Session Report

ユーザー体験を革新する
高速・高精度の検索を可能にする最新技術

生成 AI のマルチモーダル技術―「視覚を持った LLM」ができること

Google Cloud
デベロッパー リレーションズ
デベロッパー アドボケイト
佐藤 一憲

Google Cloud

Google Cloud

セッション レポート概要

生成 AI の新技術「マルチモーダル」では、文章と画像それぞれの意味を LLM が人間並みに深く理解することで、文章で商品画像を検索したり、その逆を可能にしたりなど、より高度化された検索を実現できます。人手で多くのタグやラベル、説明文を付けることなく、瞬時にデータベース検索ができます。ここでは、画像検索のデモを例に取りながら、生成 AI のマルチモーダル機能がビジネスにもたらす価値を解説します。

プレゼンター紹介

Google Cloud
デベロッパー リレーションズ
デベロッパー アドボケイト
佐藤 一憲

AI/ML 担当のデベロッパー アドボケイトとして Google Cloud US 本社のブログ記事の執筆やデモ開発を担当。2016 年より 20 件以上を掲載し、ジェフディーン、エリックシュミット、ニューズウィーク誌、ニューヨーカー誌により紹介された。また Google I/Oや Cloud Next SF などのイベントにも例年登壇。X（旧：Twitter）では 1 万 7 千人のフォロワーに向けて AI/ML の情報を日々共有している。

目次

LLM の「グラウンディング」とは何か

ChatGPT や Bard といったツールは、確率的なモデルに従って人間が自然に感じるやりとりを生成するものです。それゆえ、例えばあるデータベースの中に数万〜数十万件の製品データが格納されていたとき、その各製品の型番や属性などを検索して正しい応答を返すというようなことは生成 AI 単体としては不得手であり、これは検索エンジンやデータベースにしかできません。

また、ハルシネーション（事実に基づかない誤った情報を AI が生成してしまうこと）の問題もあります。ChatGPT や Bard では、一見すると自然な文章を返すものの、検証すると内容としては誤っていることもあり、これは現状の LLM の技術では避けられません。

この問題をどう解決すれば良いかを考えた際に大切なのが「グラウンディング」です。AI が誕生した 1960 〜 1980 年代からある言葉で、現実世界と AI をどのようにつなげるか、という意味のキーワードです。

具体的に、自社が使用するデータベースやシステムに AI をどのようにグラウンディングさせるか。そのために欠かせないのが後述する「エンベディング」と「ベクトル検索」という 2 つの技術です。

LLM でさらに強化されるセマンティック検索

エンベディングとベクトル検索の効果をご紹介するために、まずは、ある質問を投げかけてそれと類似するものをデータベースから検索するというデモ環境を例に挙げます。

当社が用意したこのデモサイトでは、エンジニアの方に馴染みの深い Q&A サイト「Stack Overflow」の質問と回答の組み合わせが 800 万件入っています。そこで「How to shuffle rows in SQL（SQL にて行をシャッフルする方法）」という質問を検索してみたとしましょう。すると、約 800 万件の中から 0.02 秒ほどで、検索した文と似たような質問・回答の候補を抽出して提示してくれます。

旧来型の検索エンジンとこのデモの違いは、このデモでは「セマンティック検索」、すなわちユーザーの検索意図を検索エンジン側が理解して回答を提示する技術が使われていることです。Google においても 2015 年頃から Google 検索に同技術を活用しています。

例えば、従来のキーワード検索では、「Movie」と検索しても同義語である「Film」を検索結果に提示することができませんでした。しかし、セマンティック検索では、キーワードの意味やユーザーの意図を理解し、意味が同じ、または近いものを検索結果に表示できます。

ここで重要なのは、このセマンティック検索を LLM の技術と組み合わせることで、より高度な検索を実現できるようになるということです。先ほど検索した「How to shuffle rows in SQL」という語句は、エンジニアでなければわからないような専門的なものです。しかし、このデモでは Stack Overflow のデータなど専門的な文章を事前に機械学習しているわけではありません。あくまで LLM は一般的な知識をあらかじめ学習し、その理解力のみを用いてセマンティック検索を高度化することに成功しています。

生成 AI は「生成機能」を利用しなくても有効

生成 AI の話題でよく出てくるのが「プロンプト」です。上記のデモでは生成 AI を利用しているものの、プロンプトは利用していません。また何かコンテンツを生成することもしていません。

生成 AI が持つ 2 つの側面として、例えば与えられた文章や絵などを人間と同じように「理解すること」と、それを踏まえて、絵や文章というアウトプットを「返すこと」があります。今回のデモで利用しているのは前者のみです。生成 AI に関する誤解の 1 つでもありますが、生成することだけが生成 AI の本質ではありません。

実は、これだけでもビジネスにおいては有用です。生成しないということは、ハルシネーションが発生することがありません。与えられた質問を理解し、データベースに格納された近いものを検索しているだけなので、普段企業が使用しているデータベースやシステムにも導入しやすいでしょう。

