PREZENTACJA

1. ADAM – Slajd Tytułowy`  
     
   Dzień dobry – prezentację chcielibyśmy zacząć od podziękowań za naukę podczas tegorocznych studiów podpylomowych w tych ciężkich, zdalnych czasach. Tematem naszego projektu jest szeroko rozumiane pojęcie text-to-image generation. W obecnych czasach znaleźć możemy różnorodne sposoby podjęcia tego tematu – od GANów, poprzez transformery i inne rozwiązania state-of-the-art. W naszej prezentacj postaramy się przybliżyć najważniejsze elementy testowanego przez nas rozwiązania, a także wyniki eksperymentów, który przy użyciu tego rozwiązania przeprowadziliśmy.
2. ADAM – Dall-e

Główną motywacją dla podjęcia tego tematu było opublikowanie przez firmę OpenAI w 2021 publikacji dotyczącej modelu Dall-e czyli ich największego modelu pozwalającego na generowanie obrazów na podstawie tekstu. Autorzy opisują ten model jako model o 12-miliardach parametrów, który potrafi nie tylko generować obrazy przedstawiające rzeczywiste obiekty, ale też tworzyć abstrakcyjne twory, które mogą łączyć w sobie wiele elementów jak widoczne na przykładzie krzesło w kształcie awokado. W dużym skrócie – model dall-e składa się z 2 elementów – części generatywnej, która nie została na ten moment upubliczniona, a także części odpowiedzialnej za powiązanie tekstu z obrazem nazwanej przez autorów CLIP. Ta część modelu została upubliczniona i odniesiemy się do niej w dalszej części prezentacji.

1. MACIEK – GANy  
     
   …
2. ADAM – CLIP

Jak wspominaliśmy wcześniej jednym z elementów tworzących model DALL-E jest CLIP czyli Contrastive Language-Image Pre-Training. Jest to sieć neuronowa wyuczona na ponad 400 milionach par obraz-podpis, które zostały pobrane z internetu przez autorów. Służy on do zambedowania obrazu oraz podpisu w taki sposób by otrzymane wektory przechowywały jak najwięcej informacji z obu elementów. W naszym rozwiązaniu wykorzystujemy propowany w CLIPie sposób porównywania embeddingów wygenerowanego przez model zdjęcia, wraz z zaembedowanym podanym przez użytkownika zapytaniem przy użyciu podobieństwa cosinusowego.   
Na zdjęciu widzimy przykład użycia CLIPa jako algorytmu klasyfikacyjnego – podając mu zdjęcie oraz listę potencjalnych podpisów możemy znaleźć najlepszy z nich.

1. MACIEK – Algorytmy ewolucyjne

…

1. MACIEK – Framework – opis

…

1. ADAM – Framework – przykład  
     
   Kiedy omówiliśmy już nasz framework – przeanalizujmy prosty przykład wykonania naszego algorytmu. Zaczynamy od konfiguracji algorytmu – w tym przypadku postanowiliśmy wybrać algorytm StyleGAN2 w wersji dla generowania twarzy, dodatkowo algorytm GA, oczywiście algorytm CLIP i ustalamy 200 iteracji algorytmu.  
   Zadanym przez nas sformułowaniem jest „a blond girl with a smile”. Na 1 zdjęciu możemy zobaczyć batch z modelu, który otrzymaliśmy po 100 epoce – widzimy już, że przy tak prostym przykładzie dla „StyleGAN2” większość zdjęć przypomina bez problemu blondwłosą, uśmiechniętą kobietę. Zdjęcie ostateczne oddaje nam najlepsze zdjęcie z algorytmu genetycznego na podstawie metryki jaką jest cos-sim.

1. MACIEK – GA vs DE

…

1. MACIEK – StyleGAN vs BigGAN   
     
   …
2. ADAM – Ewaluacja  
     
   Pomijając pojedyncze eksperymenty, które miały nam odpowiedzieć na postawione przez nas pytania, postanowiliśmy również sprawdzić jak testowane przez nas rozwiązanie sprawdzi się w szerszej skali. W tym celu postanowiliśmy przeprowadzić ewaluację, którą zasugerował nam paper dotyczący modelu CLIP. Mianowicie: dla dwóch popularnych datasetów – CIFAR10 oraz ImageNet znaleźliśmy wyuczone na nich klasyfikatory o jak największej skuteczności, następnie wygenerowaliśmy dla 10 wybranych klas z obu datasetów po 512 zdjęć, podaliśmy je do klasyfikatorów i obliczyliśmy jaki % wygenerowanych przez nas zdjęć potrafił oszukać klasyfikatory by uznały nasze zdjęcia za podobne do tych ze zbioru treningowego. Wyniki prezentują się następująco:  
   - Pierwszy dataset – CIFAR10 wybraliśmy ze względu na małą liczbę klas, którą posiada – 10. Otrzymany średni % accuracy na wygenerowanych przez nas zdjęciach nie przekroczył 25. Analizując pokazywaną macierz błędów doszliśmy do wniosków, że zwierzęta były często mylone ze sobą (zagęszczenie na środku), a maszyny ze sobą (airplane, automobile).   
   - Drugim datasetem był ImageNet – z jednej strony w porównaniu do CIFAR10 zawiera on zdecydyowanie więcej klas – ponad 2000, z drugiej strony, model BigGAN, który wykorzystujemy był na nim uczony, więc cięzko było powiedzieć, czy wyniki będą prezentować się lepiej czy gorzej niż dla CIFARA. Są one następujące – nie przekroczyliśmy 23% skuteczności, ale zważywszy na poziom skomplikowania zadania, odbieramy ten wynik pozytywnie – co piąte zdjęcie wygenerowane przez nas na podstawie nazwy klasy pozwalało oszukać klasyfikator. Przykłady dla wszystkich klas, które wygenerowaliśmy w obu datasetach ewaluacyjnych można znaleźć w załączniku do naszej pracy – tutaj przedstawiamy przykładowy z nich czyli 5 zdjęć, które oszukały klasyfikator dla klasy LLAMA i 5 które nie dały rady go oszukać.
3. MACIEK – Wnioski  
     
   … W niedalekiej przyszłości planujemy przeprowadzenie dodatkowych eksperymentów i opublikowanie ich w formie artykułu na platformie medium.  
   - problem z kontkstem w zapytaniach np. kolor poszczególnego obiektu

UWAGI Z CALLA:

Siema, w prezentacji omówimy temat text-to-image generation…,  
- co omówimy  
- motywacja/geneza

1. DALL-E: ADAM

- OpenAI wypuscili GPT3+CLIP, tylko CLIP jest udostepniony, jest to jedno z wielu podejść bla bla, zero-shot classifier, dużo parametrów, state-of-the-art

1. GANy (po 2-3 kropki – kto stworzyl, do czego sluzyl, na czym wyuczone) MACIEK  
   - zastępuje czesc generacyjna w Dall-e, my uzywamy przykladow state-of-the-art. – StyleGAN2 i BigGAN – stylegan skupia się na konkretnych srodowiskach, Biggan lepiej generalizuje z zalozenia
2. CLIP (ten z pracki) (punkty) ADAM
3. Algorytmy ewolucyjne (po 2 kropki) MACIEK  
   -
4. Framework MACIEK