Wizualizacja z wykorzystaniem generatywnej sztucznej inteligencji (AI)

Tu omówiono wizualizację z wykorzystaniem generatywnej sztucznej inteligencji (AI), która obejmuje popularne narzędzia do generowania obrazów wprowadzone w 2022 i 2023 roku. Dowiesz się również o AI21 i jego platformie, a także o platformie Aleph-Alpha Playground, które oferują funkcjonalność podobną do platformy GPT-3 Playground. Przykłady kodu oparte na Pythonie wymagają Pythona 3.7, który można pobrać z internetu, jeśli jeszcze go nie masz. Pierwsza sekcja zawiera przegląd generatywnej sztucznej inteligencji (AI) i kilka przykładów jej zastosowania. Następnie poznasz dyfuzję, czyli technikę usuwania szumu z obrazu, a także jej przydatność w modelach infuzji. Dowiesz się również o modelach dyfuzji i ich porównaniu z sieciami GAN (generatywnymi modelami przeciwstawnymi). W tej sekcji przedstawiono CLIP i GLIDE, oba pochodzące z OpenAI. Druga część zawiera przegląd kilku popularnych narzędzi do generowania tekstu na obraz, takich jak Stability AI i Imagen. Dowiesz się również o Make-a-Scene i GauGAN2 oraz ograniczeniach modeli przetwarzania tekstu na obraz. Trzecia część omawia modele DALL-E, płatne konta DALL-E oraz sposób wywoływania API DALL-E. Dodatkowo poznasz różne szczegóły dotyczące DALL-E 2 oraz informacje o DALL-E 3. Ostatnia część omawia narzędzia do generowania tekstu na wideo, takie jak Make-a-Video (Meta), Text-to-Video (Imagen), a także narzędzie do generowania tekstu na mowę o nazwie Whisper firmy OpenAI.

GENERATYWNA SZTUKA, SZTUKA I PRAWA AUTORSKIE

Sztuka generatywna została pokrótce opisana wcześniej, wraz z opisem różnic między nią a sztuczną inteligencją konwersacyjną. Co ważniejsze, sztuczna inteligencja generatywna udowodniła swoje możliwości w wielu obszarach, takich jak filmy i wideo. Jednym z porównań jest wpływ CGI w latach 90. XX wieku i jego wpływ na hollywoodzkie filmy. Sztuczna inteligencja generatywna w radykalny sposób wykracza poza CGI, co doprowadzi do przełomowych zmian w wielu branżach. Osiągnęliśmy już punkt, w którym sztuczna inteligencja generatywna może tworzyć sztukę, a nawet wygrywać konkursy artystyczne. Nasuwa się naturalne pytanie: Kto jest właścicielem dzieła sztuki? Choć uzyskanie praw autorskich lub patentu na swoją pracę przez sztuczną inteligencję mogłoby wydawać się sprawiedliwe, sędzia orzekł, że tylko ludzie mogą uzyskać prawa autorskie do swojej pracy.

GENERATYWNA SZTUKA I GANS

Chociaż generatywna sztuczna inteligencja i generatywne sieci przeciwstawne (GAN) są blisko spokrewnionymi koncepcjami w dziedzinie uczenia maszynowego i głębokiego uczenia, istnieją między nimi różnice. Generatywna sztuczna inteligencja to szerokie pojęcie, obejmujące każdy model uczenia maszynowego zaprojektowany w celu generowania nowych próbek danych. Obejmuje to sieci GAN, sieci VAE i sieci RBM. Zarówno generatywna sztuczna inteligencja, jak i GAN odegrały kluczową rolę w różnych zastosowaniach, od syntezy obrazu, rozszerzania danych, transferu stylu, po odkrywanie leków i inne. Generatywna sztuczna inteligencja odnosi się do szerszej klasy modeli uczenia maszynowego, których celem jest generowanie nowych próbek podobnych do danych wejściowych. Modele te uczą się rejestrować rozkład danych treningowych, dzięki czemu mogą generować nowe, syntetyczne przykłady, zgodne z danymi, które już widziały. Istnieje kilka typów generatywnych modeli sztucznej inteligencji, z których niektóre wymieniono poniżej wraz z krótkim opisem:

- Probabilistyczne modele graficzne
- Autoenkodery wariacyjne (VAE)
- Ograniczone maszyny Boltzmanna (RBM)
- Generatywne sieci przeciwstawne (GAN)

Probabilistyczne modele graficzne to modele statystyczne, które wykorzystują graf do reprezentowania i mapowania zależności między różnymi zmiennymi losowymi w danych. Przykładami są ukryte modele Markowa i sieci bayesowskie.

Autoenkodery wariacyjne (VAE) to architektury sieci neuronowych, które uczą się kodować i dekodować dane w taki sposób, że zakodowane reprezentacje mogą być wykorzystane do generowania nowych, podobnych danych. Ograniczone maszyny Boltzmanna (RBM) to sieci neuronowe z dwiema warstwami (widoczną i ukrytą), które uczą się probabilistycznego modelu danych wejściowych.

Generatywne sieci przeciwstawne (GAN) to rodzaj modelu generatywnego, który wykorzystuje dwie sieci neuronowe zwane generatorem i dyskryminatorem.

GAN (Generative Adversarial Networks)

GAN to specyficzny typ generatywnej sztucznej inteligencji (AI), wprowadzony przez Iana Goodfellowa i jego współpracowników w 2014 roku. GAN to specyficzny typ generatywnego modelu AI, charakteryzujący się procesem trenowania przeciwstawnego, obejmującym generator i sieć dyskryminatora. Unikalną cechą GAN jest proces trenowania przeciwstawnego, który obejmuje dwie sieci neuronowe zwane generatorem i dyskryminatorem. Generator próbuje wygenerować dane w oparciu o losowy szum jako dane wejściowe, a następnie generuje próbki jako dane wyjściowe. Natomiast dyskryminator próbuje odróżnić dane autentyczne (z zestawu treningowego) od danych fałszywych wygenerowanych przez generator. Oto kroki procesu uczenia:

• Generator stara się generować fałszywe dane, które wyglądają jak najbardziej realistycznie.

