dane	188
6.4.	Inicjalizacja modelu z użyciem wag wstępnie przeszkolonego modelu	193
6.5.	Dodawanie nagłówka klasyfikacji	195
6.6.	Obliczanie straty i dokładności klasyfikacji	202
6.7.	Dostrajanie modelu na danych nadzorowanych	206
6.8.	Wykorzystanie modelu LLM jako klasyfikatora spamu	212
	Podsumowanie	214

<i>Dostrajanie modelu LLM do zadań wykonywania instrukcji</i>	215
7.1. Wprowadzenie do dostrajania do wykonywania instrukcji	216
7.2. Przygotowanie zbioru danych do nadzorowanego dostrajania pod kątem wykonywania instrukcji	218
7.3. Organizowanie danych w partie szkoleniowe	222
7.4. Tworzenie mechanizmów ładujących dane dla zbioru danych instrukcji	234
7.5. Ładowanie wstępnie przeszkolonego modelu LLM	237
7.6. Dostrajanie modeli LLM do zadań wykonywania instrukcji	241
7.7. Wyodrębnianie i zapisywanie odpowiedzi	245
7.8. Ocena dostrojonego modelu LLM	250
7.9. Wnioski	260
7.9.1. <i>Co dalej?</i>	260
7.9.2. <i>Bądź na bieżąco w szybko zmieniającej się dziedzinie</i>	261
7.9.3. <i>Na koniec</i>	261
Podsumowanie	261
<i>Dodatek A Wprowadzenie w tematykę frameworka PyTorch</i>	263
<i>Dodatek B Bibliografia i lektura uzupełniająca</i>	303
<i>Dodatek C Rozwiązania ćwiczeń</i>	315
<i>Dodatek D Usprawnianie pętli szkoleniowej</i>	328
<i>Dodatek E Skuteczne dostrajanie parametrów za pomocą LoRA</i>	337