人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第27回)
SIG-AM-27-02

CLIP を用いた画像ランキングによるパラメータ最適化に基づいた絵本の挿絵生成

Generated Images for Picture Book based on Parameter Optimization by Image Ranking Using CLIP

齋藤 優也[1]　黄 潤和[2]

Yuya Saito[1],　Runhe Huang[2]

[1] 法政大学大学院　情報科学研究科
[1]Graduate School of Computer and Information Sciences, Hosei University
[2] 法政大学　情報科学部
[2] Faculty of Computer and Information Sciences, Hosei University

Abstract: We propose a system for generating images from the text of a picture book to assist in image formation when reading. In the proposed system, each paragraph of the picture book is summarized, and an image is generated using VQGAN-CLIP of Text-to-Image model based on the summarized text on the top of the original content. In the process of generating an image, we also propose a ranking method for determining parameters (the seed value and learning times) based on the CLIP image score and the loss value of image generation. In the experiments, the relationship between the loss value and the learning time was investigated, and the learning time parameter was determined in consideration of the execution time.

1　はじめに

近年，様々なデータ化が進み，読書をする媒体も紙などの書籍から電子化された電子書籍へと変化している. スマートフォンの普及が進んだ背景もあり，場所や時間を問わずに簡単に読書を行うことが可能となっている．電子書籍を提供する Kindle など様々なサービスでは，読書をサポートするための様々な機能が搭載されている．分からない単語の意味の検索を行う単語検索機能や，メモをしたい箇所を保存しておくハイライト機能，文字を読みやすくするための拡大機能などがサポートされている．これらの機能は紙媒体の時代から行われていることを簡単に行えるようにしただけではあるが，電子書籍ならではの機能として，文章を自動で要約して読書の時間を短縮するような機能も注目を浴びている．

そこで，本研究では文章の内容を解析して，読書のサポートをする機能として，自動で本文から挿絵を生成することによって，読者のイメージ形成をサポートする機能を提案する．本研究によって提案されるシステムでは，各パラグラフに対して挿絵をそれぞれ生成する．各パラグラフの文章の内容を読み取るために，PEGASUS[1]による文章要約を利用し，パラグラフの内容を 1~2 文にまとめ上げる．要約された文章を基に Text-to-Image と呼ばれる文章からその内容に見合った画像を生成する技術で挿絵を生成する．今回は VQGAN-CLIP[2]，[3]と呼ばれるモデルを利用する．挿絵を生成する際は複数枚生成し，ランキングを行うことで最適な挿絵を生成できるようにする．本研究では，特に絵本の文章から画像の生成を行う．絵本は元々文章と画像がペアとなっているため，挿絵として画像を生成することに適している文章であると考えられるからである．最終的に絵本の各パラグラフに対して挿絵を生成し，提案システムによる絵本の作成を行えることを目的とする．

2　関連研究

2.1　文章要約

自然言語処理分野において，文章要約は入力とする文章から簡潔でかつ正確な要約を出力することを目的に研究が行われてきた．一般的に，文章要約に

は，入力の文章内から重要な部分を断片的に得て要約する抽出型要約と，入力の内容に沿った文章を生成することによる抽象的要約の 2 通りに大きく分けることができる．

抽出型要約には，文章をグラフ構造で表現し，文や単語の関係性を基に要約するグラフベースの TextRank[4]や，文章全体のトピックを算出し，そのトピックにあった文章を抽出するトピックベースの LSA(Latent Semantic Analysis)[5]を利用した手法など様々な手法が取り入れられてきた．

一方で抽象的要約においては，人が要約を作るように，文章の意味を理解したうえで適切な要約を生成する手法である. 抽象型要約を実現するためには，長文の内容理解能力，文章の情報整理能力，新たな文章を生成する能力など，自然言語処理分野でも難しい処理が求められる．これらの複雑な処理を実現するために，機械学習が多く取り入れられている．文章データと要約文章を含む高品質な教師付きデータセットが多く公開されていることも機械学習が盛んになっている理由である．特に，RNN や Transformer を応用した Encoder-Decoder モデルが主流となっている．さらに，近年自然言語処理分野で大きな注目を浴びている BERT[6] を応用した PEGASUS によって，高品質な要約を生成することが可能となった．