• Dyskryminator stara się lepiej odróżniać dane prawdziwe od fałszywych.

• Proces ten jest podobny do fałszerza próbującego stworzyć fałszywy obraz, podczas gdy detektyw sztuki próbuje wykryć, który z nich jest fałszywy. Z czasem fałszerz staje się tak biegły w tworzeniu obrazów, że detektyw nie potrafi odróżnić oryginału od podróbki.

Zalety sieci GAN obejmują:

• Mogą generować dane o bardzo wysokiej jakości, zwłaszcza obrazy.

• Nie wymagają jawnego modelowania rozkładu danych.

Wyzwania związane z sieciami GAN obejmują:

• Uczenie może być niestabilne i wrażliwe na wybór hiperparametrów.

• Mogą one paść ofiarą załamania modów, gdzie generator generuje ograniczoną liczbę próbek.

CZYM JEST DYFUZJA?

W istocie dyfuzja w generatywnej sztucznej inteligencji polega na symulowaniu procesu, w którym dane są stopniowo oczyszczane z szumu do realistycznych próbek. Proces ten jest inspirowany ideą dyfuzji, gdzie elementy rozprzestrzeniają się i mieszają w czasie, ale jest stosowany w kontekście generowania danych. Dyfuzja w kontekście generatywnej sztucznej inteligencji nie odnosi się do biernego ruchu cząsteczek w sensie fizycznym lub biologicznym. Zamiast tego często odnosi się do modeli dyfuzyjnych, które są klasą modeli generatywnych symulujących proces dyfuzji w celu generowania nowych próbek danych. Poniższa lista zawiera główne aspekty dyfuzji w generatywnej sztucznej inteligencji:

• Modele dyfuzyjne

• Proces sterowany szumem

• Oznaczanie dopasowania wyników

• Proces generatywny

Modele dyfuzyjne generują nowe próbki danych poprzez symulację procesu dyfuzji. Zaczynając od losowej próbki, iteracyjnie oczyszczają ją za pomocą serii aktualizacji szumu, aż stanie się próbką z docelowego rozkładu danych. Proces sterowany szumem odnosi się do faktu, że modele dyfuzyjne dodają i usuwają szum z/do próbek danych. Dokładniej, proces rozpoczyna się od rzeczywistego punktu danych, a następnie dodaje szum w serii kroków, aż do uzyskania czystego szumu losowego. Następnie proces generatywny polega na odwróceniu tego dodawania szumu, aby przejść od szumu losowego z powrotem do próbki przypominającej rzeczywiste dane. Technika dopasowywania wyników odnosi się do sposobu, w jaki model uczy się przewidywać różnicę (lub „wynik") między danymi zaszumionymi a danymi oryginalnymi. Poznając ten wynik, model rozumie, jak odszumiać dane, co jest przydatne w procesie generatywnym. Krok procesu generatywnego to krok wnioskowania, który ma miejsce po wytrenowaniu modelu. W tym kroku model dyfuzyjny rozpoczyna od losowej próbki szumu, a następnie iteracyjnie udoskonala tę próbkę, wykorzystując poznane wyniki odszumiania w ciągu kolejnych kroków. Proces ten stopniowo przekształca szum w próbkę, która przypomina jedną z próbek z docelowego rozkładu danych. Modele dyfuzyjne mogą generować wysokiej jakości próbki i wykazały, że działają konkurencyjnie w porównaniu z innymi modelami generatywnymi, takimi jak GAN i VAE. Podobnie jak inne modele generatywne, modele dyfuzyjne mogą być wykorzystywane do generowania danych, syntezy obrazów, malowania w obrazie, uzyskiwania superrozdzielczości i innych zadań wymagających generowania nowych próbek danych.

Przykład obrazu dyfuzyjnego

Poniższy adres URL to aplikacja online do generowania obrazów dyfuzyjnych z danego obrazu: https://replicate.com/tommoore515/material_stable_diffusion. Rysunek 1 przedstawia zawartość pliku sample3.tiff, który jest określony jako obraz wejściowy dla poprzedniego linku, a rysunek 2 przedstawia wygenerowany obraz dyfuzyjny.

Modele dyfuzyjne kontra sieci GAN

Według artykułu opublikowanego przez OpenAI w czerwcu 2021 r. pod tytułem „Modele dyfuzyjne pokonują sieci GAN w syntezie obrazu", modele dyfuzyjne mogą osiągać lepszą jakość obrazu niż modele generatywne (choć mają pewne ograniczenia).

Czym są dyfuzory i DDPM?

Termin „diffusers" odnosi się do modeli dyfuzji SOTA (napisanych w języku PyTorch), które są dostępne dla więcej niż jednej modalności, na przykład generowania obrazu i dźwięku. Probabilistyczne modele denoisingu dyfuzji (DDPM) to głębokie modele generatywne, które stanowią podstawę generowania obrazów w DALL-E 2 i Imagen

CLIP (OPENAI)

CLIP to akronim od Contrastive Language-Image Pre-training (model opracowany przez OpenAI). Chociaż sam w sobie nie jest modelem generatywnym, został zaprojektowany do rozumienia obrazów w połączeniu z językiem naturalnym, wypełniając lukę między zadaniami związanymi z widzeniem a językiem. CLIP to model zaprojektowany w celu zrozumienia i wypełnienia luki między widzeniem a językiem. Ucząc się kojarzenia obrazów z opisami w języku naturalnym, CLIP osiąga wszechstronne rozumienie, które można zastosować do różnych zadań, zarówno w kontekście wizji, jak i w scenariuszach, w których jest ono łączone z modelami generatywnymi. CLIP obejmuje następujące komponenty:

• Mechanizm uczenia

• Uczenie kontrastowe

• Uczenie bezstratne

Mechanizm uczenia obejmuje duży zestaw obrazów sparowanych z opisami tekstowymi. Celem jest zapewnienie, aby obraz i odpowiadający mu opis tekstowy zbliżyły się do siebie w przestrzeni osadzonej, podczas gdy inne, niepasujące pary zostaną rozsunięte. Uczenie kontrastowe stanowi podstawę mechanizmu uczenia. W tym podejściu model uczy się rozróżniać pary dodatnie (poprawne dopasowania obraz-tekst) i pary ujemne (losowe kombinacje obraz-tekst).