もう 1 つ、このデモで言及したいのが、スケーラブルで高速であることです。ChatGPT や Bard では応答時間に数秒かかりますが、これでは検索エンジンやシステムのレコメンデーション機能へ活用するのに現実的ではありません。これを解決する技術が、このデモで用いられている「ベクトル検索」です。この仕組みは後ほど説明します。

専門的な内容を含んでいても正確で高速な検索が可能になる

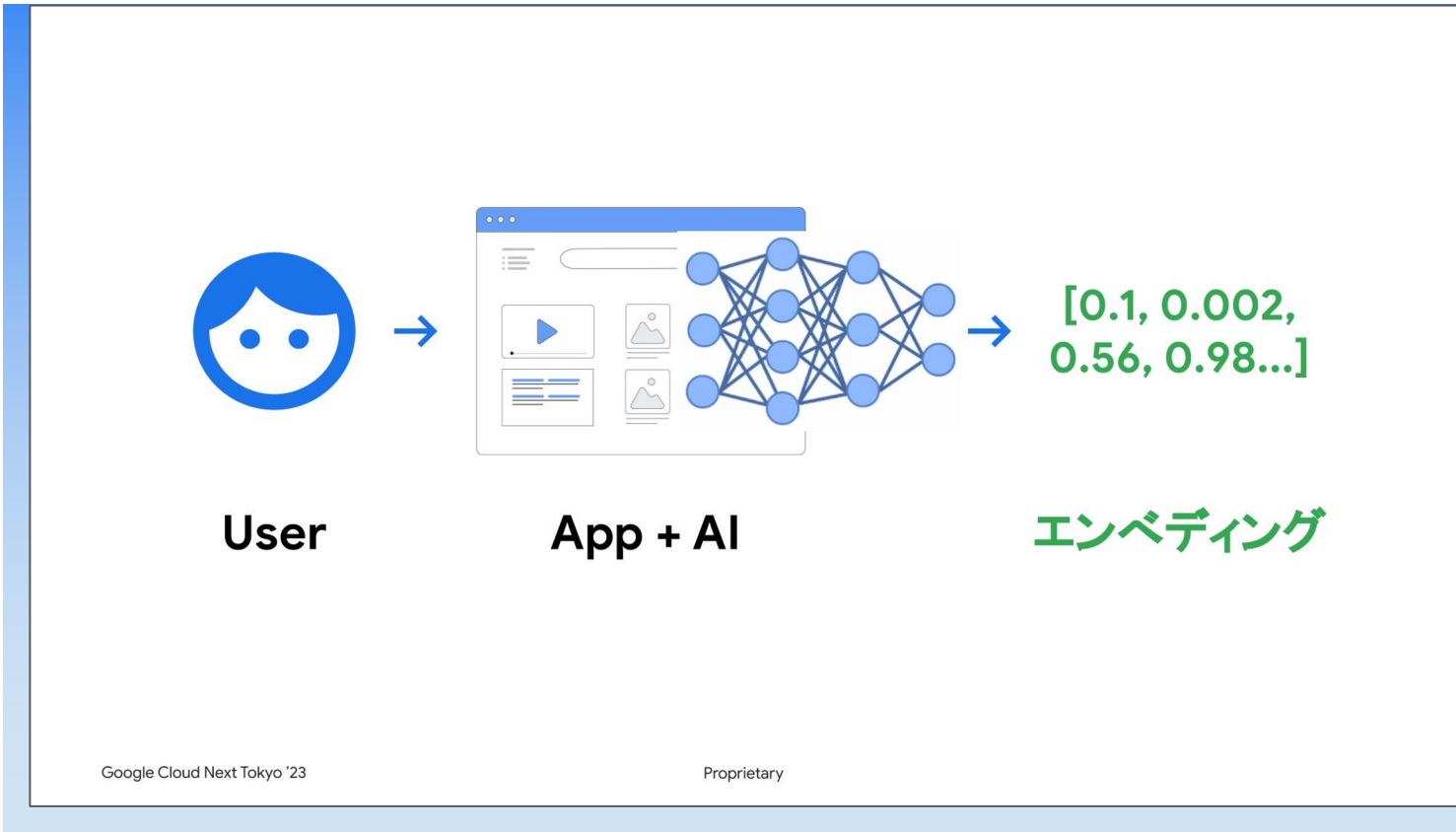
昨今の検索・レコメンデーション技術を支える「エンベディング」

先に紹介したデモにて使用されている重要な 2 つの技術が、「エンベディング」と「ベクトル検索」です。このうちまずエンベディングを紹介します。

エンベディングを理解するために、まずはエンベディングを使わないシステムを考えます（下図）。バックエンドには、従来型のデータベースのように表形式のテーブルが用いられている形です。

世の中の IT システムの 99.9%がこれに該当しますが、一方で UX の制約になっている実情もあります。例えば、レストランを探す場合には、まずカテゴリを絞り込んで、次に食べたい料理を探すといったステップを踏みます。言い換えれば、データベースに存在するカテゴリでしか検索できないということです。