2.2 Text-to-Image

Text-to-Image タスクとは，任意のテキストから，そのテキストの内容に沿った画像を生成するタスクのことである．従来，このタスクではあらかじめテキストと画像のペアを用意し，入力のテキストに対してそれらの画像を組み合わせる手法がとられてきた．しかし，機械学習分野の発展や豊富なデータセットの取得が可能となったことから，汎用的な画像が生成可能となった．このタスクは，アート生成やコンピュータによるデザインの補助などに応用できると考えられている．

様々な手法が研究されているタスクであるが，近年では特に，深層学習モデルを取り入れた研究が盛んになりつつある．深層学習モデルの一つである敵対性ネットワーク(GAN)を応用し，テキストを入力として与えることで直接画像を生成する手法がある．最初に GAN を応用した GAN-INT-CLS[7]をはじめ，高解像度画像が生成可能となる StackGAN[8]のように GAN を応用した多くの研究が行われている．その他にも機械学習モデルでは, OpenAI から発表された DALL-E[9]と呼ばれるモデルが注目を浴びている．自然言語処理分野で革新的な性能を示した GPT-3[10]を応用し，膨大なデータセットを学習させることで汎用的な画像を生成することが可能となった．これにより，モデルが未知のデータに対しても画像生成が可能となるゼロショット学習が可能となり，優れた性能を示すことができた．

3 提案手法

本研究では，絵本の文章を入力テキストとして，そのテキスト内容に見合った画像を生成するためのシステムを提案する．提案システムのアーキテクチャは図 1 となる．

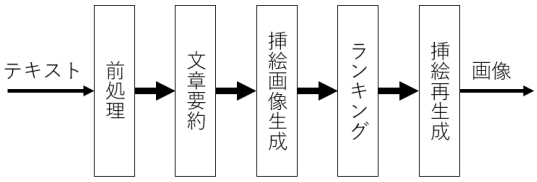

図 1：提案システムのアーキテクチャ

入力されたテキストに前処理によって整形したのち，機械学習モデルによって文章要約を行い，要約文章を獲得する．その後，Text-to-Image のモデルによって，要約文章から画像を生成する．また，複数枚画像を生成したのち，ランキング方式で画像生成に必要なパラメータを決定する．最後に，決定したパラメータを用いて出力とする挿絵を再生成する．

なお，今回使用する文章は，英文の絵本を対象としている．前処理や要約，Text-to-Image には英文を利用したデータセットを使用する．

3.1 前処理

対象とする絵本の本文に対して前処理を行う．前処理は自然言語処理分野で一般的な手法で，文章中のノイズを除去し，機械のパフォーマンスを安定させることが目的である．提案システムで行う前処理は以下のとおりである．

・ストップワードの除去

“a”や”the”，”is”などの情報量の少ない単語をテキストから取り除くことで，重要な単語に焦点を当てることが可能となる．

・不要な記号の除去

文章中に含まれる，”!”や”?”などの記号を除去し，”.”に変換させる．文章内の曖昧性を減らし，パフォーマンスを安定させる．

3.2 文章要約

文章要約では，Text-to-Image のモデルに与える入力の文章を生成することを目的とする．各パラグラフの挿絵を生成するためには，そのパラグラフの内容を最も示す文章を入力とすることが適している．そのための手段として，今回は文章要約による重要文の生成を行う. 2.1 で文章要約に関する様々な研究について述べたが，本研究では，文章要約に特化した事前学習モデルである Pretraining with Extracted Gap-sentences for Abstractive Summarization Sequence-to-sequence models (PEGASUS)を用いた．PEGASU によって，各パラグラフに対して，1~2 文程度の要約文章を得ることが可能となった．