Uczenie bezstratne to cecha, dzięki której CLIP może wykonywać zadania bez konieczności precyzyjnego dostrajania. Mając zestaw klas opisanych w języku naturalnym, CLIP może klasyfikować nowe obrazy do tych klas bez konieczności wykorzystywania danych treningowych specyficznych dla

danego zadania dla każdej klasy. CLIP może obsługiwać szeroki zakres zadań wizyjnych, używając tego samego modelu. Obejmuje to klasyfikację obrazów, wykrywanie obiektów, a nawet niektóre formy generowania obrazów w połączeniu z innymi modelami generatywnymi. Chociaż CLIP nie jest modelem generatywnym, można go łączyć z modelami generatywnymi w celu uzyskania interesujących zastosowań. Na przykład, w połączeniu z modelem takim jak VQ-VAE-2 (model generatywny), CLIP może być używany do sterowania procesem generowania za pomocą opisów tekstowych, co skutkuje systemem, który może generować obrazy z komunikatów tekstowych. Chociaż CLIP został zaprojektowany z myślą o odporności, może być podatny na zaawansowane ataki, ale generalnie oferuje lepszą odporność w porównaniu z tradycyjnymi modelami. CLIP zapewnia również transferowalność, ponieważ jego rozumienie obrazów i tekstu pozwala CLIP na transfer wiedzy do szerokiego zakresu zadań bez konieczności precyzyjnego dostrajania.

GLIDE (OPENAI)

GLIDE to akronim od Guided Language-to-Image Diffusion for Generation and Editing (Guided Language-to-Image Diffusion do generowania i edycji) i składa się z osadzenia CLIP oraz modelu dyfuzji, który korzystnie wypada w porównaniu z DALL-E, mimo że GLIDE ma mniej niż jedną trzecią rozmiaru DALL-E. GLIDE koduje komunikaty językowe za pomocą transformatora tekstu, a oprócz tekstu, GLIDE może przetwarzać obrazy modyfikowane za pomocą komunikatów NLP w celu wstawiania nowych obiektów. GLIDE może również ulepszać obrazy za pomocą różnych innych efektów, takich jak odbicia i cienie. W rzeczywistości GLIDE jest preferowany przez niektórych ze względu na zestaw funkcji.

Teraz, gdy rozumiesz już CLIP i GLIDE, przyjrzyjmy się narzędziom do generowania tekstu na obraz, co jest tematem następnej sekcji.

GENEROWANIE TEKSTU NA OBRAZ

Generowanie tekstu na obraz jest obecnie niezwykle innowacyjne dzięki dostępności nowych narzędzi do generowania obrazów, takich jak DALL-E, Craiyon i Stable Diffusion. Inne narzędzia są w fazie rozwoju, a wyścig o lepszą obsługę funkcji trwa. Generowanie obrazów przeżywa renesans, który będzie miał głęboki wpływ na artystów, projektantów i firmy dostarczające narzędzia i produkty graficzne. Wraz z sukcesem generowania tekstu na obraz pojawiły się pewne kontrowersje, na przykład dotyczące praw autorskich. Na przykład Getty Images udostępnia bibliotekę zawierającą prawie 500 milionów obrazów i zakazał przesyłania do swojej kolekcji obrazów generowanych przez sztuczną inteligencję z powodu obaw o legalność takich obrazów. Inne strony, które wprowadziły podobny zakaz, to Newgrounds i PurplePort. Inny kontrowersyjny incydent dotyczył konkursu sztuk pięknych, w którym przyznano nagrodę za dzieło sztuki wygenerowane przez sztuczną inteligencję. Wśród artystów i osób zajmujących się grafiką interfejsów użytkownika narasta również niezadowolenie dotyczące potencjalnie negatywnego wpływu grafiki i projektowania opartego na sztucznej inteligencji na ich karierę. Należy pamiętać, że niektóre narzędzia do generowania obrazów, takie jak Craiyon i DALL-E, są dostępne za pośrednictwem wywołań API lub interfejsu internetowego, natomiast Stable Diffusion można pobrać na komputer. Repozytorium Github dla Stable Diffusion jest dostępne online: https://github.com/CompVis/stable-diffusion W ostatnim czasie nastąpił szybki rozwój modeli generowania tekstu na obraz, z których niektóre (w tym DALL-E) oparte są na GPT-3. W większości przypadków techniki oparte na sztucznej inteligencji w sztuce generatywnej koncentrują się na funkcjonalnościach specyficznych dla danej dziedziny, takich jak przetwarzanie obrazu na obraz lub tekstu na obraz. Obecnie poniższe modele oferują najbardziej zaawansowane możliwości w zakresie generowania obrazu i wykorzystują techniki oparte na NLP do tworzenia imponujących obrazów:

• Stable Diffusion

- DALL-E 2 (OpenAI)
- Glide (OpenAI)
- Imagen (Google)
- Muse
- Make-a-Scene (Meta)
- Diffuse the Rest
- Latent Diffusion
- DreamBooth (Google)

Model DALL-E 2 był prawdopodobnie pierwszym z zaawansowanych modeli generowania obrazu opartych na sztucznej inteligencji.

