Generowanie tekstu na podstawie obrazu

SEMINARIUM „PRZETWARZANIE JĘZYKA NATURALNEGO”

28.02.2022

MACIEJ CHRABĄSZCZ

Agenda

- Problem generowania tekstu i opisywania obrazów
- Zbiory danych
- Modele generujące opisy do obrazów
- Ewaluacja opisów
- Tworzenie opisów na podstawie wykrytych obiektów
- Wady aktualnego rozwiązania
- Plany rozwoju

Problem generowania tekstu

- W tym przypadku naszym celem jest utworzenie modelu, który generuje zróżnicowane, poprawne gramatycznie i spójne teksty.
- Zależy nam na utworzeniu modelu, który będzie tworzył teksty nieodróżnialne od napisanych przez ludzi.
- Modelem rozwiązującym ten problem jest np. GPT-3. Na podstawie takich modeli można potem je rozszerzać do wielu problemów, które wymagają wygenerowania tekstu.
- Ciekawym zastosowaniem jest **AI Dungeon**. Strona ta wykorzystuje modele przetrenowane na korpusie z internetu i na podstawie wpisanego przez nas tekstu generuje przygodę w świecie fantasy.

Generowanie opisów obrazów

- W tym przypadku naszą informacją, na podstawie której chcemy generować tekst, jest zdjęcie. Powstaje od razu pytanie, jak przekazać do sieci generującej informacje o tym obrazie.
- W celu wydobycia informacji z obrazu zazwyczaj wykorzystuje się sieci przetrenowane do klasyfikacji obrazów. Dzięki takiemu zabiegowi nasze obrazy są rzutowane na przestrzeń, która niesie o nich dużo informacji.
- Ułatwia to modelom generatywnym łatwiejszą generalizację, dla obrazów, z którymi nie miały styczności podczas treningu.

Pomoc dla niedowidzących

Naszym celem jest utworzenie rozwiązania, które dostosowuje dokumenty pod osoby niedowidzące. Zadaniem, którym ja się zajmuje jest generowanie opisów obrazów. Opis taki musi przekazać takiej osobie informację wystarczającą dla zrozumienia co się na obrazie znajduje. Aktualnie skupiamy się na języku angielskim ze względu na brak dużych zbiorów danych w języku polskim.

Dane

Niestety, dla języka polskiego nie znaleźliśmy sensownych zbiorów danych, który by pozwoliły na wytrenowanie modeli. W języku angielskim znaleźliśmy kilka zbiorów, natomiast żaden z nich nie był idealny.

COCO-Captions

- Common Objects in Context. Jest to zbiór posiadający opisy bardzo dobrej jakości i do każdego zdjęcia mamy 5 opisów.
- Zbiór danych składa się z ponad 330 tys. obrazów. Niestety, jak sama nazwa wskazuje, na tych zdjęciach są „częste” obiekty.
- Po przeanalizowaniu tego zbioru stwierdziliśmy, że nie jest on dla nas wystarczający, ponieważ potrzebujemy bardziej ogólnych zdjęć.

WIT: Wikipedia-based Image Text Dataset

- Jest to bardzo duży zbiór danych, w którym występują opisy w wielu językach. Powstał on na podstawie Wikipedii.
- Niestety, z niego też musieliśmy zrezygnować, ponieważ występuje w nim bardzo dużo nazw własnych i opisy, które odnosiły się bardziej do kontekstu, w którym dane zdjęcie występuje, a nie opisywały zdjęcia.
- Ze względu na to nasze modele widząc kościół zazwyczaj pisały, że jest to kościół św. Anny.

Conceptual Captions

- Zawiera on około 3 mln obrazów, które zostały pobrane razem z opisami z internetu. Natomiast został on poddany wstępnemu filtrowaniu, które np. zamienia nazwy własne na nazwę obiektu (np. Johnny Depp → Actor).
- Niestety, ten zbiór danych także nie jest idealny. Ze względu na to, że jest on stworzony na podstawie opisów znajdujących się w internecie, znajdują się w nim fragmenty opisów, które nie opisują zdjęcia w sposób odpowiedni dla osób niedowidzących.
- Ponieważ jest to zbiór najbliższy naszym oczekiwaniom, postanowiliśmy podjąć próby jego poprawy.
- Modele trenowane na tym zbiorze danych okazały się dla nas najlepsze i dobrze się generalizują na obrazy spoza zbioru treningowego.