PEGASUS の基本的な構造は，BERT と同様に Encoder-Decoder モデルであるが，最大の特徴は Gap Sentences Generation(GSG)という学習方法である．この GSG という方法は, 事前学習に用いる学習方法が適用するタスクに似ていればいるほど，より高速かつ高い性能を発揮できるという方法に基づいている．つまり, 文章要約用にモデルを学習するのであれば，学習方法は実際に文章要約に似た方法を利用することが好ましいということである．BERT などの事前学習モデルの学習方法では，Masked Language Model(MLM)と呼ばれる学習方法がある．この方法は，入力文章の一部分をランダムにマスクし，マスクした箇所を学習モデルが予測する方法で学習を行う．GSG では MLM の方法を，より文章要約のタスクに適するように工夫が加えられている．まず，MLM ではランダムにマスクする箇所を決定していたが，文章要約にはランダムに決定する方法は適していない．そこで GSG では，重要な文章をあえてマスクし，その箇所を予測させるように学習を行う．

3.3 画像生成

Text-to-Image の生成モデルは 2.2 で述べたように様々なモデルがあるが，今回は VQGAN と CLIP を組み合わせた VQGAN-CLIP と呼ばれるモデルを使用する．VQGAN-CLIP は様々な画像を生成することが出来ること，オープンソースとして誰でも使用可能となっていることから，AI によるアート生成の分野でも注目を浴びているモデルとなる．

このセクションでは，構成要素となっている VQGAN と CLIP について，そして VQGAN-CLIP としての全体のアーキテクチャについて述べる．

3.3.1 VQGAN

VQGAN は Vector Quantization GAN である．豊富な表現度を持つ（GAN）によって得た特徴マップに対して，ベクトル量子化（VQ）のプロセスを加えることで，高解像度の画像を生成したり，合成したりすることが可能となったモデルである．

従来の研究では，より高解像度の画像を生成するためのモデルとして，画像のピクセルの情報を直接 Transformer に与え，Transformer の利点である長距離の依存関係を応用するアプローチが検討された．画像内のコンテキストをより詳しく理解することが出来る手法ではあるが，Transformer の構造上，画像のピクセル数が増えるに応じて, 必要となる GPU のスペックが高くなってしまう問題点があり，高解像度の画像を生成することが難しかった．そこで，VQGAN はベクトル量子化の方法で利用し，画像内のピクセル情報をクラスタリングし，各グループにまとめ上げた．これにより，Transformer に与える情報量が削減し，より高解像度の画像の生成を行うことが可能となった．

本研究では，Variational Autoencoder(VAE)として VQGAN を利用することで，画像を生成するため Image Generator としての役割を担うこととなる．

3.3.2 CLIP

画像とテキストの関係性を得ることのできる事前学習モデルとして Contrastive Language-Image Pre-training(CLIP)が提案された．CLIP はインターネット上から収集した 4 億もの画像とテキストをペアとした事前学習モデルである．テキストには Transformer を用いた Text Encoder と, 画像には ResNet-50，Vision Transformer(ViT)を用いて Image Encoder を学習させる．それぞれの Encoder を通して，画像とテキストの Embedding を取得し，それぞれのコサイン類似度が最大となるように Encoder を学習させることで，与えられた画像に対してキャプションを生成する事前学習モデルを実現した．

さらに，一般的な事前学習モデルの精度は，学習したデータセットに依存し，未知のデータに対して精度を上げることが出来なかったが, CLIP は学習したデータ以外でも精度が高くなるような Zero-shot 学習が可能となった．

本研究では，入力テキストの Embedding を得るための Text Encoder と，VQGAN によって生成された画像の Embedding を得るための Image Encoder として使用される．また，CLIP は画像とテキストの相関性をスコアで獲得することが可能であり，後述するランキングによるパラメータ決定の際にも利用する．

3.3.3 VQGAN-CLIP

先ほど説明した, VQGAN と CLIP を組み合わせた

VQGAN-CLIP によって Text-to-Image を行う．VQGAN-CLIP のアーキテクチャは，図 2 となる．

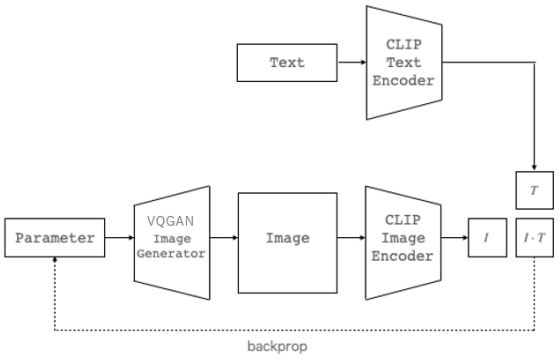

図 2：VQGAN-CLIP のアーキテクチャ

VQGAN-CLIP は，生成したい内容のテキストと VQGAN が画像生成するために必要なパラメータを入力として与える．パラメータには，学習回数の上限を決める iter 値，生成される画像の内容を決定する seed 値，学習率を示す step size などがある．これらの入力を基に，画像を生成するためのプロセスは次の通りとなる．