Stability AI i Stable Diffusion

Stable AI to firma nastawiona na zysk, która nawiązała współpracę z RunwayML (startupem zajmującym się edycją wideo) w celu stworzenia Stable Diffusion, generatora tekstu na obraz o otwartym kodzie źródłowym. Obecnie Stable Diffusion zyskało przewagę nad konkurentami, takimi jak DALL-E 2 i Midjourney. Społeczność open source umożliwiła Stable Diffusion stanie się liderem (w tym momencie) wśród konkurencyjnych modeli konwersji obrazu na tekst.

Imagen (Google)

Google stworzył Imagen, model dyfuzji tekstu do obrazu (podobny do GLIDE), który koduje również podpowiedź językową za pomocą transformatora tekstu: https://imagen.research.google/. Badacze Google ustalili, że generyczne modele LLM, wstępnie wytrenowane na korpusach tekstowych, są bardzo skuteczne w kodowaniu tekstu na potrzeby syntezy obrazu. Dwa inne godne uwagi szczegóły: Imagen osiąga wynik SOTA w zbiorze danych COCO, a ludzie oceniają go wyżej niż inne narzędzia do generowania obrazów. Poniższe repozytorium Github zawiera implementację Imagen w PyTorch, która przewyższa DALL-E 2: https://github.com/lucidrains/imagen-pytorch. Imagen wykorzystuje tekstowe opisy scen do generowania obrazów wysokiej jakości. Google stworzył również DrawBench, czyli benchmark do oceny modeli przekształcania tekstu na obraz, wraz z obszerną listą dostępnych podpowiedzi dla Imagen.

Make-a-Scene (Meta)

Make-A-Scene z Meta oferuje multimodalną technikę, która łączy język naturalny i szkice w stylu dowolnym w celu generowania reprezentacji. Co więcej, Make-A-Scene działa z danymi wejściowymi, które mogą być zarówno tekstem, jak i szkicami. Zasadniczo podejście stosowane przez Make-A-Scene generuje obrazy z bardziej szczegółowym kontekstem, takim jak położenie, rozmiar i relacje między obiektami. Make-A-Scene wykorzystuje podejście wielomodelowe, które łączy przetwarzanie języka naturalnego (NLP) ze szkicami w stylu dowolnym. W przeciwieństwie do innych modeli przetwarzania tekstu na obraz, Make-A-Scene umożliwia dostarczenie szkicu, który uzupełnia podpowiedzi tekstowe, aby generować obrazy.

Rozproszenie reszty

Inną opcją jest darmowa aplikacja Diffuse the Rest. Jako prosty przykład działania Diffuse the Rest, Rysunek 3 przedstawia ręcznie narysowaną „ludzika z patyczków” wraz z komunikatem „ciało obce”, po czym wygenerowano obrazy z Rysunku 4 i Rysunku 5.

GauGAN2 (NVIDIA)

GauGAN2 to wczesny model głębokiego uczenia maszynowego, który wykorzystuje tekst do generowania fotorealistycznych obrazów. W przeciwieństwie do DALL-E 2 (omówionego w następnej sekcji), GauGAN2 nie bazuje na GPT-3, ale mimo to potrafi łączyć tekst z innymi typami danych wejściowych, a następnie generować wysokiej jakości obrazy. Użytkownicy mogą wpisywać krótkie frazy, a następnie GauGAN2 generuje obraz na podstawie treści tekstu. Na przykład, jeden z „podstawowych” przykładów firmy NVIDIA obejmuje ośnieżone pasmo górskie, które można dostosować, dodając inne funkcje.

PromptBase

PromptBase to kolejne narzędzie do generowania tekstu na obraz, które oferuje bezpłatne kredyty na generowanie obrazów: https://promptbase.com/. Jako prosty przykład działania PromptBase, rysunek 6 przedstawia obraz wygenerowany w odpowiedzi na następujący komunikat:

„Salvador Dali jadący na Harleyu Davidsonie na plaży Daytona na Florydzie, w czapce żartownisia i palący ogromne cygaro”.

Ograniczenia modeli Text-to-Image

Modele Text-to-Image (T2I) cieszą się ogromnym powodzeniem i popularnością. Pomimo swoich możliwości, istnieją ograniczenia co do tego, co te modele mogą generować, takie jak generowanie wielu znaków, tworzenie cieni i odbić oraz jakość generowanego tekstu.

MODELE TEKSTU NA OBRAZ

Synteza tekstu na obraz odnosi się do procesu generowania realistycznych i trafnych obrazów z opisów tekstowych. Zadanie to jest trudne ze względu na inherentną złożoność zrozumienia kontekstu, semantyki i niuansów w opisie tekstowym, a następnie przełożenia tego zrozumienia na reprezentację wizualną. W ciągu ostatnich kilku lat głębokie uczenie maszynowe, a zwłaszcza sieci GAN, wykazały się dużym potencjałem w tym obszarze. Do ważnych modeli tekstu na obraz należą:

• AttnGAN

• StackGAN

• DeepArt

Model AttnGAN wykorzystuje wieloetapowe udoskonalanie oparte na uwadze, aby generować drobnoziarniste obrazy o różnych rozdzielczościach z opisów tekstowych. Mechanizm uwagi pozwala modelowi skupić się na różnych częściach tekstu podczas generowania różnych części obrazu. Model StackGAN dzieli zadanie generowania tekstu na obraz na dwa etapy. W pierwszym etapie generuje obraz o niskiej rozdzielczości z opisu tekstowego. W drugim etapie obraz o niskiej rozdzielczości jest udoskonalany w celu wygenerowania obrazu o wysokiej rozdzielczości.

