Plan eksperymentów

Michał Iwaniuk, Bartłomiej Borycki

November 2025

1 Rozpatrywane konfiguracje hiperparametrów

1.1 Architektura

Przyjmijmy nastepujace oznaczenia

- l^t – liczba warstw enkodera (liczba bloków Transformer).
- d^h – rozmiar wektora ukrytego (*hidden size*)
- d^f – rozmiar warstwy pośredniej w feed-forward network (FFN), czesto nazywany *intermediate size*. W klasycznym BERT zwykle wynosi $4 \cdot d^h$.
- h – liczba głowic w mechanizmie wielogłowicowej uwagi.
- $d^{q|k|v}$ – wymiar przestrzeni zapytań (*query*), kluczy (*key*) oraz wartości (*value*) w każdej głowicy uwagi. (W klasycznym BERT przyjmuje sie zazwyczaj $d^q = d^k = d^v = \frac{d^h}{h}$)

1.2 Wspólne parametry treningowe

Architektura: MLP dropout: 0.1, Embedding dropout: 0.1, Pozycje: `rope` (base=10000, scale=1.0). Atencja: projection_bias=true, attn_out_dropout=0.1, attn_dropout=0.0.

MLM head: tie_mlm_weights=true, mask_p=0.15, mask_token_p=0.8, random_token_p=0.1.

Classification head: num_labels=2, classifier_dropout=0.1, pooling=`cls`, pooler_type=`bert`.

Trening: batch=32 (`seq_len`=512) / 2 (`seq_len`=8192), lr=2e−5, warmup=0.1, wd=0.01, max_grad_norm=1.0, grad_accum=1, AMP=true, loss=cross_entropy.

1.3 Architektura Bazowa

Jako standardowa architekture małego modelu BERT przyjmiemy $BERT_{SMALL}$ opisana w artykule *Well-Read Students Learn Better: On the Importance of Pre-training Compact Models*

$BERT_{SMALL}$:

l^t	d^h	d^f	h	$d^{q\|k\|v}$
4	512	2048	8	512

1.4 Szukanie optymalnych parametrow

W oparciu o architekture bazowa przeprowadzimy eksperymenty majace na celu wyznaczenie optymalnych hiperparametrów dla mechanizmów *FAVOR+* oraz LSH, które zostana nastepnie wykorzystane w fazie właściwego treningu. Optymalizacja obejmie nastepujace parametry:

- **Mechanizm *FAVOR+*:** Optymalizacji poddany zostanie hiperparametr `nb_features`, określajacy liczbe ortogonalnych wektorów projekcyjnych. Rozpatrzone zostana wartości proporcjonalne do wymiaru d_k głowy atencji: $0.5d_k$, $1.0d_k$, $2.0d_k$ oraz $4.0d_k$.
- **Mechanizm LSH (*Locality-Sensitive Hashing*):** Analizie poddane zostana trzy kluczowe parametry:
  - **Rozmiar fragmentu (m):** Testowane wartości to 32, 64 oraz 128.
  - **Liczba funkcji haszujacych (`num_hashes`):** Sprawdzone zostana konfiguracje z 4 oraz 8 haszami.
  - **Maskowanie wewnatrz fragmentu (`mask_within_chunks`):** Parametr logiczny, gdzie:
    - `True`: zapytania (*queries*) wewnatrz fragmentu moga zwracać uwage (*attend*) wyłacznie na klucze z tego samego kubełka LSH w obrebie okna.
    - `False`: zapytania moga zwracać uwage na dowolny klucz w obrebie okna.

2 Architektura właściwa (do klasyfikacji)

Wykorzystamy architektury modeli wyłonione w artykule *AutoTinyBERT: Automatic Hyper-parameter Optimization for Efficient Pre-trained Language Models* jako podstawe do treningu właściwych modeli klasyfikacyjnych.

Na powyższych architekturach przeprowadzone zostana procesy uczenia dla

- **SDPA**,

l^t	d^h	d^f	h	$d^{q\|k\|v}$
5	564	1054	8	512
4	396	624	6	384
4	432	384	4	256
3	320	608	4	256

- **FAVOR** (z wykorzystaniem optymalnych parametrów wyznaczonych w poprzednim etapie),
- **LSH** (z wykorzystaniem optymalnych parametrów wyznaczonych w poprzednim etapie).

Dodatkowo, w celu stworzenia punktu odniesienia (*baseline*), wszystkie trzy wymienione typy atencji (SDPA, FAVOR, LSH) zostana wytrenowane również na architekturze bazowej $BERT_{SMALL}$.

Taki sposób przeprowadzenia eksperymentów ma na celu weryfikacje czy wskazane architektury faktycznie zapewniaja najkorzystniejszy stosunek szybkości działania do jakości predykcji również w przypadku metod FAVOR i LSH, analogicznie jak ma to miejsce w standardowych modelach BERT.

3 Trening

3.1 Korpus Wikipedii do pretreningu MLM

Wybrano podzbiór ok. 200 000 artykułów.
Wybierane dokumenty zawierajace przynajmniej jedno ze słów kluczowych:

```
["film", "sport", "business", "science", "technology", "news"]
```

3.2 TAPT + Fine-tuning(klasyfikacja): IMDB, AG News i korpus ArXiv

- **IMDB:** `seq_len`=512
- **AG News:** `seq_len`=512
- **ArXiv Classification:** `seq_len`=8192

4 Podsumowanie

- Każdy z trzech mechanizmów atencji (SDPA, FAVOR, LSH) zostanie pretrenowany w pieciu wariantach architektury (bazowy $BERT_{SMALL}$ oraz cztery konfiguracje z AutoTinyBERT). Co daje łacznie **15 modeli**.

- Po procesie TAPT + Fine-tuning trzymamy łacznie **45 finalnych modeli** , co przełoży sie na 45 wyników klasyfikacji podlegajacych późniejszej analizie.

‘