①与えられたパラメータから Image Generator である VQGAN が画像を生成する．

②①で生成した画像に対して，CLIP の Image Encoder によって画像の Embedding を得る．同様に，入力として与えられたテキストに対しては，CLIP の Text Encoder によってテキストの Embedding を得る．

③画像 Embedding とテキスト Embedding を比較し，2 つの loss 値を求める．

④③で求めた loss 値が小さくなるように逆誤差伝播法で繰り返し学習を行う．

以上の①~⑤のプロセスを経て，生成する画像が入力されたテキストの内容に近づいていく．ただし，上記で説明したパラメータは自分で決定する必要があり，最適なパラメータを探すことは困難である．そのため，次セクションで述べる方法によって，パラメータを決定する方法を提案する．

3.4 ランキングによるパラメータ決定

3.3.2 で画像生成に必要なパラメータを決定するために複数枚画像を生成し，その中でランキングを行い，ランクが高い画像を生成したときに使用したパラメータを使用する．今回のランキングによって決定するパラメータは，生成される画像の内容を決定する seed 値と，画像を生成する際の学習回数を決定する iter 値の 2 つとする．3.3.3 で述べた step size は学習率を示すパラメータであり，生成される画像の内容に対する影響が少ないため，今回は 0.01 に固定して毎回学習を行う．

Seed 値の決定には，CLIP の画像スコアを利用する．3.3.2 で述べたように CLIP には，テキストと画像を与えることで，画像の内容がテキストにどれくらい即しているかを図るスコアを算出することが出来る．このスコアを生成したすべての画像で比較することによって，最もよい Seed 値を調べることが可能となる．

本研究では，与えられたテキストに対して，ランダムな seed 値を与え，25 枚の画像を生成する．各画像は生成後，CLIP による画像のスコアがそれぞれ算出される．それに加え，各画像を学習し生成した際の loss を基に以下の式(1)でランキングスコアを出す．

$$\text{Rank score} = \frac{CLIPによる画像スコア}{画像を生成したときのloss} \quad (1)$$

CLIP による画像スコアは，テキストと画像の相関性を示すスコアで，高い数値であればあるほど相関性があるといえる．また，画像を生成したときの loss は，低い値であればあるほど良い画像が生成できているといえる．つまり，Rank Score は高い値であるほど良い画像が生成されたといえる．

続いての iter 値の決定には，画像が生成されたときの loss 値を利用する．loss 値は先述した通り，低い値であれば，テキストと画像の間で誤差が少ないと評価することが出来る．つまり，loss 値が低い際の iter 値が学習回数として最も適しているといえる．今回は iter 値が 50 ずつ増えるごとに画像を保存し，loss 値を記録する．4 で行う実験によって，各 iter 値での画像の loss を参考にし，最も適した iter 値を決定する．

3.5 再生成

3.4 で決定したパラメータであるランキングが最も高い Seed 値を VQGAN-CLIP のパラメータとし，同じテキストで画像を再度生成する．学習回数 50 回ごとに画像と loss 値を記録し，学習終了後に loss 値が最も低いときの画像をシステムの出力とする．

4 実験

提案システムにおける，パラメータ決定方法の精度を確かめるために，実際に画像を生成した際の iter と loss の関係性を調査する．実験に使う絵本のデータは，無料の公開されている絵本から文章のみを抽出する．抽出した文章は，1 本分の絵本で 21 パラグ

ラフ分を使用し，21 枚の画像を生成する．学習回数の上限を 100 回，250 回，500 回として学習を行う．その際に，iter 値が 50 ずつ増えるごとに loss 値を計測する．