DeepArt koncentruje się na transferze stylu, gdzie opis tekstowy lub inny obraz jest używany do generowania obrazu w określonym stylu artystycznym. Model MirrorGAN wykorzystuje możliwości semantycznego osadzania tekstu i przestrzennych konfiguracji scen obrazu do generowania obrazów z opisów tekstowych. Składa się z trzech modułów: semantycznego osadzania tekstu, modułu Global-Local Collaborative Attentive oraz Streamlined Object Generator. Modele tekst-obraz o otwartym kodzie źródłowym można znaleźć w następujących lokalizacjach:

• GitHub

• Model Zoos

• Artykuły z kodem

• TensorFlow Hub

GitHub to główny hub projektów open source. Można tam znaleźć implementacje niemal wszystkich głównych modeli syntezy tekstu-obrazu, w tym AttnGAN, StackGAN i wielu innych. Model Zoos często towarzyszą frameworkom głębokiego uczenia (takim jak TensorFlow i PyTorch) i pełnią funkcję „centrów", w których dostępne są wstępnie wytrenowane modele. W przypadku zadań związanych z przetwarzaniem tekstu na obraz można znaleźć kilka modeli, ale GitHub oferuje zazwyczaj bardziej kompleksowe rozwiązania dla tego konkretnego zadania.

Papers with Code zawiera starannie wyselekcjonowaną listę artykułów na temat uczenia maszynowego wraz z kodem. To doskonałe źródło najnowocześniejszych modeli i ich implementacji open source. TensorFlow Hub firmy Google oferuje moduły uczenia maszynowego wielokrotnego użytku. Można znaleźć moduły związane z syntezą tekstu na obraz, choć są one bardziej ograniczone w porównaniu z GitHubem. Przypominamy, że zawsze sprawdzaj licencję powiązaną z kodem/modelem, gdy korzystasz z modeli open source lub tworzysz na ich podstawie, aby upewnić się, że używasz ich w sposób zgodny z intencjami twórców i wszelkimi powiązanymi warunkami.

MODELI DALL-E

Microsoft doda DALL-E do swojego pakietu Office i Azure AI, a Adobe planuje dodać narzędzia generatywnej AI do programu Photoshop. Generator treści Jasper również ogłosił ogromną rundę finansowania w wysokości 125 milionów dolarów. Sprawa GitHub Copilot „zaczyna wyglądać na pierwszą sprawę dotyczącą uczenia maszynowego i dozwolonego użytku w Stanach Zjednoczonych". Dostępnych jest kilka modeli generowania obrazu opartych na sztucznej inteligencji, a trzy modele omówione w poniższych podsekcjach są wymienione tutaj:

• DALL-E

• DALL-E 2

• craiyon (dawniej DALL-E Mini)

DALL-E

DALL-E to model autoregresyjny, którego nazwa jest połączeniem imienia i nazwiska słynnego artysty Salvadora Dalí oraz WALL-E z filmu Pixar. DALL-E to wariant modelu GPT-3 o 12 miliardach parametrów, który generuje bezstratne generowanie tekstu na obraz. DALL-E został wytrenowany na zbiorze danych par tekst–obraz. Projekt DALL-E o otwartym kodzie źródłowym można pobrać ze strony https://github.com/openai/dall-e. DALL-E wykorzystuje architekturę GPT do generowania obrazów pikselowych z tekstu, wiersz po wierszu. Zainspirowany architekturą GPT, traktuje słowa i piksele jako jedną sekwencję tokenów, dzięki czemu może kontynuować generowanie obrazu z komunikatu tekstowego. OpenAI prowadzi znaczące badania nad blokowaniem generowania tekstu lub obrazów za pośrednictwem DALL-E, które zawierają nieodpowiednie treści. DALL-E 2 rozszerza istniejące możliwości DALL-E 1 (np. fotorealizm), a także oferuje nowe możliwości wykraczające poza DALL-E 1.

Płatne konta dla DALL-E

DALL-E był początkowo programem beta dostępnym wyłącznie na zaproszenie, ale później OpenAI podjęło decyzję o obsłudze płatnego dostępu do DALL-E, który został zaprojektowany w następujący sposób:

• Monit o DALL-E kosztuje jeden kredyt.

• Użytkownicy otrzymują 50 darmowych kredytów w pierwszym miesiącu.

• Użytkownicy otrzymują 15 darmowych kredytów miesięcznie w kolejnych miesiącach.

• Użytkownicy mogą zakupić 115 kredytów za 15 USD miesięcznie.

W przeciwieństwie do programu beta dla DALL-E, użytkownicy mają pełne prawa do komercjalizacji obrazów tworzonych za pomocą DALL-E 2, co obejmuje możliwość przedrukowywania, sprzedaży i promowania wygenerowanych obrazów. Co więcej, użytkownicy uzyskują te prawa niezależnie od tego, czy wygenerowali obrazy za pomocą darmowego, czy płatnego kredytu, pod warunkiem przestrzegania polityki treści i warunków użytkowania OpenAI, dostępnych online:

https://labs.openai.com/policies/content-policy

https://labs.openai.com/policies/terms

Istnieją pewne ograniczenia: OpenAI odrzuci wszystkie obrazy przesłane do DALL-E, które przypominają treści o charakterze pornograficznym, treści z nazwami lub realistyczne twarze. Mając

na uwadze powyższe punkty, utwórz bezpłatne konto OpenAI, a następnie przejdź do poniższego linku, aby uzyskać dostęp do edytora DALL-E: https://labs.openai.com/editor

Jako prosty przykład, Rysunek 7 przedstawia wygenerowany obraz w odpowiedzi na następujący monit: Wyświetl klauna na obrazie Mony Lisy, który pali cygaro, ma na sobie duże okulary przeciwsłoneczne i pije piwo.

Jako inny przykład, wpisz tekst „narysuj krzesło awokado” w polu wprowadzania tekstu, wybierz 4 jako „Liczba obrazów, które chcesz” i kliknij przycisk „Generuj”. Po kilku minutach DALL-E wygeneruje cztery obrazy w odpowiedzi na monit, jak pokazano na rysunku 4.

Wywoływanie API DALL-E

OpenAI niedawno dodało publiczne API do dostępu do DALL-E. Listing wyświetla zawartość pliku dalle_api3.py, który pokazuje, jak wywołać to API i następnie wygenerować wiele obrazów.

