Zadanie 2: Modele językowe

Opracowanie: dr inż. Arkadiusz Janz, mgr inż. Piotr Miłkowski

Część I: Modelowanie języka (6 pkt)

A. Należy zapoznać się z implementacją następujących klas w bibliotece *Transformers* (https://huggingface.co/docs/transformers/index)
[0,25 pkt]:
- *BertModel,*
- *BertForMaskedLM,*
- *BertForSequenceClassification,*
- *BertForTokenClassification.*

B. Następnie, należy zapoznać się z biblioteką *PEFT* oraz przykładami z repozytorium
[0,25 pkt]:
- https://huggingface.co/docs/peft/quicktour
- https://github.com/huggingface/peft/tree/main/examples

C. Proszę również zapoznać się z tutorialem z następującego źródła i powtórzyć ten tutorial we własnym środowisku uruchomieniowym
[0,25 pkt]:
- https://medium.com/@nubyra/parameter-efficient-fine-tuning-peft-of-bert-base-model-to-predict-medical-diagnosis-5086a1828f4b

D. Na koniec proszę zapoznać się również z następującymi klasami w bibliotece *Transformers*
[0,25 pkt]:
- *GPT2Model,*
- *GPT2LMHeadModel,*
- *GPT2ForSequenceClassification,*
- *GPT2ForTokenClassification.*

Plan realizacji ćwiczenia

1. Należy wykonać strojenie modelu BERT-base na danych pozyskanych w ramach Zadania 1. Można również wykorzystać inne dane treningowo-testowe. Proszę wykorzystać model polskojęzyczny (np. `allegro/herbert-base-cased`). **Uwaga: proszę zastosować również adapter PEFT!!! [1,5pkt]**
    a. Wykonać strojenie dla zadania klasyfikacji tekstu (*BertForSequenceClassification*)
    b. Wykonać strojenie dla zadania klasyfikacji tokenów (*BertForTokenClassification*)
    c. Wydzielić niewielki zbiór testowy i ocenić jakość predykcji.

2. Zwizualizować przestrzeń wektorową dla przykładów testowych (**wystarczy jedynie klasyfikacja tekstu**) i zrobić interaktywny wykres przedstawiający to, w jaki sposób przypadki testowe organizują się w przestrzeni wektorowej względem przypisanych etykiet. **[1pkt]**

- W przypadku wykorzystania adaptera PEFT proszę wykorzystać reprezentacje tokenu [CLS] jako reprezentację tekstu.

3. Powtórzyć podpunkty 1 i 2 z wykorzystaniem klas *GPT2ForSequenceClassification i GPT2ForTokenClassification.* Proszę wykorzystać model polskojęzyczny (np. `sdadas/polish-gpt2-medium`) **Uwaga: proszę zastosować również adapter PEFT!!! [1pkt]**

- W przypadku wykorzystania adaptera PEFT proszę wykorzystać reprezentacje ostatniego tokenu sekwencji tekstowej jako reprezentację tekstu.
- Proszę również zamaskować tokeny paddingu i znacznik <eos> za pomocą etykietki -100.

4. Wykorzystać token **[MASK]** w modelu BERT do powiększenia zbiorów treningowych (ang. data augmentation). **[0,5pkt]**

5. Wykorzystać generatywne zdolności GPT-2 i różne hiperparametry generacji, np. *temperatura, top-p, top-k*, powiększenia zbiorów treningowych (ang. data augmentation). **[1pkt]**

- Uwaga! W podpunktach 4 i 5 można również połączyć wykorzystanie modelu BERT i GPT-2 w ramach zagadnienia powiększania zbiorów danych. Wystarczy wytrenować jeden model na powiększonym zbiorze danych.

Część II Analiza własności modeli językowych (4 pkt)

A. Na wstępie należy zapoznać się z następującym zbiorem danych
- https://huggingface.co/datasets/clarin-knext/wsd_polish_datasets

B. Proszę zapoznać się również z podanym artykułem **[0,25pkt]**:
- http://ai.stanford.edu/blog/contextual/
- https://aclanthology.org/D19-1006.pdf

C. Na koniec proszę zapoznać się z techniką *Parameter Projection* zaprezentowaną w niniejszym artykule **[0,25pkt]**:
- https://arxiv.org/pdf/2209.02535
- https://github.com/guy-dar/embedding-space

Plan realizacji ćwiczenia

1. Należy wykorzystać dołączony do zadania zbiór danych z podpunktu ***A*** i odtworzyć badanie z podpunktu ***B*** dla modeli BERT i GPT-2 . **[2pkt]**
    a. wykonać analizę anizotropii (Anisotropy)
    b. wykonać analizę zależności od kontekstu (Context-Specificity)

2. Proszę wykorzystać technikę *Parameter Projection* omówioną w artykule z podpunktu **C** niniejszej instrukcji, następnie dokonać analizy modeli BERT-base i GPT-2 **[1,5pkt]**
    a. do analizy należy wybrać modele polskojęzyczny `sdadas/polish-gpt2-medium`,
    b. proszę utworzyć listy z załączników C.1 oraz C.2 przedstawionych w artykule,
    c. należy przeanalizować utworzone listy i podsumować uzyskane rezultaty.