それに対して、Google 検索や YouTube のレコメンド機能はもちろん、Google 以外のビッグテックが提供しているサービスの多くは、こうした従来型の技術だけではなく「エンベディング」というものが利用されています。従来のデータベースはもちろん存在しますが、AI が作り出したエンベディングにより、ユーザーがリアルタイムに求めている情報をシステムが提供することができます。このエンベディングの技術は、LLM や生成 AI とは並行して起きており、AI のイノベーションとも言えます。

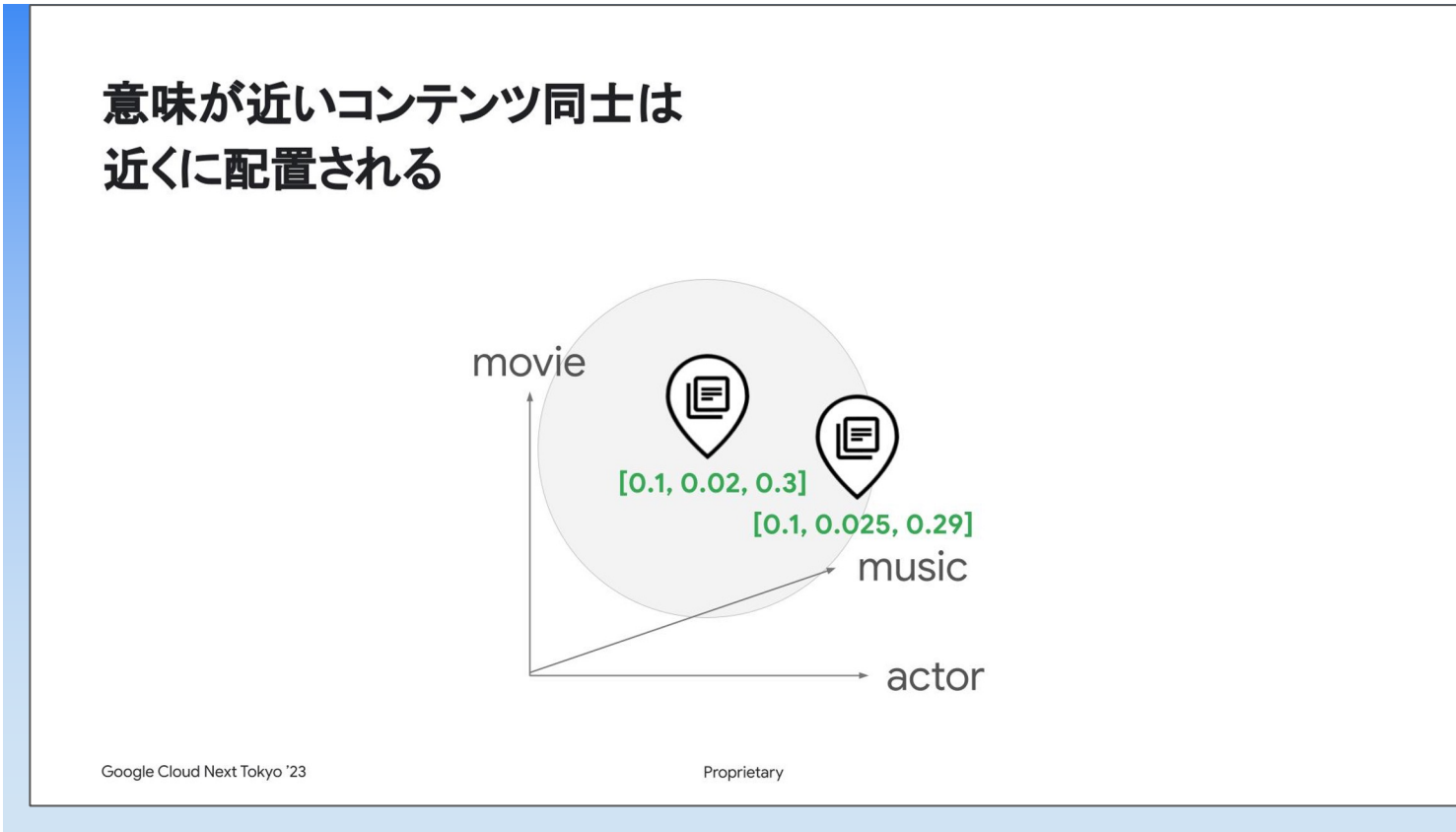


昨今の高度な検索技術を提供するシステムにはデータがエンベディングとして格納されている

エンベディングとは意味情報をベクトルとして表したもの

では、実際にエンベディングとはどのようなものでしょうか。これは端的に言えばベクトルで表されたデータであり、「意味の地図における座標」と捉えられます。例えば、映画や音楽、俳優についてそれぞれの 10%、2%、30% の割合で言及しているコンテンツがあったとしましょう。そのコンテンツに対してディープラーニング モデルが 3 次元の意味の地図上に［0.1, 0.02, 0.3］といった座標を作成するイメージです。

エンベディングというベクトルの形では、意味が近いコンテンツがその周辺に配置されるため、類似したものを瞬時に検索できるようになります。従来のデータベースでは、意味を扱うにはそれぞれのデータに対して人間が意味をラベリングするしかありませんでしたが、AI によってさらに細かい粒度で意味情報を表せるようになりました。