Listing : dalle_api3.py

```
# pip3 install openai

import openai

openai.api_key = "your-openai-key"

# pętla do generowania wielu obrazów:

basename="avocado_couch"

image_count = 4

for ndx in range(image_count):

response = openai.Image.create(

prompt="an avocado couch,"

n=1,

# image sizes must be in ['256x256', '512x512',

'1024x1024']

size="512x512"

)

# generate URL with a DALL-E image:
```

```
image_url = response['data'][0]['url']

# save image:

import requests

img_data = requests.get(image_url).content

filename= basename+str(ndx)+".jpg"

with open(filename, 'wb') as handler:

handler.write(img_data)
```

Listing rozpoczyna się od instrukcji import, a następnie inicjuje zmienną open.api_key kluczem API. Następnie inicjowane są zmienne basename i image_count, po czym następuje pętla, która wywołuje metodę open.Image.create() czterokrotnie (co odpowiada wartości image_count), która następnie inicjuje zmienną response. Podczas każdej iteracji pętli zmienna image_url jest wypełniana adresem URL poprzez dostęp do jednego z elementów zmiennej response. Kolejna część pętli inicjuje zmienną img_data (zawierającą zawartość wygenerowanego obrazu), która jest zapisywana w systemie plików komputera. Zwróć uwagę, że nazwa pliku każdego obrazu zależy od bieżącej iteracji pętli, używając tego fragmentu kodu:

```
filename= basename+str(ndx)+".jpg"
```

Rysunek 8, Rysunek 9, Rysunek 10 i Rysunek 11 przedstawiają cztery obrazy wygenerowane przez kod z Listingu .

DALL-E 2

Ta sekcja jest podzielona na następujące podsekcje, które omawiają różne aspekty DALL-E 2:

• Omówienie DALL-E 2

• Model DALL-E 2

• Przygotowanie treści DALL·E 2

• DALL-E 2 i projektowanie komunikatów

• Tokenizator DALL-E

• DALL-E 2 i projektowanie komunikatów

• Tokenizator DALL-E

• DALL-E 2 Bot

Omówienie DALL-E 2

Chociaż DALL-E 2 jest następcą DALL-E, to jednak jest modelem opartym na dyfuzji (a nie autoregresyjnym, jak DALL-E). W rzeczywistości DALL-E 2 przypomina GLIDE, ponieważ oba modele generują obrazy z kodowania obrazu CLIP. DALL-E 2 wykorzystuje trzy główne kroki:

• Wykorzystanie modeli CLIP wytrenowanych na milionach obrazów.

• Zmodyfikowany model GLIDE generuje obrazy z osadzeń CLIP.

• Modele dyfuzyjne wstrzykują informacje tekstowe do obrazów.

Dla porównania, DALL-E łączy metody CLIP i dyfuzyjne, aby generować fotorealistyczne obrazy z wykorzystaniem danych wejściowych opartych na tekście, podczas gdy DALL-E 2 zawiera również technikę „unClip", która wykorzystuje koder do przetwarzania tekstu wejściowego. Koder najpierw generuje przestrzeń reprezentacji, która jest przekazywana do modelu, który mapuje dane wejściowe na zakodowany obraz, a następnie dekoder generuje obraz końcowy. Co więcej, DALL-E 2 zapewnia lepsze efekty wizualne w porównaniu ze swoim poprzednikiem, takie jak wyższa wierność i bardziej realistyczne obrazy, ponieważ wykorzystuje nowsze techniki SOTA text-to-image. Oprócz funkcjonalności DALL-E, DALL-E 2 obsługuje kilka nowych funkcji:

• tworzenie wariantów istniejącego obrazu

• dodawanie obiektów do istniejącego obrazu

• edycja zaznaczonego obszaru obrazu

• tworzenie obrazu łączącego dwa obrazy wejściowe

Przejdź na stronę główną i zarejestruj się, aby założyć bezpłatne konto. Wszelki tekst, który określisz, oraz wszelkie obrazy wygenerowane za pomocą DALL-E należą do Ciebie i mogą być wykorzystywane zarówno do użytku osobistego, jak i komercyjnego, podczas gdy DALL-E 2 ogranicza obrazy wyłącznie do użytku osobistego (tj. do użytku niekomercyjnego). Oprócz generowania obrazów na podstawie opisów tekstowych, DALL-E 2 może modyfikować istniejące obrazy, np. za pomocą opisu tekstowego. Ponadto DALL-E 2 może ulepszać istniejące obrazy, tworząc ich wariacje (również za pomocą opisów tekstowych). Kolejnym ważnym aspektem DALL-E 2 jest obsługa nieodpowiednich treści. W szczególności DALL-E 2 zawiera filtry, które analizują komunikaty użytkowników, aby zapobiegać treściom poniżającym, szerzącym nienawiść lub szkodliwym, skierowanym do osób lub grup osób. W rzeczywistości DALLE-2 posiada mechanizmy blokowania kombinacji tekstu lub obrazów, które są zabronione, mimo że poszczególne słowa lub obrazy mogą być nieszkodliwe. DALL-E 2 stara się również wykrywać i blokować treści sugerujące uprzedzenia płciowe, a także stereotypy rasowe. Początkowe konta DALL-E 2 były dostępne tylko na zaproszenie, co ograniczało wykorzystanie generowanych obrazów do użytku niekomercyjnego. Jednak w połowie 2022 roku OpenAI umożliwiło płatne subskrypcje DALL-E 2, które pozwalają na komercyjne wykorzystanie dowolnych obrazów tworzonych w DALL-E 2 (co ma sens w przypadku płatnej subskrypcji). Według różnych artykułów, na początkowej liście oczekujących na DALL-E 2 mogło znajdować się nawet milion osób, z których wszystkie będą miały możliwość zapisania się na płatną subskrypcję. W powiązanym temacie, co najmniej jeden robot jest zdolny do tworzenia sztuki. Ai-Da (nazwany na cześć Ady Lovelace) jest pierwszym robotem zdolnym do malowania jak ludzki artysta. Algorytmy sztucznej inteligencji (AI) zachęcają robota do przesłuchania, wyboru i podjęcia decyzji w celu stworzenia obrazu.

