AI 特許紹介(74)
AI 特許を学ぶ！究める！
～Imagen 特許～

2025 年 3 月 10 日
河野特許事務所
所長弁理士　河野英仁

「AI 特許紹介」シリーズは、注目すべき AI 特許のポイントを紹介します。熾烈な競争となっている第４次産業革命下では AI 技術がキーとなり、この AI 技術・ソリューションを特許として適切に権利化しておくことが重要であることは言うまでもありません。

AI 技術は Google, Microsoft, Amazon を始めとした IT プラットフォーマ、研究機関及び大学から毎週のように新たな手法が提案されており、また AI 技術を活用した新たなソリューションも次々とリリースされています。

本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

1.概要
特許権者　Google
出願日　2023 年 5 月 19 日
登録日　2024 年 5 月 7 日
登録番号 US11978141
発明の名称　生成ニューラルネットワークのシーケンスを使用した画像生成

141 特許は、テキストエンコーダに大規模言語モデルである T5 を用い、T5 の後段に複数の拡散モデルを組み合わせることで、高レベルのフォトリアリズムと深いレベルの言語理解能力を備えた画像生成 AI である Imagen 技術に関する。

2.特許内容の説明
図 1A は、画像生成システム 100 の一例のブロック図を示す。

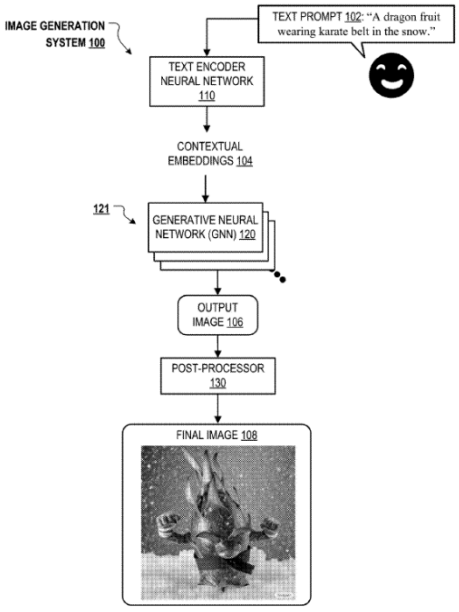


FIG. 1A

画像生成システム 100 は、テキストエンコーダニューラルネットワーク 110、生成ニューラルネットワーク(GNN)のシーケンス 121、およびポストプロセッサ 130 を含む。システム 100 は、テキストプロンプト 102 を入力として受信し、最終画像 108 を出力として生成する。

システム 100 は、シーンを説明するテキストプロンプト$\mathcal{T}$102 を受信する。テキストプロンプト 102 は、自然言語 $\mathcal{T}=(\mathcal{T}_1,\mathcal{T}_2,\ldots)$の複数のテキストトークン$\mathcal{T}_{1,2,\ldots}$を含むテキストシーケンスである。たとえば、図 1A に示すように、テキストプロンプト 102 は、「雪の中で空手帯を巻いたドラゴンフルーツ」である。

テキストエンコーダ 110 は、テキストプロンプト 102 を処理して、テキストプロンプト 102 のコンテキスト埋め込み(u)のセットを生成する。テキストエンコーダ 110 は、事前トレーニング済みの自然言語テキストエンコーダ、たとえば、T5-XXL などの T5 テキストエンコーダ、CLIP テキストエンコーダ、大規模言語モデル(LLM)などである。

コンテキスト埋め込み 104 は、システム 100 による処理のために計算しやすい表現を提供するテキストプロンプト 102 のエンコードされた表現とも呼ばれる。たとえば、コンテキスト埋め込み 104 は、値のセット、ベクトル、または配列（たとえば、UNICODE または Base64 エンコード)、英数字の値、記号、または任意の便利なエンコードにすることができる。