Modele

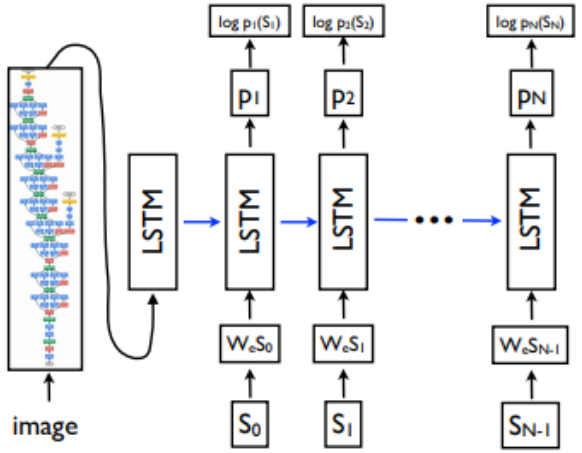
Podczas tworzenia rozwiązania przeanalizowaliśmy wiele podejść, ale nie wszystkie były przez nas przetestowane. Staraliśmy się skupić na najnowszych modelach w tej dziedzinie.

Jak reprezentować obraz

- Chcemy do modelu przekazać informację o obrazie, która niesie jak najwięcej informacji o nim.
- Można przekazać do modelu cały obraz po spłaszczeniu. Jest to bardzo nieoptymalne, ze względu na wielkość obrazów.
- Informacją przekazaną do modelu może być wyjście z sieci CNN, przetrenowanej do klasyfikacji obrazów.
- Wykorzystując sieci rozpoznające obiekty na obrazie (np. RCNN), możemy utworzyć ciąg tych wykrytych obiektów i taka sekwencja może nam posłużyć do reprezentacji obrazu.

Model z wykorzystaniem RNN

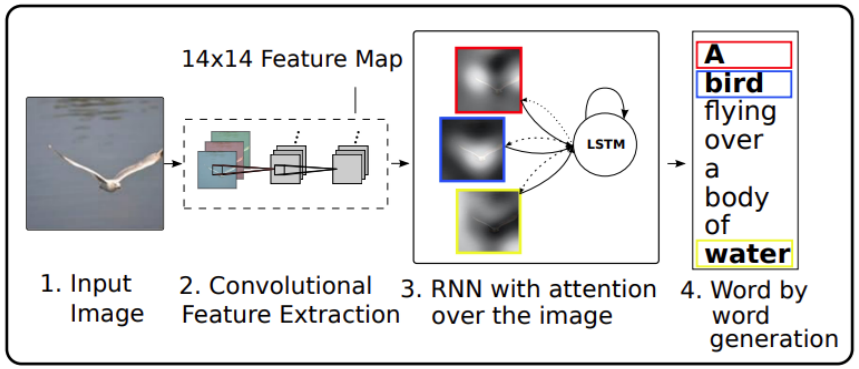
- W 2015 roku Google zaproponował, żeby jako pierwsze wejście do sieci rekurencyjnej wykorzystywać wyjście z sieci konwolucyjnej.
- Model ten ma problem z propagowaniem informacji o obrazie, ponieważ informacja ta ma problem z przechodzeniem do dalekich tokenów.

