AI 特許紹介(71)
AI 特許を学ぶ！究める！
～MUSE 特許～

2024 年 12 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
特許出願人　Google
出願日　2023 年 12 月 15 日
公開日　2024 年 6 月 20 日
公開番号 WO2024130137
発明の名称　マスクされた生成トランスフォーマーによるテキストから画像への生成

137 特許は、LLM から抽出したテキスト埋め込みを用いて、離散トークン空間でのマスクされたモデリングタスクでトレーニングすることにより、拡散モデルまたは自己回帰モデルよりも大幅に効率よくテキストから画像へ変換処理を行うことが可能となる Muse 技術に関する。

2.特許内容の説明
拡散モデルまたは自己回帰モデルなどの従来の技術は効果的であるが、多くの場合、膨大な計算リソースおよびサンプリング反復を必要とする。本明細書で説明するシステムおよび方法は、離散トークン空間でマスクされたモデリングタスクを実装することに

より、これらの課題に対処し、モデルの効率を大幅に向上させる。

図 1 は、テキストから画像への生成モデルの例と、そのモデルのトレーニングに使用されるフレームワークの例を示している。

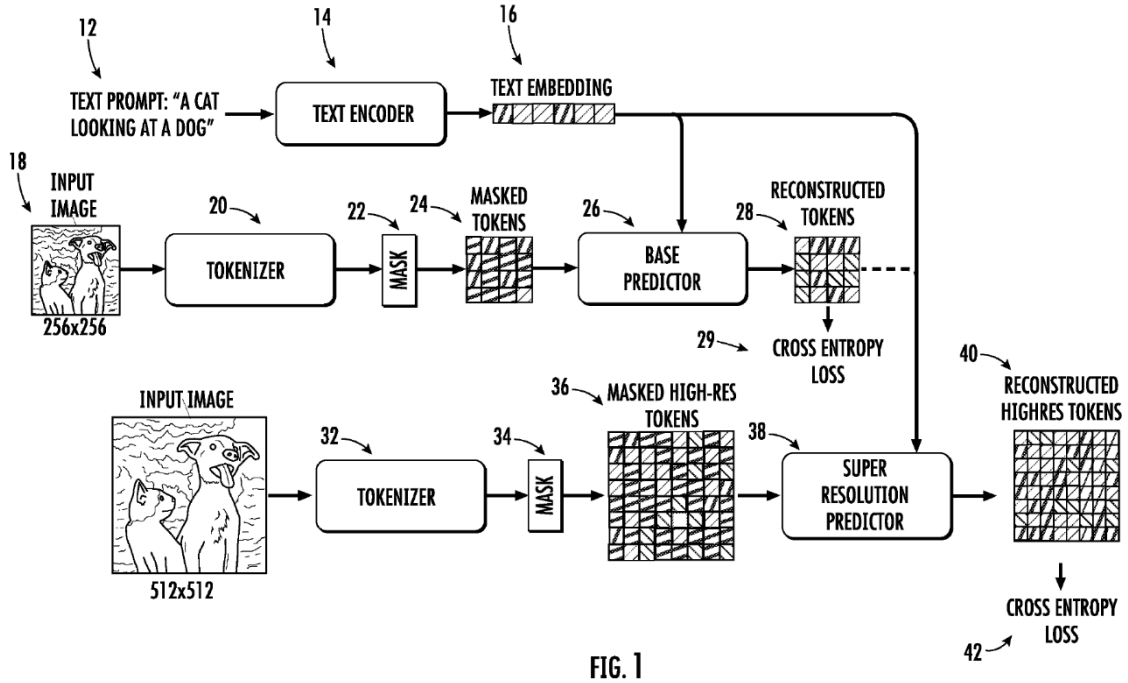


FIG. 1

図 1 に示すように、トレーニングプロセスの最初のステップでは、コンピューティングシステムによってトレーニングタプルが取得される。トレーニングタプルは、テキストプロンプト（図の 12）と、テキストプロンプトで説明されているコンテンツを視覚的に表す対応するトレーニングイメージを含む。

その後、コンピューティングシステムは、図 1 の 14 で示されるテキストエンコーダを使用してテキストプロンプトを処理する。このテキストエンコーダは、大規模なテキストデータコーパスでトレーニングされた Google の T5 モデルなどの既存の言語モデルである。テキストエンコーダは、テキストプロンプトを受け取り、図 1 の 16 で示されるテキスト埋め込みに変換する。このテキスト埋め込みは、潜在空間におけるテキストプロンプトの数値表現であり、その意味情報と構文情報をキャプチャする。