エンベディングは似たようなデータを探す上で有用な形式

エンベディングにおけるデータは、例えば、画像や音声、IoT のセンサーなど AI で扱えるものであれば種類を問いません。実際に Google でも、例えば Google アナリティクスや Google ショッピング上のユーザーの行動データから、その行動の意味やユーザーの嗜好をエンベディングに置き換え、さまざまなサービスに応用しています。

このような技術が、昨今のビッグテックが実現・提供しているユーザー体験の根底にあります。いまや Googleの 収益の大半は、このエンベディングと後述するベクトル検索の技術のおかげと言っても過言ではないでしょう。

エンベディングの難易度を LLM が大幅に低減

エンベディング技術によって、優れたユーザー体験を提供するコンシューマ サービスが生まれていますが、これに慣れてしまうと、業務システムにおいても、上記と同等のユーザー体験を期待するのがもっともです。

それにもかかわらずエンベディングが流行しなかった理由は、エンベディングを作成する難易度が高かったという背景があります。データ サイエンティストがディープ ラーニングのモデルをゼロから作ったり学習データを用意したりと時間や費用を要するものでした。

しかし、生成 AI に関する技術が進化したことで、文章や画像を API に投げることで、人間と同じように意味を捉えたエンベディングを、専門的な知識を要せず瞬時に生成可能になりました。現在のエンベディングは従来のものと次元数は変わりませんが、捉える意味の解像度が格段に高まっていることが特徴です。

このようなエンベディングの高度化により、商品説明文を人間が作成したものと同じように細かく表現できたり、おすすめ音楽を細かく分類できたり、商品のレコメンデーションをより高い精度で行えたりと、ジャンルを問わずさまざまなことが可能となりました。

実際に、LLM を組み合わせたエンベディングの例を、先ほどのデモ環境でご紹介した Stack Overflow の質問データで見ていきましょう。例えば、下図では、1 列目の質問文と類似する質問を 2 列目に例示しています。これは、LLM が類似度を判定して分類していることを示しています。

質問文	エンベディング空間で近くにある質問	LLM の理解レベル
How do I write a class that instantiates only once	How to create a singleton class based on its input?	The model knows "instantiating only once" and "singleton" mean the same.
Does moving the request line to a header frame require an app change?	Does an application developed on HTTP/1.x require modifications to run on HTTP/2?	HTTP/2 ではヘッダフレームの変更が必要であることを理解している
How to increase IO speed with TensorFlow?	Tensorflow GPU/CPU Performance Suddenly Input Bound	The model knows both questions intent to improve TensorFlow IO performance

`for num in range(1,101):`
`  string = ""`
`  if num % 3 == 0:`
`  string = string + "A"`
`  if num % 4 == 0:`
`  string = string + "B"`
`  if num % 4 != 0 and num % 3 != 0:`
`  string = string + str(num)`
`  print(string)`
	Write a program that prints the integers from 1 to 100 (inclusive) in one line of code
	I am new to python and would like to write a program that prints the integers from 1 to 100 (inclusive) in 1 line using python:
	`for i in xrange(1, 101):`
	`  if i % 15 == 0:`
	`  print "shellfish"`
	`  elif i % 3 == 0:`
	`  print "shell"`
	`  elif i % 5 == 0:`
	`  print "fish"`
	`  else:`
	`  print i`
		The model thinks both Python code looks similar in what they are doing (printing fizzbuzz-like sequences), although they are not exactly the same.