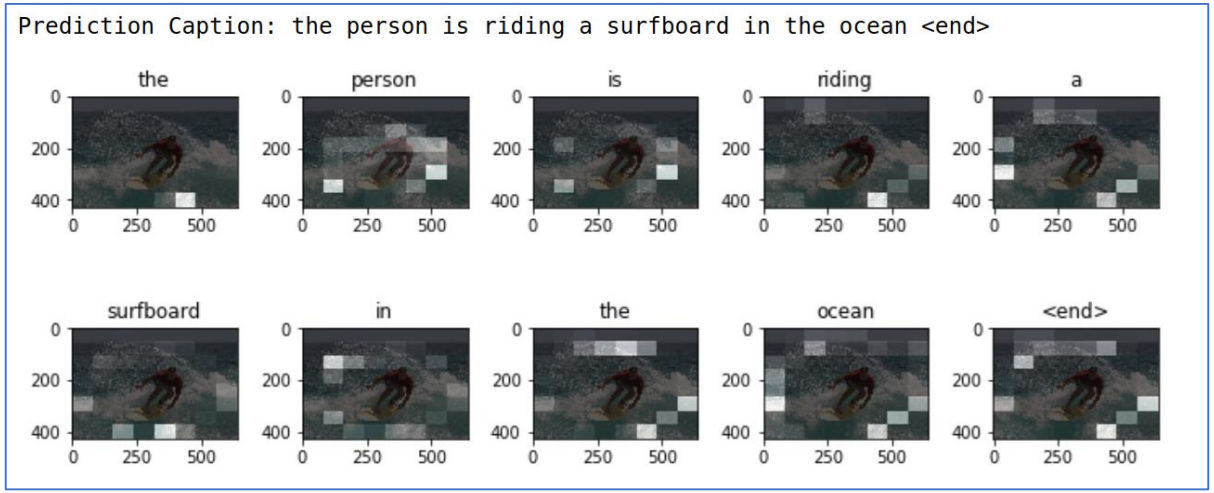


$$
\begin{aligned}
x_{-1} &= \mathrm{CNN}(I) \\
x_t &= W_e S_t, \quad t \in \{0 \dots N-1\} \\
p_{t+1} &= \mathrm{LSTM}(x_t), \quad t \in \{0 \dots N-1\}
\end{aligned}
$$

Modele z atencją

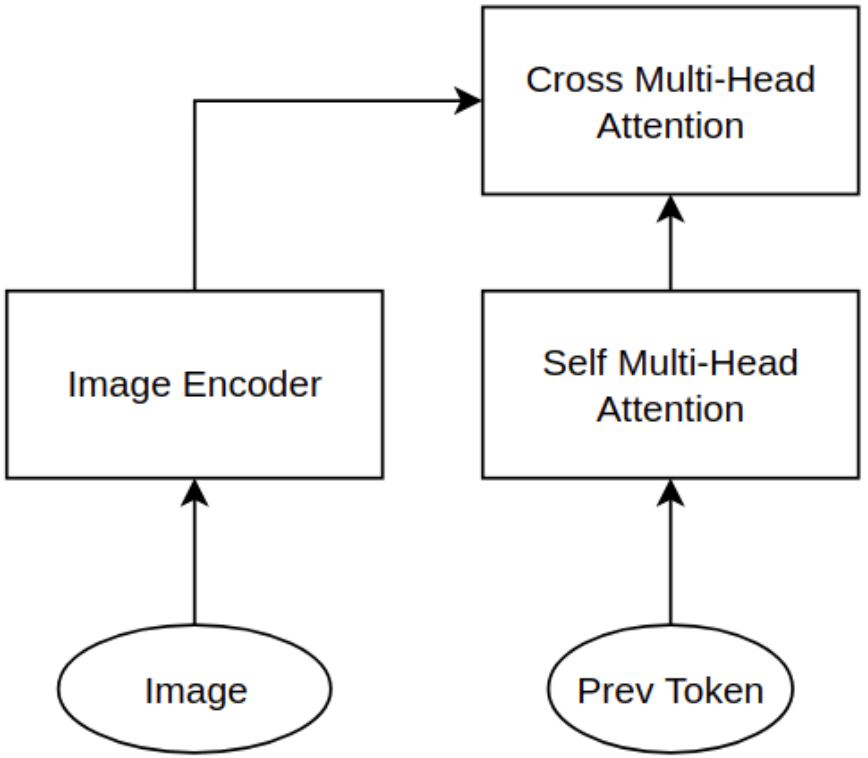
- Pierwszymi modelami przez nas testowanymi były modele składające się z sieci typu RNN z mechanizmem atencji.
- Atencja w tym przypadku jest wyliczana na podstawie wyjścia wcześniej przetrenowanej sieci konwolucyjnej.
- Dzięki temu model uczył się wybierać najpotrzebniejsze dla niego części obrazu. W ten sposób model ten nie ma problemu z zapominaniem informacji o obrazie, tylko na bieżąco wybiera informację, która jest mu aktualnie potrzebna.