テキスト埋め込みが生成されると、コンピューティングシステムは、図 1 の 18 で示されるトレーニング画像の低解像度バージョンを処理する。この処理は、図の 20 でマークされている最初のトークナイザーモデルを使用して実行される。このトークナイザーモデルは、画像を離散トークンのシーケンスに変換するようにトレーニングされた VQGAN (Vector Quantized Generative Adversarial Networks) モデルなどの画像エ

ンコーダにすることができる。トークナイザーモデルは、画像の中間表現として機能する第1の調整トークンセットを生成する。トークナイザーは、量子化された空間でトークンを生成できる。可能な値の連続体ではなく、可能なトークン値の（量子化された)リストを持つことで、クロスエントロピー損失をエンコーダ出力に適用できる。

VQGANモデルは、入力画像を学習済みコードブックからのトークンのシーケンスに変換する量子化レイヤーを備えたエンコーダとデコーダを含む。一部の実装では、エンコーダとデコーダを畳み込みレイヤーのみで構築して、さまざまな解像度の画像をエンコードできるようにする。2つのVQGANモデルをトレーニングして、1つはダウンサンプリング比率16、もう1つは比率8で、それぞれベースモデルと超解像度モデルのトークンを取得できる。これらのトークンは離散的な性質を持っているため、出力時にクロスエントロピー損失を使用して、後続のステージでマスクされたトークンを予測できる。

図1の22で示される次のステップでは、コンピューティングシステムが最初の調整トークンセットの1つ以上をランダムにマスクし、24とラベル付けされた最初のマスクトークンセットを生成する。このマスクプロセスは、画像の不完全な表現を作成するために実行され、モデルは後続のステップでこの表現の完成を試みる。

次に、マスクされたトークンの最初のセットは、図1の26でマークされている基本ビジュアルトークン予測モデルによって処理される。このモデルは、以前に生成されたテキスト埋め込みを条件としている。基本ビジュアルトークン予測モデルは、マスクされたトークンを予測するようにトレーニングされ、画像の低解像度バージョンに関連付けられた潜在的なビジュアルトークンの最初のセット（図1の28）を生成する。この予測では、マスクされたビジュアルトークンとテキスト埋め込みの両方が考慮され、基本的にテキストの説明とビジュアル表現の間のギャップが埋められる。

ベース予測モデル26は、投影されたテキスト埋め込みと画像トークンを入力とするマスクされたトランスフォーマーである。画像トークンはランダムにマスクされ、特別な[MASK]トークンに置き換えられる。ベースモデルは、セルフアテンションブロック、相互アテンションブロック、MLPブロックなど、いくつかのトランスフォーマー層を使用して特徴を抽出する。出力層でMLPを使用して、マスクされた各画像埋め込みをVQGANコードブックのサイズに対応するロジットのセットに変換する。

基本視覚トークン予測モデル26のトレーニングは、図1の29で表される第1損失関数を使用して実行できる。この損失関数は、第1セットの条件付けトークンからマス

クされたトークンを、第 1 セットの潜在視覚トークンから予測されたトークンと比較する。この比較は予測モデルの改良に役立ち、予測されたトークンがマスクされたトークンを正確に表すことを保証する。クロスエントロピー損失を、グラウンドトゥルーストークンラベルをターゲットとして適用できる。基本モデル 26 は、トレーニング中の各ステップですべてのマスクされたトークンを予測するようにトレーニングする。

コンピューティングシステムは、図 1 の 30 で示されるトレーニングイメージの高解像度バージョンを処理する。この処理は、図の 32 で示される第 2 のトークナイザーモデルを使用して実行できる。第 1 のトークナイザーモデルと同様に、この第 2 のトークナイザーモデルも一連の調整トークンを生成するが、これらのトークンはイメージの高解像度バージョンに関連付けられている。

前のマスキングステップと同様に、コンピューティングシステムは、第 2 の調整トークンセットの 1 つ以上をマスクし、図 1 で 36 とラベル付けされた第 2 のマスクトークン セットを生成する。これらのマスクトークンは、第 1 の潜在視覚トークンセットとともに、図で 38 とラベル付けされた超解像度視覚トークン予測モデルによって処理される。このモデルは、画像の高解像度バージョンに関連付けられた第 2 の潜在視覚トークン セット（図 1 の 40）を生成する。この超解像度モデルは基本視覚トークン予測モデルによって生成された低解像度トークンを高解像度トークンに変換またはマッピングする。