Google Cloud Next Tokyo '23 Proprietary

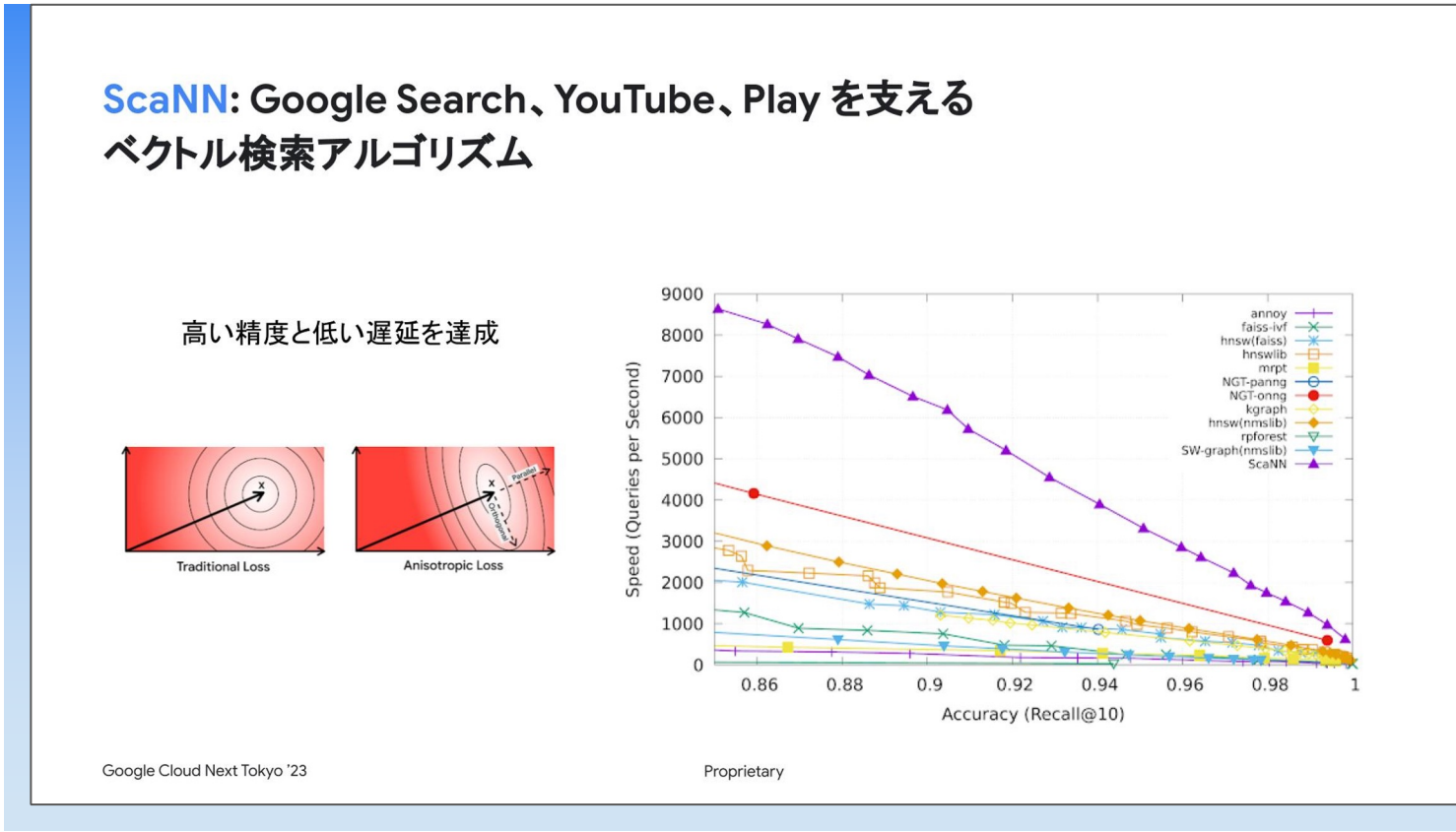
エンベディング空間で近くにあるとLLMが判定したもの

エンベディングどうしの距離を高速に計測するベクトル検索

ここまでエンベディングの概要とそのインパクトについて説明してきましたが、もう 1 つ重要な観点として、類似するエンベディングをいかに瞬時に検索できるかという課題があります。

類似するエンベディングを検索するということは、言い換えればエンベディングどうしの距離を測ることです。それ自体は難しいことではありませんが、問題なのは、ベクトルの計算が膨大化してしまうということです。

これに対して、近年研究が進んでいるのが、近似最近傍探索(Approximate Nearest Neighbor:ANN）という技術です。この技術は、広大なエンベディング空間をツリー構造に分解し、検索する範囲を絞り込むことで高速化を実現するものです。Google ではこの技術を用いて、現在業界で最速と呼ばれている ScaNN と呼ばれるベクトル検索アルゴリズムを開発しました。普段、日々皆さんが使用している Google 検索や Google マップなどの各種サービスで、あまり遅延なく待たされることがないのも、この ScaNN があるからこそです。