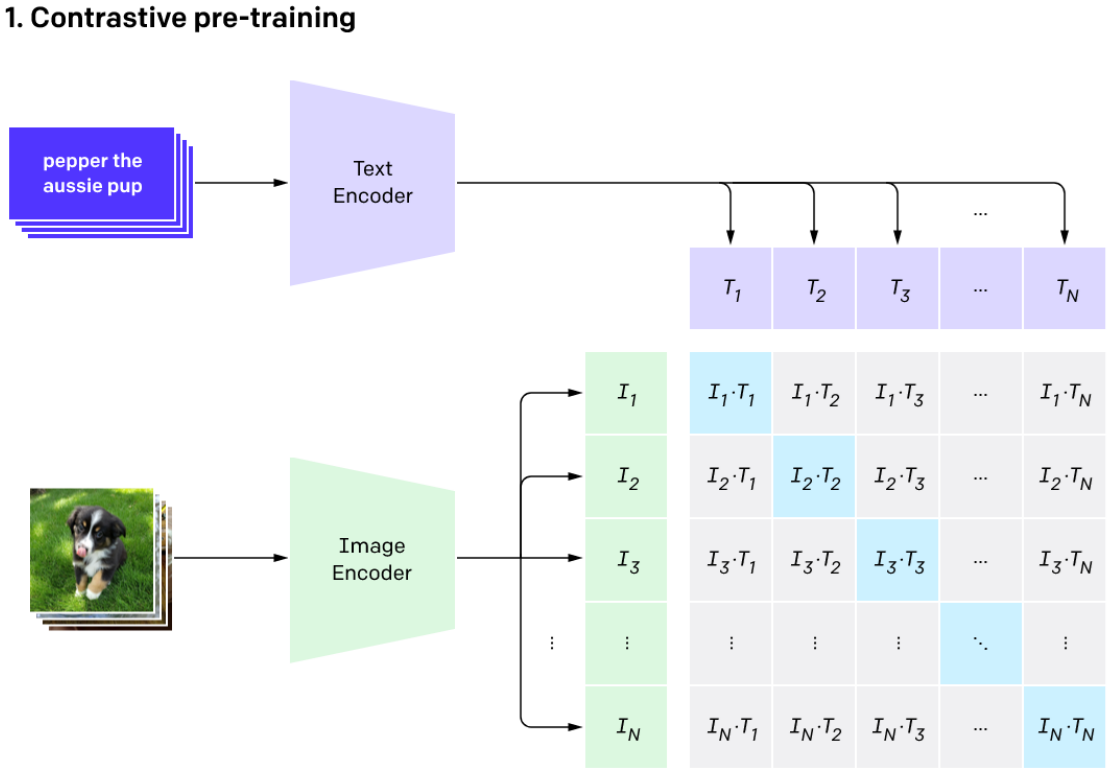
Model oparty o transformery

- W celu generacji opisów zaimplementowaliśmy sieć transformer, której encoder kodował wyjście z CNN, natomiast decoder wykorzystywał wyjście z encodera w celu generowania następnego tokenu.
- Przeprowadziliśmy też próbę pozbycia się modelu CNN i wykorzystania jako encoder modelu ViT (visual transformer).
- Model ten tak samo jak model z atencją nie ma problemu z przekazywaniem informacji o obrazie.
- Zaletą zastąpienia RNN przez Transformery jest przyśpieszenie uczenia.



