

Fine-Tuning vs. Training from Scratch

Lingwistyka Obliczeniowa | Laboratorium 3

Wojciech Bartoszek

1 Cel i zakres zadania

Celem laboratorium było porównanie dwóch podejść do trenowania modeli językowych typu decoder-only w zadaniu klasyfikacji tekstu: trenowania małego modelu od podstaw (from scratch) oraz dostrajania (fine-tuning) modelu pretrenowanego.

2 Zbiór danych

Wybrano zbiór danych **Polish Youth Slang Classification** (`jziebura/polish_youth_slang_classification`).

- Liczba klas: 3 (Negative, Neutral, Positive)
- Podział: Train / Val / Test
- Liczba przykładów treningowych: 3469
- Liczba przykładów walidacyjnych: 434
- Liczba przykładów testowych: 434
- Rozkład klas (Train): 0: 1013, 1: 1777, 2: 679

3 Modele

3.1 Model trenowany od podstaw (From Scratch)

Mały model typu Transformer (Decoder-only) z głowicą klasyfikującą. Parametry modelu zostały dobrane w procesie optymalizacji hiperparametrów (Optuna, 500 prób) w celu minimalizacji straty walidacyjnej.

- Architektura: Transformer Decoder
- Liczba parametrów: 3,410,627
- Optymalizacja: AdamW, Scheduler z rozgrzewką (warmup)

3.2 Model dostrajany (Fine-Tuned)

Model pretrenowany pobrany z Hugging Face: `sdadas/polish-gpt2-small`.

- Nazwa modelu: `sdadas/polish-gpt2-small`
- Liczba parametrów całkowita: 125,952,768
- Liczba parametrów trenowalnych: 7,090,176
- Metoda dostrajania: Frozen Backbone (oprócz ostatniej warstwy) + Head Fine-Tuning

4 Wyniki eksperymentów

4.1 Porównanie jakości klasyfikacji

Model	Accuracy	F1 Score
From Scratch	0.5645	0.5047
Fine-Tuned	0.6866	0.6852

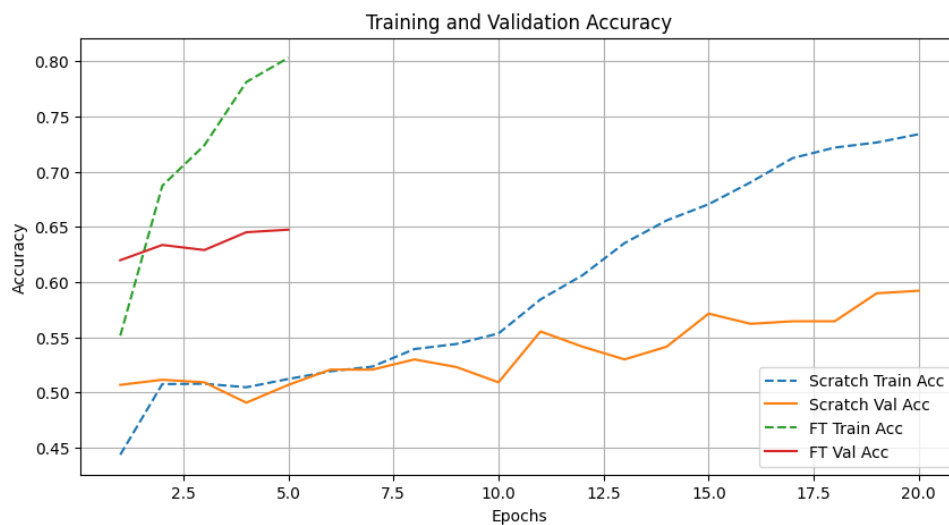
Tabela 1: Wyniki klasyfikacji na zbiorze testowym

4.2 Czas treningu i wnioskowania

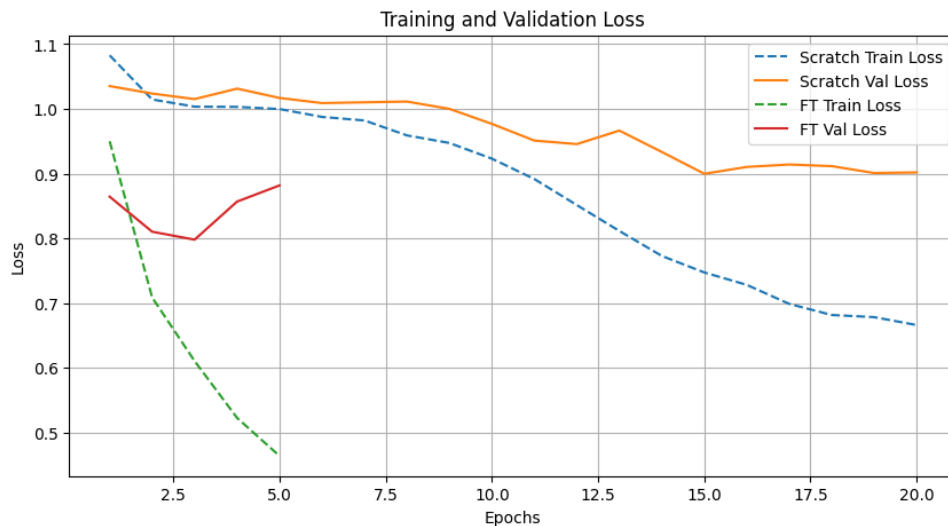
Model	Training Time (Total)	Training Time (per epoch)	Inference Time (ms/sample)
From Scratch	111.31 s	≈ 5.5 s	1.09 ms
Fine-Tuned	2391.58 s	≈ 478 s	39.45 ms

Tabela 2: Porównanie czasów treningu i wnioskowania

4.3 Wykresy uczenia



Rysunek 1: Historia dokładności (Accuracy) podczas treningu



Rysunek 2: Historia funkcji straty (Loss) podczas treningu

5 Wnioski

Zgodnie z oczekiwaniami, model dostrajany (Fine-Tuned) osiągnął lepsze wyniki jakościowe (Accuracy 0.6866, F1 0.6852) niż model trenowany od podstaw. Jednakże, model From Scratch (Accuracy 0.5645, F1 0.5047) poradził sobie zaskakująco dobrze, co jest zasługą przeprowadzonej optymalizacji hiperparametrów (500 prób w Optuna). Pozwoliło to na znalezienie konfiguracji, która skutecznie unika przeuczenia (overfittingu), co widać po stabilnym spadku funkcji straty na zbiorze walidacyjnym.

W przypadku modelu dostrajanego, mimo lepszych wyników końcowych, zaobserwowano lekkie tendencje do przeuczenia w ostatnich epokach (wzrost straty walidacyjnej z 0.79 do 0.88), podczas gdy strata treningowa nadal malała.

Warto również zwrócić uwagę na koszty obliczeniowe. Model From Scratch jest znacznie mniejszy (3.4M vs 126M parametrów) i trenuje się błyskawicznie (niecałe 2 minuty vs 40 minut), a jego czas inferencji jest rzędu 1 ms na próbkę, co czyni go bardzo wydajnym rozwiązaniem w środowiskach o ograniczonych zasobach, mimo nieco niższej dokładności.