文章要約に用いた PEGASUS は，cnn_dailymail を使用する．CNN と Daily Mail によって書かれた 30 万件強のニュース記事を含む英語のデータセットである．文章要約の機械学習に使われる一般的なデータセットである．

VQGAN には，ImageNet のデータセットを学習に用いたモデルを使用する．ImageNet は画像認識分野における標準的なデータセットであり，訓練用の画像データが 120 万枚用意されており，学習に十分なデータセットである．

CLIP は，提案元である OpenAI が公開されている事前学習モデルを使用する．3.3.2 で述べたようにインターネット上から収集した 4 億にも及ぶ画像とテキストのペアを学習に利用している．

使用する GPU は, NVIDIA GeForce GTX 1080, 8GB を使用し，画像生成し，実行時間を計測する．

5 実験結果と考察

学習回数を上限 100， 250， 500 回として学習を行った際に, 学習回数が 50 ずつ増えるごとに loss を計測する．その後，どの iter 値が最も低い loss 値となったかを，パラグラフごとにまとめる．実行時間に関する結果を表 1，iter 値と loss 値の関係を示した結果は表 2 となった．

表 1：上限学習回数を行った際の平均実行時間

学習回数	100	250	500
実行時間(s)	48.1166	114.4604	225.50585
100との差		66.34385	177.3893
250との差			111.04545

表 1 の実行時間から，学習回数が増えると，実行時間も増加していることが分かる．これは，画像を生成するプロセスを繰り返し行っているためである．

次に，表 2 の iter 値と loss 値の関係性についての結果を見る．iter 値を 100 とした時は，ほとんどのパラグラフで 100 回学習したときの loss 値が最も低くなった．次に，iter 値が 250 の時も同様に，ほとんどのパラグラフで 250 回学習した時の loss 値が最も低くなった．続いて，iter 値が 500 とした時は，平均して iter 値が 350 の時が最も loss 値が低くなる事が分かる．このことから，最も低い loss 値の時の iter 値を探すには最低でも 500 回程度の学習が必要であると考えることが出来る．

また，各 iter 値での loss 値の平均を見ると，上限 100 回から上限 250 回の時に約 0.06 下がっているのに対して，上限 250 回から 500 回の時には 0.02 ほどしか下がっていない．

表 2：各パラグラフに対して，学習回数を定めたときの loss 値が最も低くなる iter 値

	学習回数					
	100		250		500	
パラグラフ	iter	loss	iter	loss	iter	loss
1	100	0.755	250	0.692047	400	0.6839
2	100	0.858606	150	0.777282	400	0.738503
3	100	0.841877	250	0.734142	350	0.710746
4	100	0.783286	250	0.747524	250	0.744691
5	100	0.840254	250	0.78228	250	0.7875
6	50	0.76071	250	0.67835	300	0.662174
7	100	0.776059	250	0.689184	300	0.686342
8	100	0.806283	200	0.801432	250	0.720158
9	100	0.842293	250	0.743442	300	0.740903
10	50	0.75486	250	0.702702	400	0.686614
11	100	0.731423	200	0.689242	300	0.695985
12	100	0.733282	250	0.709531	450	0.702995
13	100	0.775316	250	0.758003	350	0.746047
14	100	0.774019	250	0.708915	500	0.709487
15	100	0.757584	250	0.720492	250	0.710087
16	100	0.824111	250	0.694436	500	0.666173
17	100	0.857855	250	0.758764	450	0.714617
18	100	0.792461	250	0.757109	400	0.721365
19	100	0.789375	250	0.73341	350	0.712016
20	100	0.789269	250	0.737588	300	0.714016
21	100	0.75595	250	0.688265	300	0.676064
平均	95.2381	0.79047	240.4762	0.728769	350	0.710971

表 1 と表 2 の結果から，ランキング前の画像を複数枚生成する際には，学習回数の上限を 250 回にし，ランキング後の再生成では，500 回にすることが最も良いと考えた．複数枚の画像を生成するときは，実行時間が増えすぎず，loss も最適なときにほとんど変わらないためである．一方，ランキング後は質の良い画像を生成したいため，loss 値が最も低くなる iter 値が見つけられる 500 回が最適であると考えた．