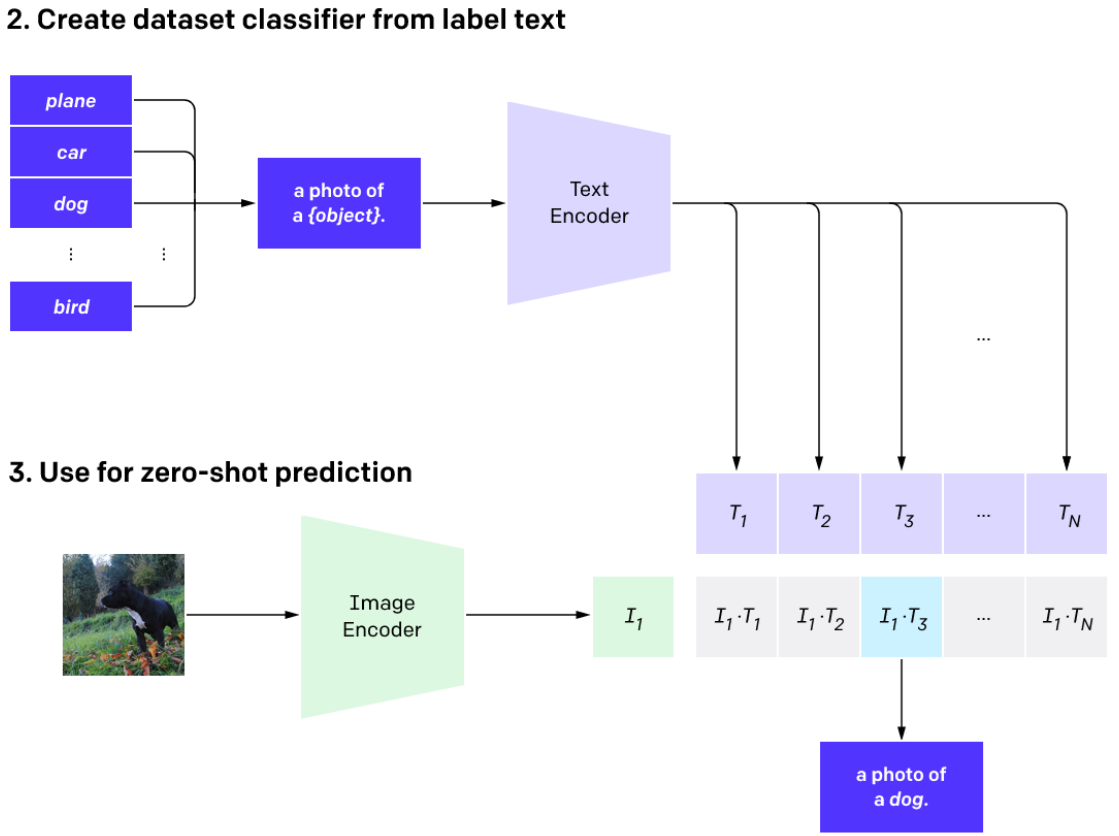
Model CLIP

- CLIP (eng. Contrastive Language–Image Pre-training) jest trenowany na podstawie zebranych z internetu obrazu oraz ich podpisów.
- Model ten jest trenowany do klasyfikowania czy dany tekst pasuje do podanego opisu.



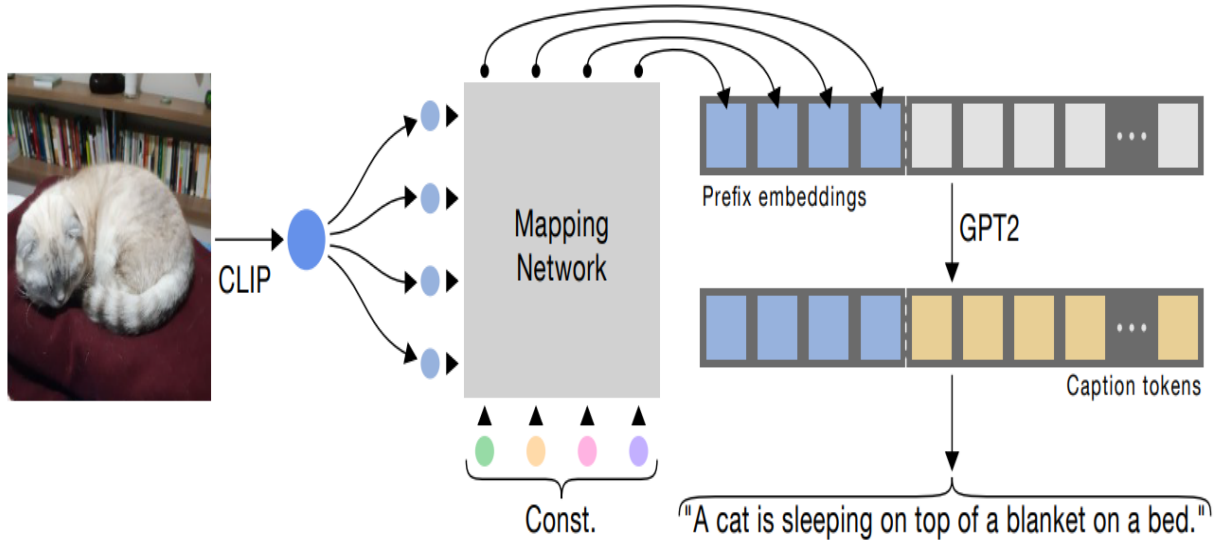
Zero-shot prediction

- Tak przetrenowany model można później wykorzystać np. do klasyfikacji obiektów na obrazie.
- Należy stworzyć „zbiór” tekstów odpowiadających obiektom.
- Przekazać tak przygotowane teksty i obraz do modelu, który zwróci prawdopodobieństwa.