GNN121 のシーケンスは、それぞれが各入力(c)を受信するように構成された複数のGNN120 を含む。各 GNN120 は、それぞれの入力を処理してそれぞれの出力画像を生成する。一般に、シーケンス 121 には、初期出力画像（たとえば、低解像度）を生成する初期 GNN と、初期出力画像の解像度を徐々に高める 1 つ以上の後続 GNN が含まれる。最初の GNN の各入力にはコンテキスト埋め込み 104 が含まれ、後続の各 GNN の各入力には、シーケンス 121 内の前の GNN によって生成された出力画像が含まれる。

後続の GNN の 1 つ以上のそれぞれの入力には、後続の GNN がテキストプロンプト 102 を条件とできるようにするコンテキスト埋め込み 104 も含まれる。たとえば、最初の GNN はテキストプロンプト「猫の写真」に関連付けられたコンテキスト埋め込みのセットを受け取り、後続の GNN の 1 つ以上はテキストプロンプト「猫の油絵」に関連付けられたコンテキスト埋め込みのセットを受け取る。

システム 100 は、シーケンス 121 を通じてコンテキスト埋め込み 104 を処理し、アーティファクトがほとんどない高解像度の出力画像 106 を生成する。出力画像 106 は、ポストプロセッサ 130 によってさらに処理され、最終画像(x)108 が生成される。

参考までに、図 1A に示されているサンプル画像は、1024×1024 ピクセルの解像度で生成されている。サンプル画像は、ベース画像生成モデルを採用した最初の DBGNN と、超解像度モデルを採用した 2 つの後続 DBGNN を含む、3 つの拡散ベース GNN (DBGNN)のシーケンスを実装する画像生成システムによって生成された。最初の DBGNN は、初期解像度 64×64 で初期出力画像を生成し、後続の 2 つの DBGNN は、解像度を 4×4 倍ずつ連続的に増加させ、最初の後続 DBGNN は 64×64→256×256 を実装し、2 番目の後続 DBGNN は 256×256→1024×1024 を実装する。最初の DBGNN には 20 億のパラメータがあり、最初の後続 DBGNN には 6 億のパラメータがあり、2 番目の後続 DBGNN には 4 億のパラメータがあり、合計で 30 億のニューラルネットワークパラメータがある。

3.クレーム

141 特許のクレーム 1 は以下の通りである。

1. 1 台以上のコンピュータによって実行される方法であって、

自然言語のテキストトークンのシーケンスを含む入力テキストプロンプトを受信し、

入力テキストプロンプトのコンテキスト埋め込みのセットを生成するために、テキストエンコーダニューラルネットワークを使用して入力テキストプロンプトを処理し、

入力テキストプロンプトで説明されているシーンを描写する最終的な出力画像を生成するために、拡散ベースの生成ニューラルネットワークのシーケンスを通じて文脈埋

め込みを処理し、該拡散ベースの生成ニューラルネットワークは以下を含む、

初期拡散ベースの生成ニューラルネットワークは、以下の動作を行うように構成されている:

コンテキスト埋め込みを受信し、

出力として初期解像度を持つ初期出力画像を生成するために、コンテキスト埋め込みを処理し、

1 つ以上の後続の拡散ベースの生成ニューラルネットワークはそれぞれ、以下のように構成されており、

それぞれの入力を受信し、該入力は、以下を含み、

(i)文脈埋め込み、および (ii) それぞれの入力解像度を持ち、シーケンス内の先行する拡散ベースの生成ニューラルネットワークによって出力として生成されたそれぞれの入力画像、

それぞれの入力を処理して、それぞれの入力解像度よりも高いそれぞれの出力解像度を有するそれぞれの出力画像を出力として生成し、

ここで、後続の拡散ベースの生成ニューラルネットワークごとに、それぞれの入力を処理してそれぞれの出力画像を生成することは、以下を含む：

それぞれの出力解像度を有する潜在画像をサンプリングし、

一連のステップにわたって潜在画像をそれぞれの出力画像にノイズ除去するステップであって、一連のステップの最終ステップではない各ステップについて、以下のステップを含む：