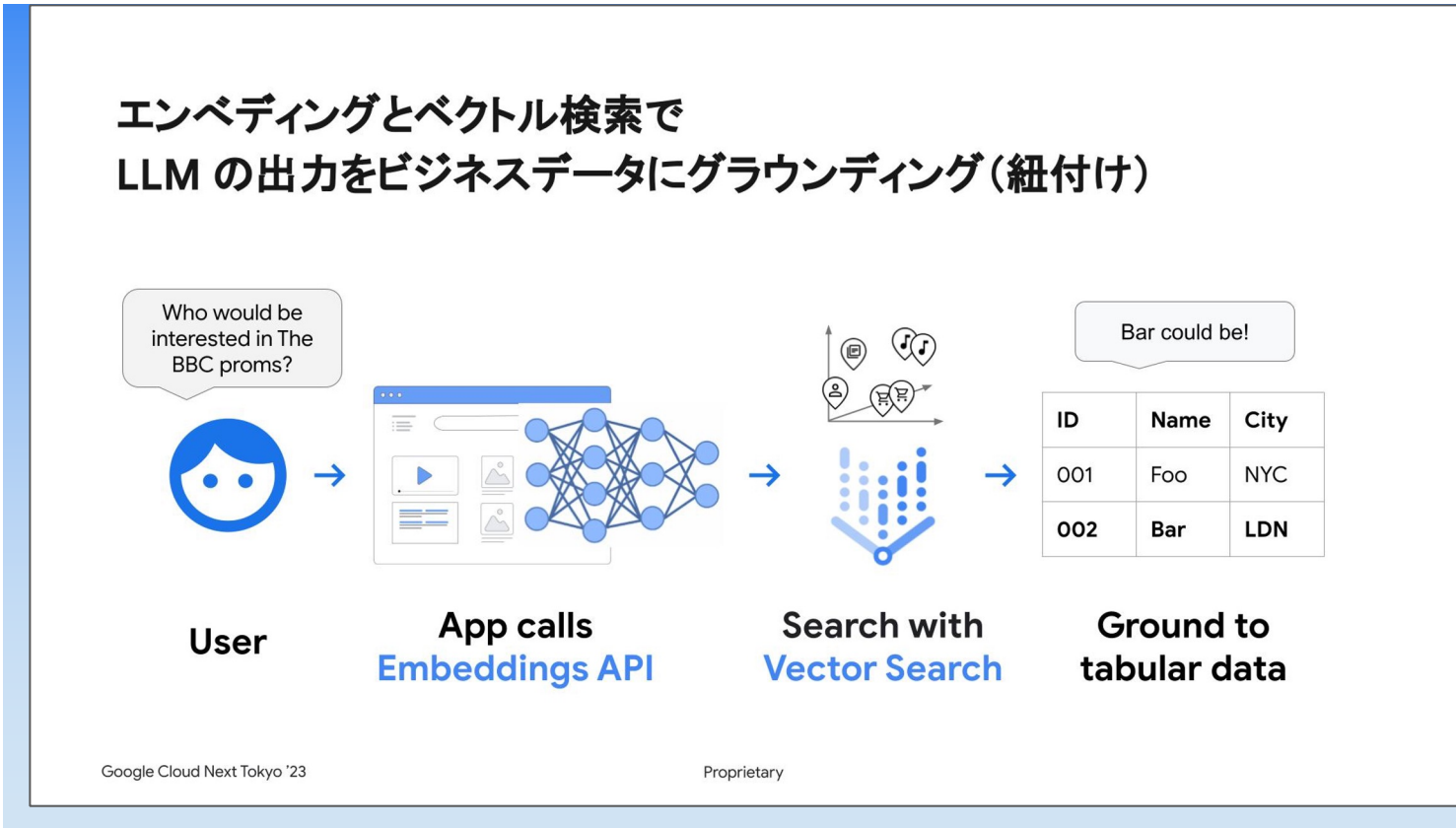


ScaNN は Google の数多くのサービスを支えている

Google 自身が扱うベクトル検索を利用できる Vertex AI Vector Search

この技術をユーザーが利用できるようにしたのが、Google Cloud の「Vertex AI Vector Search」というサービスです。これを使用することで、Google の屋台骨となっているシステムをそのまま自社の既存のシステムに組み込むことができます。

また、冒頭に触れたグラウンディングにも応用できます。例えば、ユーザーから問い合わせがあった際、つまり何らかの文章で入力を受け付けたとき、それを API でエンベディングに変換し、ベクトル検索の形で検索を行います。その結果、例えば商品 ID や文章 ID などに変換できるため、データベース上のビジネスデータに簡単に紐づけることが可能です。データ サイエンティストがいなくても、このようなシステムを Google Cloud の Vertex AIの中で完結して実現できるのです。