Rzutowanie reprezentacji obrazu

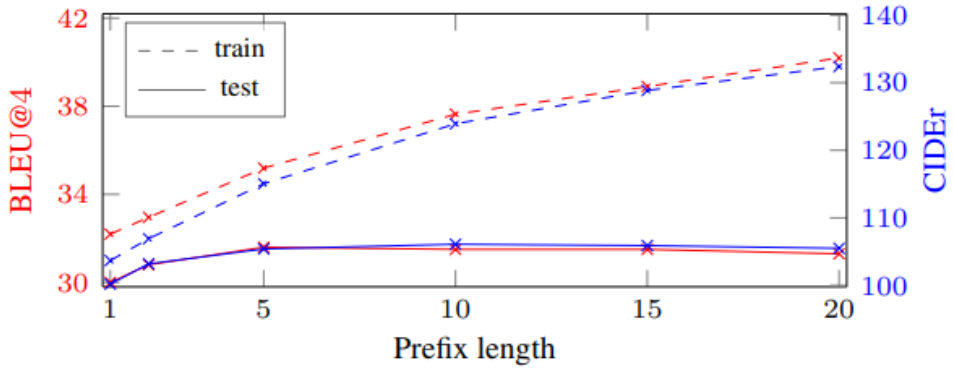
- Najnowszym pomysłem było wykorzystanie modelu CLIP, który był trenowany na podstawie tekstu i obrazów i GPT-2.
- Wzorując się na artykule naukowym embedding zdjęcia, który ten model zwraca przy użyciu perceptrona rzutowany był na przestrzeń embeddingów GPT-2.
- Tak przetrenowany model dał nam najlepsze wyniki. Kolejną zaletą było znaczne skrócenie uczenia (około 4x szybciej w porównaniu do modelu opartego o transformer).



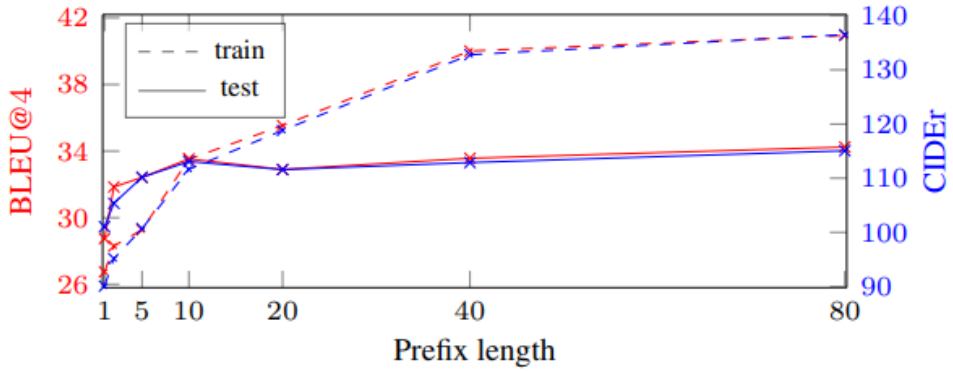
Interpretacja prefixu

						
GPT-2 tuning	Caption	a motorcycle is on display in a showroom.	a group of people sitting around a table.	a living room filled with furniture and a book shelf filled with books.	a fire hydrant is in the middle of a street.	display case filled with lots of different types of donuts.
	Prefix	com showcase motorcycle A ray motorcycle-posed what polished Ink	blond vegetarian dishes dining expects smiling friendships group almost	tt sofa gest chair Bart books modern doorway bedroom	neon street Da alley putistan colorful nighttime	glass bakery dough displays sandwiches2 boxes Prin ten
w/o tuning	Caption	motorcycle that is on display at a show.	a a group of people sitting at a table together.	a living room with a couch and bookshelves.	a fire hydrant in front of a city street.	a display case full of different types of doughnuts.
	Prefix	oover eleph SniperÃÂÃÂ motorcycle synergy undeniably achieving\n	amic Delicious eleph SukActionCode photographers interchangeable undeniably achieving	orianclassic eleph CameroonÃÂÃÂÃÂroom synergy strikingly achieving\n	ockets Pier eleph SniperÃÂÃÂ bicycl synergy undeniably achieving\n	peanuts desserts elephbmÃÂÃÂÃÂ cooking nodd strikingly achieving\n

Długość prefixu



(a) MLP mapping network with fine-tuning of the language model.