超解像度視覚トークン予測モデル 38 は、図 1 の 42 で示される第 2 の損失関数を使用してトレーニングする。この損失関数は、第 2 の条件付けトークンセットからマスクされたトークンと、第 2 の潜在視覚トークンセットから生成されたトークンを比較する。この比較により、超解像度モデルがさらに改良され、高解像度トークンがマスクされたトークンを正確に表すことが保証される。

3.クレーム

137 特許のクレーム 1 は以下の通りである。

1. 計算効率を向上させてテキストから画像への生成を実行するように構成されたコンピューティングシステムにおいて、

1 つ以上のプロセッサと、および

以下をまとめて保存する 1 つ以上の非一時的なコンピュータ読み取り可能な媒体を含み、

基本視覚トークン予測モデル、超解像度視覚トークン予測モデル、および視覚トー

クンデコーダーモデルを含む機械学習によるテキストから画像への生成モデルと、

コンピューティングシステムによって実行されると、コンピューティングシステムに操作を実行させる命令とを含み、操作には以下が含まれる、

画像の内容をテキストで説明するテキストプロンプトに関連付けられたテキスト埋め込みを取得し、

第１の解像度に関連付けられた潜在的な視覚トークンの第１のセットの１つ以上を予測するために、基本視覚トークン予測モデルを使用してテキスト埋め込みを処理し、

第１解像度より大きい第２解像度に関連付けられた潜在視覚トークンの第２セットの１つ以上を生成するために、超解像度視覚トークン予測モデルを使用して潜在視覚トークンの第１セットを処理し、

合成画像を生成するために、潜在的な視覚トークンの第２セットを視覚トークンデコーダーモデルで処理し、合成画像は、テキストプロンプトによってテキストで説明された画像コンテンツを表す。

4. 本特許に関連する論文

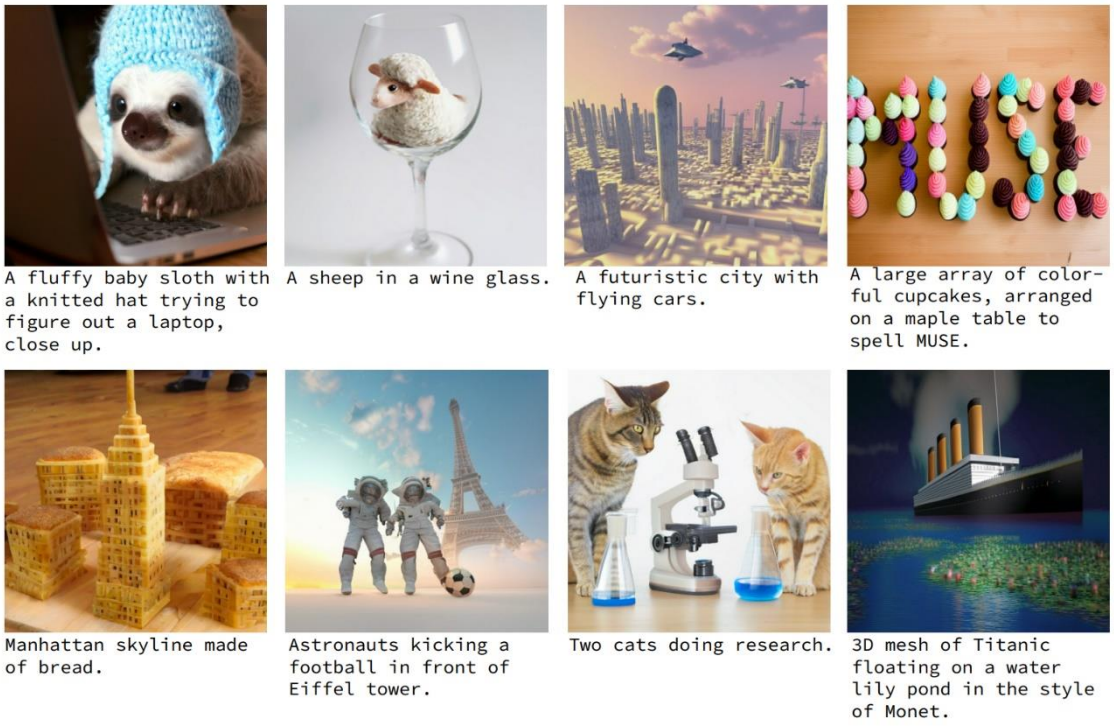
本特許に関する論文“Muse: Text-To-Image Generation via Masked Generative Transformers”[1]が、Google の Huiwen Chang 氏らにより公表されている。