Model DALL-E 2

Podstawowe komponenty DALL-E 2 to:

• model CLIP

• poprzedni model

• model unCLIP (dekoder)

DALL-E 2 jest wynikiem połączenia poprzedniego modelu z modelem unCLIP.

• CLIP: Model, który pobiera pary obraz-podpis i tworzy „mentalne” reprezentacje w postaci wektorów, zwane osadzeniami tekst/obraz

• Model poprzedni: Pobiera osadzenie tekstu podpis/CLIP i generuje osadzenia obrazów CLIP

• Model dyfuzji dekodera (unCLIP): Pobiera osadzenie obrazu CLIP i generuje obrazy

• DALL·E 2: Połączenie modeli poprzednich i dekodera dyfuzji (unCLIP)

Przygotowanie treści DALL-E 2

OpenAI przeprowadził zadania wstępnego trenowania na zbiorze danych DALL-E 2, aby zapewnić zgodność z polityką firmy dotyczącą treści, w tym:

• usuwanie obrazów o charakterze seksualnym i przemocy

• redukcję stronniczości (mniej filtrowania)

• usuwanie powtarzających się obrazów

OpenAI odkrył interesujący (zaskakujący?) wynik: zwiększenie ilości filtrowania obrazów czasami prowadzi do bardziej niezrównoważonego zbioru danych (np. większej liczby mężczyzn niż kobiet). OpenAI usunął również powtarzające się obrazy w zbiorze treningowym z powodu efektu zwanego „regurgitacją obrazów”, w wyniku którego powtarzające się obrazy czasami pojawiają się w wynikach zamiast generować nowe. OpenAI filtrował dane treningowe za pomocą techniki GLIDE, która obejmuje następujące kroki:

• Stworzenie specyfikacji kategorii obrazów do etykietowania.

• Zebranie kilkuset pozytywnych obrazów dla każdej kategorii.

• Zebranie kilkuset negatywnych obrazów dla każdej kategorii.

• Wykorzystanie procedury aktywnego uczenia się w celu zebrania większej ilości danych.

Celem procedury aktywnego uczenia się było polepszenie kompromisu między precyzją a odtworzeniem. Ostatni krok polega na zastosowaniu klasyfikatora do całości w taki sposób, aby preferować odtworzenie zamiast precyzji. OpenAI skupił się na usuwaniu błędnych danych, zamiast zachowywać dobre dane z następującego powodu: łatwiej jest dostroić model niż sprawić, by model „oduczył się” czegoś, czego nauczył się w trakcie treningu. Przejdź do poniższego adresu URL, gdzie można uzyskać bardziej szczegółowe informacje na temat interesujących kombinacji technik zastosowanych przez OpenAI do wygenerowania wysokiej jakości zbioru danych obrazów dla DALL-E 2:

https://openai.com/blog/DALL-E 2-pre-training-mitigations

DALL-E-Bot

DALL-E-Bot umożliwia robotowi reorganizację obiektów w scenie, najpierw poprzez wywnioskowanie opisu tekstowego tych obiektów, następnie wygenerowanie obrazu przedstawiającego naturalny, przypominający ludzki układ tych obiektów, a na końcu fizyczne ułożenie obiektów zgodnie z tym obrazem. Istotne jest to, że osiągamy ten efekt zerowego ujęcia za pomocą DALL-E, bez konieczności dalszego gromadzenia danych ani trenowania. Zachęcające wyniki badań z udziałem ludzi pokazują, że jest to ekscytujący kierunek dla przyszłości algorytmów uczenia się robotów w skali sieci. Proponujemy również listę rekomendacji dla społeczności zajmującej się przetwarzaniem tekstu na obraz, aby

dostosować dalszy rozwój tych modeli do zastosowań w robotyce: https://www.robot-learning.uk/dall-e-bot.

DALL-E 3

Pod koniec 2023 roku OpenAI ogłosiło DALL-E 3, który znajdował się w „wersji testowej”. DALL-E 3 ma być dostępny w październiku 2023 roku dla użytkowników ChatGPT Plus (co wiąże się z subskrypcją w cenie 20 USD miesięcznie), a także dla klientów korporacyjnych. Co więcej, DALL-E 3 będzie dostępny za pośrednictwem interfejsu API. DALL-E 3 jest oparty na ChatGPT i wydaje się być znaczącym postępem w stosunku do swojego poprzednika, a także produktów konkurencji. Krótko mówiąc, DALL-E 3 umożliwia generowanie bardzo szczegółowych obrazów, które wiernie odzwierciedlają treść komunikatów.

WERSJE DEMONSTRACYJNE DALL-E

Możesz przetestować zarówno DALL-E, jak i DALL-E 2, wprowadzając ciągi tekstowe pod następującymi adresami URL:

https://main-dalle-client-scy6500.endpoint.ainize.ai/

https://gpt3demo.com/apps/openai-dall-e

https://gpt3demo.com/apps/DALL-E 2-by-openai

Jako zabawny przykład, wpisz tekst „narysuj samochód klauna z ananasami” w polu wprowadzania tekstu, wybierz 4 jako „Liczba obrazów, które chcesz” i kliknij przycisk „Generuj”. Po kilku minutach DALL-E wygeneruje cztery obrazy w odpowiedzi na Twoje polecenie, jak pokazano na rysunku 12.

Jako inny przykład, wpisz tekst „narysuj krzesło w kształcie awokado” w polu wprowadzania tekstu, wybierz 4 jako „Liczba obrazów, które chcesz” i kliknij przycisk „Generuj”. Po kilku minutach DALL-E wygeneruje cztery obrazy w odpowiedzi na Twój monit, jak pokazano na rysunku 13.