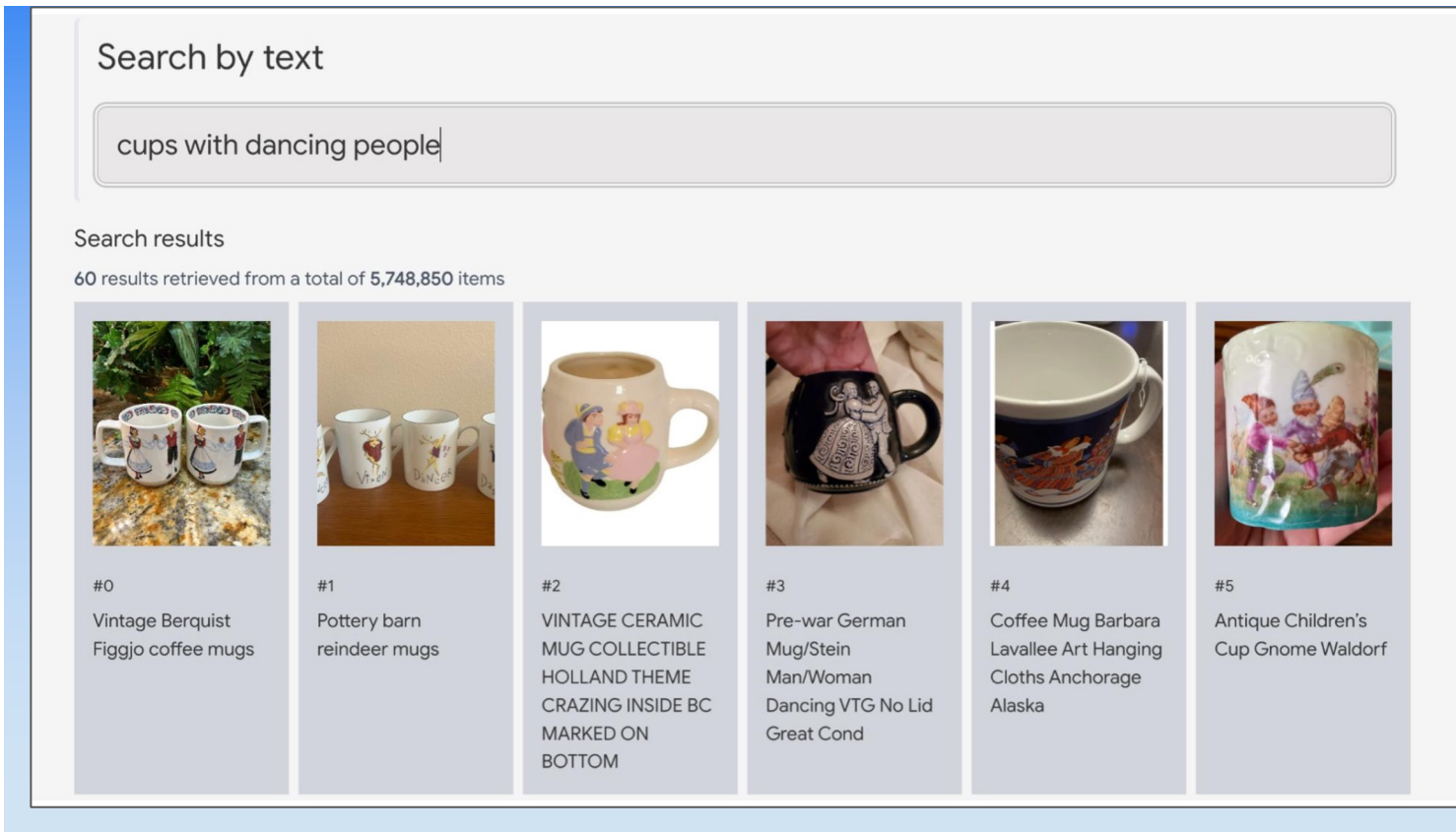


「視覚を持った LLM」がもたらす検索の革新

最後に、「視覚を持った LLM」について解説します。例えば、数百万枚の商品画像や製品画像に対して人間と同じように意味を理解し、検索や推薦システムに活用する、またはセキュリティ カメラや工場の検査画像に活用したい場合にはどうするのでしょうか。

このような場合でも、エンベディングとベクトル検索を組み合わせることで、実現可能になります。ここでは「マルチモーダル エンベディング」といったものを活用します。マルチモーダル エンベディングとは、テキストだけではなく画像や動画、音声などのモーダルに対して分け隔てなく意味を理解し、その類似度を表現できるエンベディングのことです。

例えば、下図は、メルカリで使用されている画像をお借りして作成したデモ環境にて「踊ってる人が書かれているコップ」と検索し、600 万件のデータの中から 0.02 秒程度で結果を表示したものです。ここで特筆したいのは、商品のタイトルや説明文といったカテゴリやラベルは使用せず、画像のみを見て生成 AI が人間と同じ理解力で検索結果を表示したものになります。このように、最新の技術では画像の意味を人間と同レベルで理解するところまでを数ミリ秒で検索できるようになっています。