Muse は、離散トークン空間でのマスクされたモデリングタスクでトレーニングされる。つまり、事前トレーニング済みの大規模言語モデル(LLM)から抽出されたテキスト埋め込みが与えられれば、Muse はランダムにマスクされた画像トークンを予測するようにトレーニングされる。Imagen や DALL-E 2 などのピクセル空間拡散モデルと比較すると、Muse は離散トークンを使用し、必要なサンプリング反復回数が少ないため、大幅に効率が高くなる。

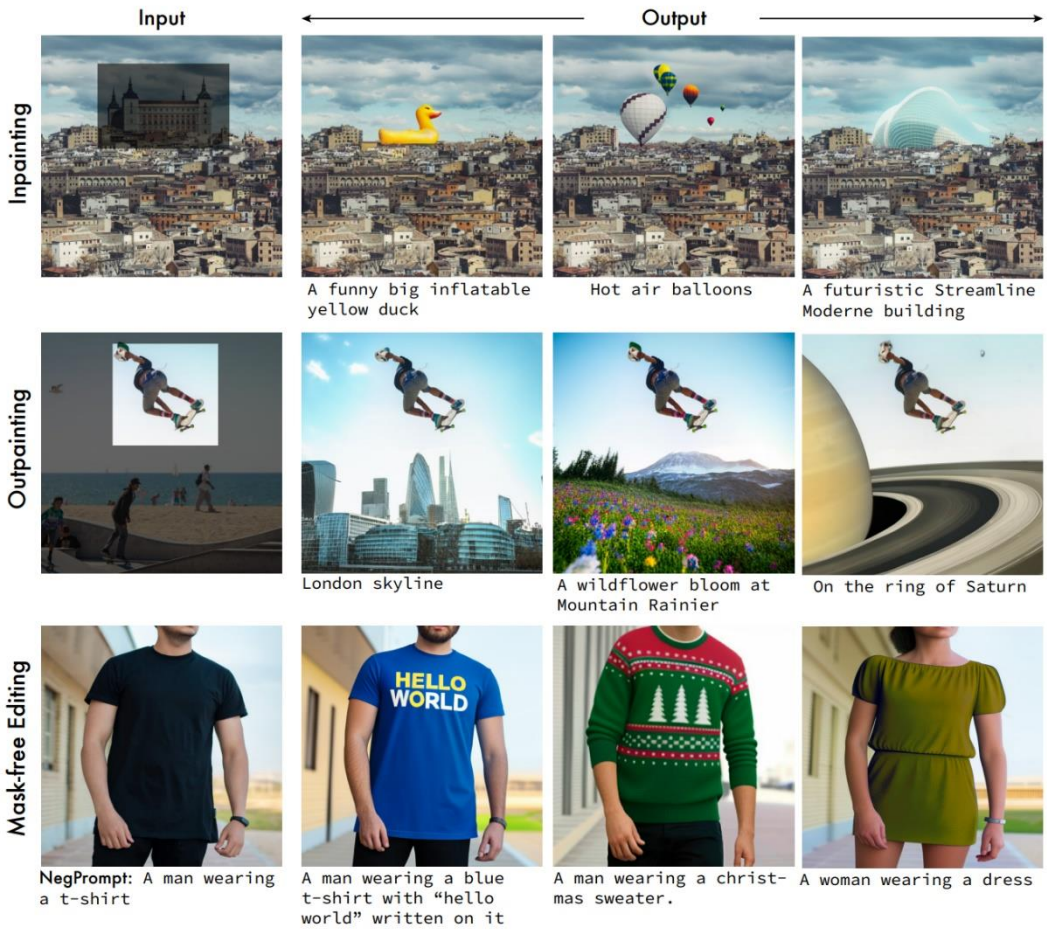
事前トレーニング済みの LLM を使用すると、きめ細かな言語理解が可能になり、忠実度の高い画像生成や、オブジェクト、その空間関係、ポーズ、カーディナリティなどの視覚概念の理解につながる。本論文の 900M パラメータモデルは、CC3M で新しい SOTA を達成し、FID スコアは 6.06 である。

下記図はテキストからの画像生成を示す。

[1] Huiwen Chang, et al. “Muse: Text-To-Image Generation via Masked Generative Transformers” arXiv:2301.00704v1 [cs.CV] 2 Jan 2023



また Muse は、下記図に示すように、モデルのファインチューニングや反転を行うことなく、インペインティング、アウトペインティング、マスクフリー編集など、さまざまな画像編集アプリケーションを直接実行することもできる。



Muse については下記サイトにて詳しく解説されている。

https://muse-model.github.io/

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。