# Dostrajanie zapytania w problemie odpowiedzi na pytania na podstawie informacji wizualnej.

promotor dr inż. Jacek Komorowski

Emilia Zawadzka-Gosk

Numer albumu: 01183161

# Cel pracy

- Wskazanie fragmentu obrazu zawierającego odpowiedź na pytanie zadane w języku naturalnym.
- Zainspirowany „Toloka Visual Question Answering Challenge” https://toloka.ai/challenges/wsdm2023/

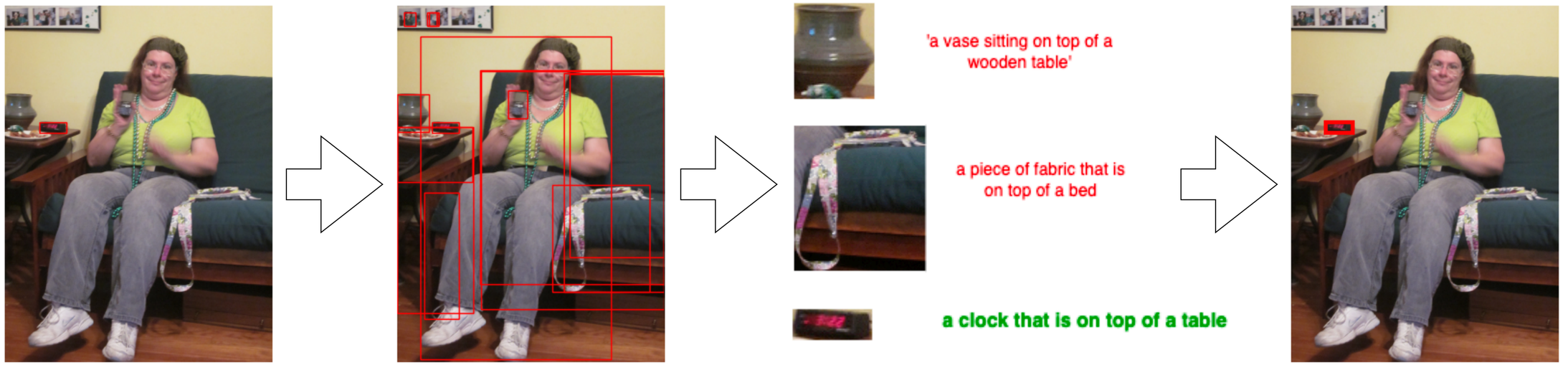
**Pytanie:** Which is different from the group?

**Koordynaty:** 409, 172, 432, 206

**Pytanie:** What do we drive for personal use?

**Koordynaty:** 161, 181, 569, 367

# Zaproponowane rozwiązanie




**Pytanie: „Where do we look to see time”**

1. Detekcja obiektów zrealizowana za pomocą modelu DETR

2. Opis obszarów wykrytych w poprzednim kroku za pomocą modelu VisionEncoder-DecoderModel

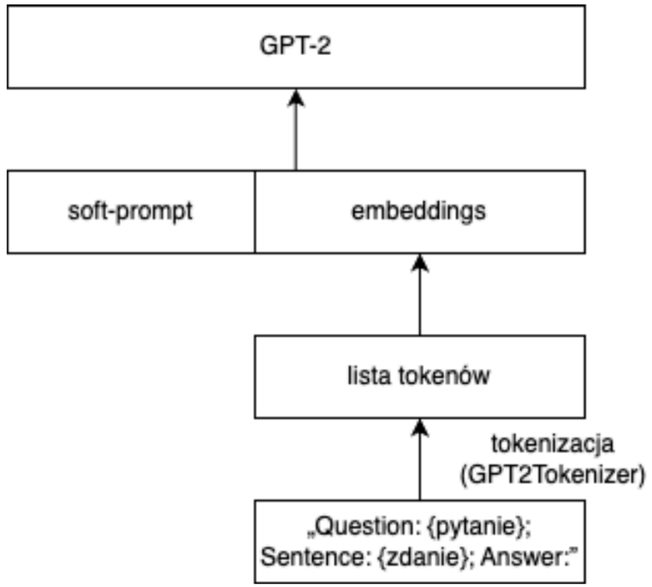
3. Ocena przez model językowy GPT-2

# Strategie tworzenia zapytań do modelu GPT-2

1. Stały prompt:
   *„Does the sentence contain the answer for the question?”*

2. Stały prompt zawierający przykłady:
   *„Does the sentence contain the answer for the question?*
   *Question: What animal is fluffy and furry?; Sentence: the cat sleeps on a windowsill; Answer:yes*
   *Does the sentence contain the answer for the question?*
   *Question: What can be used to cut bread?; Sentence: a person on a bike; Answer:no*
   *Does the sentence contain the answer for the question?”*

3. Soft-prompt:
   *Tensor o rozmiarze 10x1024*




Schemat konstruowania promptu z użyciem soft-promptu.

# Dostrojenie zapytania, a dostrojenie modelu

| | soft-prompt tuning | fine-tuning |
|---|---|---|
| Liczba parametrów | 10240 | 345M |
| Liczba epok | 3 | 1 |
| Czas trwania 1 epoki | 1 godz. | 3 godz. |
| Całkowity czas treningu | 3 godz. | 3 godz. |

Porównanie procesów dostrajania soft-promptu i całego modelu językowego GPT-2.

# Wyniki

| Zastosowane rozwiązanie | Dokładność na zbiorze walidacyjnym |
|---|---|
| Prefix bez przykładów | 0% |
| Prefix (2 przykłady) | 14% |
| Prefix (4 przykłady) | 18,7% |
| Dostrajanie zapytania (prompt-tuning) | 77% |
| Dostrajanie modelu (fine-tuning) | 81,9% |

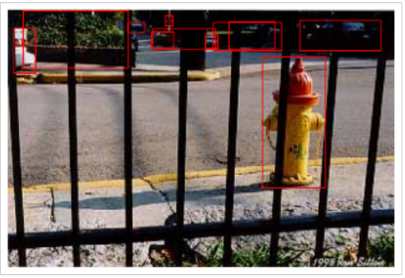
Zaproponowany w projekcie prompt-tuning uzyskał dokładność na zbiorze walidacyjnym o 5% niższą, jednak liczba parametrów w przypadku trenowania soft-promptu jest o cztery rzędy wielkości niższa, co ma istotny wpływ na zapotrzebowanie obliczeniowe eksperymentu.

# Wyniki

Detekcja obiektów (DETR)

Opis obiektów (VisionEncoderDecoderModel) oraz wybór właściwej odpowiedzi (GPT-2)

What is installed along side roads by fire department?



a white fire hydrant sitting on the side of a road

a yellow fire hydrant sitting on the side of a road

What can we use to see time?

a clock that is on a wall

a sign that is on the side of a building

# Dziękuję