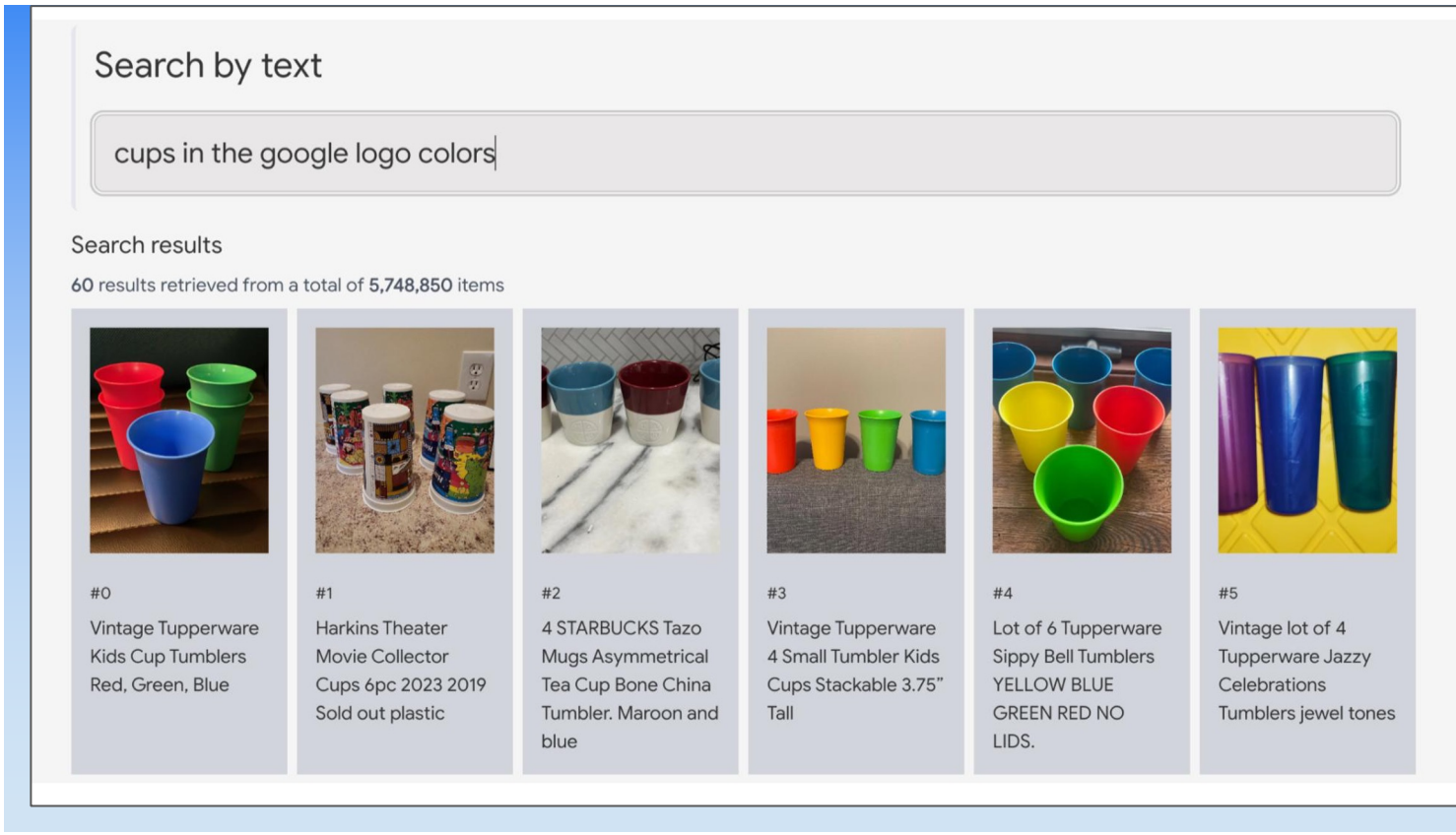
画像を理解した上で高速な検索が可能になる

このデモの特徴としては、画像の意味を人間と同じように理解するものの、生成を行っているわけではないので、ハルシネーションが発生することはありません。もちろん、スケーラブルで高速である点もこれまで説明した特徴と同様です。

マルチモーダル AI による検索がビジネスに価値をもたらす

マルチモーダルでは、同じエンベディング空間内にテキストだけでなく、画像も意味の近さで分類して配置できます。これにより画像からテキストを探したりテキストから画像を探したり、さらには画像から画像を探すことも可能となりました。

下図は「Google ロゴの色をもったコップ」と入力した際の検索結果です。学習を行わずに事前に蓄えた知識のみで、Google のロゴの色を理解して結果を表示しています。



Google ロゴの色のコップを検索したイメージ

マルチモーダル AI の用途は多岐にわたります。例えば EC サイトに出品するものの画像をアップロードし、それを生成 AI の API にわたすような仕組みを構築すれば、商品のタイトルや説明文を、画像から判断してある程度自動的に生成できるようになるかもしれません。

ほかにも、セキュリティカメラにおいて、「ドアが開いた」など自然言語で条件を指定し、その条件に該当する様子が録画されると自動でアラートを通知するといった仕組みも可能になるでしょう。

このマルチモーダル AI の機能は、Google Cloud では、すでに複数のプロジェクトに組み込まれています。Vertex AI Search という非常に容易に扱える検索エンジンもありますし、機械学習エンジニアの方であれば、ベクトル検索とエンベディングを組み合わせることで、自分で検索バックエンドを作成してもよいでしょう。

	Vertex AI Search (website/unstructured app)	Vertex AI Vision Warehouse	Vertex AI Search (structured app)	Vertex AI Vector Search
対象ユーザ	IT engineers, requires no ML expertise	IT engineers, requires no ML expertise	IT engineers, requires basic ML expertise	ML engineers/Data Scientists
検索アルゴリズム	Keywords + embeddings-based search	Embeddings-based search	Keywords + embeddings-based search	Embeddings-based search
Vision Language Model (VLM)	Google Internal	Google Internal	Vertex AI Multimodal Embeddings (CoCa)	Vertex AI Multimodal Embeddings (CoCa)
簡単に開発	Yes	Yes	Yes (for the search engine)	No
検索対象データ	Web pages/PDF files indexed on Search	Media (images and videos) indexed on Vision Warehouse	Tabular data indexed on Search	Any images or texts indexed on Vector Search
エンベディングの取り出し	No	No	Yes	Yes
エンベディング生成コスト	No	Yes	Yes	Yes

Google Cloud で提供されるソリューション

今回は、生成 AI のマルチモーダル技術について、エンベディングとベクトル検索の解説から始まり、視覚を持った LLM でできることを紹介しました。このようなマルチモーダル技術を活用して、ぜひ自社のより良いサービス開発を行っていただければと思います。

参照リンク

1. Cloud FinOps の概要
2. AI と ML のソリューションの概要
3. AI と機械学習のプロダクトの概要
4. 生成 AI のマルチモーダル技術ー「視覚を持った LLM」ができること アーカイブ動画視聴ページ

製品、サービスに関するお問い合わせ

goo.gl/CCZL78

Google Cloud の詳細については、上記 URL もしくは QR コードからアクセスしていただくか、同ページ「お問い合わせ」よりお問い合わせください。

Google Cloud