(b) Transformer mapping network with frozen language model.

Figure 7. Effect of the prefix length on the captioning performance over the COCO-captions dataset. For each prefix length, we report the BLEU@4 (red) and CIDERr (blue) scores over the test and train (dashed line) sets.

Trenowanie modeli

- Wszystkie modele były trenowane na podstawie par *zdjęcie–opis*. Podczas treningu modele starały się zminimalizować krzyżową entropię.
- Model otrzymywał na wejściu zdjęcie i wszystkie tokeny poza ostatnim, a jego zadaniem było przewidzenie poprawnie następnego tokenu.

Ewaluacja wygenerowanych opisów

W jaki sposób ocenia opisy wygenerowane przez model? Można zastosować metryki takie jak BLEU, ROUGE, METEOR, CIDEr.

Ewaluacja ludzka

- Zauważyliśmy, że wcześniej wymienione metryki nie do końca odzwierciedlają, czy opis jest dobry dla osoby niedowidzącej.
- Wszystkie wcześniej podane metryki są oparte na n-gramach, przez co mogą przyjąć dużą/małą wartość mimo złego/dobrego opisywania obrazu.
- Ze względu na to nasze modele są oceniane przez człowieka w skali 1–5 (kompletnie nie pasujący – idealnie pasujący).

Generowanie opisów na podstawie wykrytych obiektów

W celu utworzenia modelu generującego opisy można też wykorzystać modele wykrywające obiekty na obrazie.

RCNN a opis

- Korzystając z modelu RCNN możemy z obrazu wyciągnąć obiekty znajdujące się na obrazie razem z ich pozycjami.
- Mając pozycje obiektów i ich pozycje możemy utworzyć opis schematyczny wypisujący ile obiektów i gdzie pojawiły się na obrazie.
- Zrezygnowaliśmy z takiego podejścia, ze względu na to że jego uogólnienie wymaga modelu RCNN, który wykrywa wiele różnych obiektów.
- Niestety nie byliśmy w stanie znaleźć modelu wykrywającego bardzo ogólne klasy.
- Jest to sposób Facebooka, z którego korzystają do tworzenia tekstów alternatywnych w swoim serwisie.

Przeniesienie rozwiązania na inne języki

- Aktualne modele generują opisy w języku angielskim. Co zrobić, aby rozszerzyć to rozwiązanie na inne języki?
- Ze względu na brak danych w języku polskim naszymi pomysłami jest automatyczne tłumaczenie z angielskiego na polski wygenerowanych opisów lub automatyczne przetłumaczenie zbioru danych i na tym zbiorze danych przetrenowanie modelu generującego opisy w języku polskim.

Wady aktualnego rozwiązania

- Nie jest przez nas wykorzystywany kontekst w jakim występuje zdjęcie, które opisujemy.
- W przypadku pojawienia się obrazu z bardzo specyficznej dziedziny (np. obraz z mikroskopu) nasz model raczej generuje błędny opis.
- Ze względu na niedoskonałość zbioru Conceptual Captions model jest trenowany z pewnym szumem, który utrudnia nauczenie się dobrego opisywania zdjęć.

Plany rozwoju modelu

- Dodanie do procesu uczenia dyskryminatora oceniającego dopasowanie opisu do zdjęcia i wykorzystanie funkcji straty z Reinforcement Learning w celu poprawienia jakości modelu.
- Stworzenie modelu oceniającego zdjęcia w skali 1–5 na podstawie ocen anotatorów.
- Poprawa zbioru danych przez lepsze przefiltrowanie lub pobranie większej liczby obrazów z opisami.

Bibliografia

- Show and Tell: A Neural Image Caption Generator
- Show, Attend and Tell: Neural Image Caption Generation with Visual Attention
- An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
- Model CLIP
- Coco Captions
- WIT
- Conceptual Captions