前記ステップの潜在画像を受信し、

前記ステップのそれぞれの入力と潜在画像を処理して前記ステップの推定画像を生成し、

前記ステップの推定画像のピクセル値を動的に閾値化し、

前記ステップの推定画像を少なくとも使用して次のステップの潜在画像を生成する。

4. 本特許に関連する論文

本特許に関する論文“Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”[1]が、Google の Chitwan Saharia 氏らにより公表されている。

[1] Chitwan Saharia, et al. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding” arXiv:2205.11487v1 [cs.CV] 23 May 2022

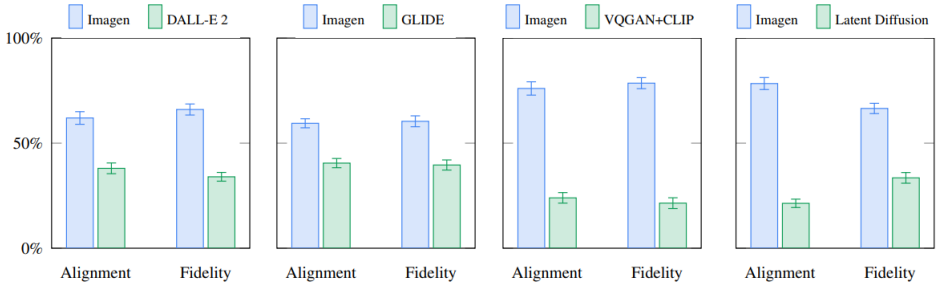
本論文では高いレベルのフォトリアリズムと深いレベルの言語理解を備えたテキストから画像への拡散モデルである Imagen が紹介されている。Imagen は、テキストを理解するための大規模なトランスフォーマ言語モデルの力に基づいて構築されており、高忠実度画像生成における拡散モデルの強さに依存している。本論文での重要な点は、テキストのみのコーパスで事前トレーニングされた汎用の大規模言語モデル（例:T5）が、画像合成のためのテキストのエンコードに驚くほど効果的であるということである。Imagen の言語モデルのサイズを大きくすると、画像拡散モデルのサイズを大きくするよりも、サンプルの忠実度と画像とテキストの配置の両方が大幅に向上する。

下記テーブルは、FID(Fréchet Inception Distance)スコアを使用して COCO 検証セットで Imagen を評価した結果を示す。

Model	FID-30K	Zero-shot FID-30K
AttnGAN [76]	35.49	
DM-GAN [83]	32.64	
DF-GAN [69]	21.42	
DM-GAN + CL [78]	20.79	
XMC-GAN [81]	9.33	
LAFITE [82]	8.12	
Make-A-Scene [22]	7.55	
DALL-E [53]		17.89
LAFITE [82]		26.94
GLIDE [41]		12.24
DALL-E 2 [54]		10.39
Imagen (Our Work)		**7.27**

Imagen は、COCO で最先端のゼロショット FID を 7.27 で達成し、DALL-E 2 の同時作業や COCO でトレーニングされたモデルよりも優れている。

次に、DrawBench を使用して、Imagen を DALL-E2、GLIDE、Latent Diffusion、および CLIP ガイド付き VQ-GAN と比較する。下記図は、Imagen と 3 つの各モデルを一対一で比較した人間の評価結果を示している。



評価者がモデル A、モデル B を好んだ時間、または画像忠実度と画像とテキストの配置の両方について無関心だった時間の割合を報告する。すべてのカテゴリと評価者にわたってスコアを集計する。人間の評価者は、画像とテキストの配置と画像忠実度の両

方で、他のすべてのモデルよりも Imagen を非常に好んでいることがわかった。下記図は、Imagen により生成された画像である。

Imagen の詳細は下記サイトに示されている。

imagen.research.google

以上

著者紹介

河野英仁

河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コース、生成 AI ビジネスコース修了。

AI 特許コンサルティング、医療 AI 特許コンサルティングの他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 3」、「ブロックチェーン 3.0(共著)」がある。