6 今後の展望

今回提案するシステムの評価として，画像生成の精度に関する評価を行う予定である．今回使用するモデルでの VQGAN-CLIP は，教師なし学習であるためモデルの評価が難しい．一般的に GAN などの教師なし学習では，教師あり学習と異なって，正解データがないためにモデルをどのように評価するかという問題がある．一般的に使われる性能評価の指

標として，Inception Score や FID などの指標があるが，ある程度の実際の画像が必要である．本研究のシステムでは，入力される文章に制限を設けていないため，生成される画像も多種多様となることもあり，特定の画像を用意することが困難であると考えた．そのため，生成された画像の評価には客観的評価として，アンケート方法を用いる予定である．アンケートによって評価する内容は，①文章から画像の整合性があっているか(Text→Image)，②画像から文章を想像できるか(Image→Text)の評価を行う．アンケートで評価する内容は以下の通りである．

①要約文章から生成した画像は，その文章の内容にあっているかを検討する．生成された画像内のオブジェクト(人や建物など)， 背景や風景，時間などはあっていると思うかついて評価する．

②生成された画像から，どのような文章を生成できるかを検討する．画像内にどのようなオブジェクトがあるように見えるか，風景や背景などの様子はどのように見えるかについて評価を行う．

①と②の評価によって，テキストと画像の双方向からの評価が出来ると考えている．

また，よりシステムの生成する画像の質を向上するために以下のような方法を検討する．

・生成画像の統一性

現在の画像は各パラグラフを別々に生成するために，それぞれの画像に統一性がない．生成画像の統一性は，絵本を作成するうえでは必要なものであると考えられる．画像生成モデルには，自分で画像を当てることによってその画像を参考に，入力テキストに合わせた画像を新たに生成することが可能であるため，これによって統一性を保証することが出来ると考えている．

・生成画像の画風変換

画像生成モデルは多様なデータを学習させているため，入力テキストに特定のテキストを追加することで，画風を変えることが可能である．例えば，生成する画像を「葛飾北斎」が書いたようにしたい場合は，入力テキストに「by Hokusai Katsushika」と追加することで，あたかも「葛飾北斎」が書いたような画像を生成することができる．

7 まとめ

今回，読者のイメージ形成のサポートとして絵本の文章から画像を生成するシステムを提案した．提案システムでは，生成するための文章を決めるために，文章要約の方法を用いて，事前学習モデルでのある PEGASUS を利用して，要約文章の生成を行った．その後，要約文章を入力テキストとして Text-to-Image のモデルである VQGAN-CLIP から，テキストの内容にあった画像を生成することが出来た．また，画像生成の際に必要なパラメータの決定を，複数枚の画像をランク付けし，最も適したパラメータを決定した．今後は，生成された画像の評価をアンケートによって行い，システムの調査を行っていく．

謝辞

本研究を進めるにあたり，ご指導いただいた黄潤和教授に感謝します．また，日頃の議論を通じて多くの知識や示唆をいただいた黄研究室の皆様に感謝いたします．

参考文献

[1] Zhang， Jingqing， et al. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." International Conference on Machine Learning. PMLR，2020.

[2] Esser， Patrick， Robin Rombach， and Bjorn Ommer. "Taming transformers for high-resolution image synthesis." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[3] Radford， Alec， et al. "Learning transferable visual models from natural language supervision." arXiv preprint arXiv:2103.00020 (2021).

[4] Mihalcea， Rada， and Paul Tarau. "Textrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[5] Ozsoy，Makbule Gulcin，Ferda Nur Alpaslan，and Ilyas Cicekli. "Text summarization using latent semantic analysis." Journal of Information Science 37.4 (2011): 405-417.

[6] Devlin， Jacob， et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[7] Reed， Scott， et al. "Generative adversarial text to image synthesis." International Conference on Machine Learning. PMLR， 2016.

[8] Zhang， Han， et al. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.

[9] Ramesh， Aditya， et al. "Zero-shot text-to-image generation." arXiv preprint arXiv:2102.12092 (2021).

[10] Brown， Tom B.， et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).