GENEROWANIE TEKSTU NA WIDEO

Rozwój narzędzi do przetwarzania tekstu na obraz szybko doprowadził do rozwoju narzędzi do przetwarzania tekstu na wideo. Poniżej wymieniono niektóre firmy, które stworzyły produkty do przetwarzania tekstu na wideo:

• Meta (Make-a-Video)

• Imagen

• Text-to-Video (Stable Diffusion)

• Ruben Villegas i Google (Phenaki)

Produkty każdej z powyższych firm zostały krótko opisane w poniższych podsekcjach.

Meta (Make-a-Video)

Meta stworzyła Make-A-Video, które wykonuje tłumaczenie tekstu na wideo (T2V), wykorzystując T2I (tekst na obraz) w połączeniu z nienadzorowaną treścią wideo i obecnie reprezentuje SOTA w T2V pod względem jakości wideo. Jedną z zalet Make-A-Video jest to, że nie wymaga sparowanych danych tekstowo-wideo. Po drugie, Make-A-Video zapewnia wyższą jakość generowanych filmów dzięki możliwości uchwycenia bogactwa obrazów, z których generuje treści wideo. Make-A-Video wykorzystuje połączenie sieci U-Net, tensorów uwagi i potoku do tworzenia filmów.

Imagen Text-to-Video

Modele tekstu na obraz, takie jak DALL-E, mają niezwykłą zdolność

tworzenia zachwycających efektów wizualnych. Z kolei przetwarzanie tekstu na wideo jest wciąż w powijakach, prawdopodobnie dlatego, że to zadanie jest znacznie bardziej złożone w porównaniu z generowaniem obrazów. Niemniej jednak, Imagen Video od Google to model przekształcania tekstu na wideo, który wykorzystuje modele dyfuzyjne do tworzenia filmów w wysokiej rozdzielczości. Imagen Video składa się z 7 podmodeli, które realizują generowanie wideo warunkowego pod względem tekstu, superrozdzielczości przestrzennej i superrozdzielczości czasowej. Rezultatem tego procesu są filmy w wysokiej rozdzielczości 1280×768 z szybkością 24 klatek na sekundę.

Ruben Villegas i Google (Phenaki)

Ruben Villegas współpracował z Google, aby stworzyć Phenaki, narzędzie przekształcające tekst na wideo, które może generować filmy o dowolnej długości. Sam tekst może mieć strukturę przypominającą historię. Idea Phenaki opiera się na traktowaniu poszczególnych obrazów tak, jakby były klatkami w filmie (co stanowi prawdziwą kompozycję filmów). Następnie te „filmy" są łączone z krótkimi filmami z napisami, co skutkuje powstaniem długich i zaawansowanych filmów poprzez powtórzenie poprzedniego procesu. Niższe szczegóły Phenaki obejmują:

• koder generujący osadzenia wideo

• model językowy generujący osadzenia tekstu

• dwukierunkowy transformator

• dekoder tłumaczy osadzenia wideo na piksele

Należy zauważyć, że dwukierunkowy transformator generuje nowe osadzenia wideo na podstawie tekstu i istniejących osadzeń wideo. Powstały model Phenaki ma liczbę parametrów bliską 2 B, która została zredukowana w celu utworzenia modelu o liczbie parametrów bliskiej 1 B.

GENEROWANIE TEKSTU NA MOWĘ

Generowanie dźwięku z tekstu było dostępne jako TTS (tekst na mowę) i można znaleźć w Internecie przykłady kodu Pythona, które realizują to zadanie. Jednak niektóre nowe narzędzia oparte na sztucznej inteligencji umożliwiają generowanie tekstu na mowę w celu tworzenia podcastów. Jedną z firm oferujących tę funkcjonalność jest Play.ht, której strona główna to https://podcast.ai. Jeden z podcastów dotyczy Joe Rogana i Steve'a Jobsa.

Whisper (OpenAI)

Whisper to model kodera-dekodera oparty na transformatorach, o otwartym kodzie źródłowym, który wykonuje wiele zadań, w tym automatyczne rozpoznawanie mowy (ASR), wielojęzyczne rozpoznawanie mowy, tłumaczenie mowy i identyfikację języka. Zbiór danych audio dla Whisper składa się w około 2/3 z języka angielskiego, a pozostała 1/3 z danych z różnych innych języków. Whisper sprawdza się wyjątkowo dobrze w rozpoznawaniu mowy w języku angielskim. Whisper transkrybował filmy z Twittera oraz wykłady z dokładnością ponad 99%.

PODSUMOWANIE

Ta sekcja rozpoczęła się od omówienia wizualizacji z wykorzystaniem generatywnej sztucznej inteligencji (AI) i popularnych narzędzi do generowania. Następnie zapoznałeś się z AI21 i Aleph-Alpha oraz ich odpowiednimi platformami, podobnymi do platformy GPT-3. Następnie poznałeś dyfuzję i modele dyfuzji oraz ich kontrast z sieciami GAN. Dodatkowo poznałeś narzędzia do generowania tekstu na obraz, takie jak Stability AI i Imagen. Co więcej, poznałeś DALL-E i jego następcę DALL-E 2, które generują obrazy z danych tekstowych. Na koniec poznałeś narzędzia do generowania tekstu na wideo oraz narzędzie do generowania tekstu na mowę. Gratulacje! Dotarłeś do końca dynamicznej podróży, która rozpoczęła się od mechanizmu uwagi i architektury transformatora. Omówiliśmy BERT, GPT-3 oraz ChatGPT i GPT-4. W tym momencie jesteś w dobrej pozycji, aby wykorzystać zdobytą wiedzę jako punkt wyjścia do pogłębienia swojej wiedzy na temat generatywnej sztucznej inteligencji.