生成型要約に基づくWebページのサムネイル生成

前田　直宏†　　山本　岳洋†

† 兵庫県立大学 大学院情報科学研究科　〒651–2197 兵庫県神戸市西区学園西町 8-2-1
E-mail: †ad22p062@gsis.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp

あらまし　本研究では Text2Image モデルを用いて Web ページの主題を表したサムネイルを生成する手法を提案する．具体的には，モデルに入力するプロンプトを Web ページを要約することで複数生成し，（1）記事の主題を表している，（2）画像として自然な出力が期待できる，という 2 点の観点から適したプロンプトを分類して選択する．この条件を満たすプロンプトを取得するために本研究では，ランダムフォレストを用いて分類モデルを構築する．実験の結果，自然言語処理タスクで高い性能を示す BERT よりも高い精度で分類を行うことができた．また，構築した分類モデルを用いて CNN ニュース記事と旅行ブログなどの Web ページを対象としたサムネイル生成を行い，人手による評価を行う．サムネイル評価実験の結果，記事のタイトルをプロンプトとするベースラインに比べ，提案手法によって生成されるサムネイルの方が記事のサムネイルに合っているとして好まれた．

キーワード　情報検索，Text2Image，プロンプトエンジニアリング，画像生成，要約文生成

1　はじめに

Web ページの概要を表すサムネイルは，ユーザ所望の情報を獲得する助けになる．例えば検索結果の表示をテキストベースだけではなく，Web ページ内の画像やデザインを用いた 1 枚のサムネイルも表示することで，探している情報と関係のある情報をより正確に獲得することが可能である [4]．また，一度閲覧した Web ページの約 80% は再度閲覧することが明らかになっている [1] [3]．ブラウザに備わったブックマーク機能や閲覧履歴，記憶を辿ることで再度閲覧したページにたどり着くことができる一方，検索結果表示欄にサムネイルも表示することで，一度訪れたサイトへの再訪問が容易に行えることが報告されている [22]．このようにユーザが情報を獲得するとき，Web ページの概要を表したサムネイルは情報検索の効率を上昇させる．

Web ページの内容を表したサムネイルの生成方法としては，Web ページの上部のスクリーンショットにロゴを付与するものや，ページ内に掲載されている写真を組み合わせたもの，インターネット上からその Web ページと関連性が高い画像を用いたものなど様々なサムネイルが提案されている [9] [15] [22]．また近年の検索エンジンでは，Web ページの上部に掲載されているアイキャッチ画像とも呼ばれる画像をサムネイル代わりとして使用されることもある．このアイキャッチ画像はフリー素材サイトなどで取得する必要があるが，必ずしも記事の内容に合った画像が存在するわけではない．また内容と異なる画像を用いた場合，誤った情報の獲得を促してしまう可能性がある．そのため，Web 検索におけるサムネイルは記事の内容や主題に合っていることが重要である．

そこで本研究では Text2Image モデルを用いて，Web ページの主題を表したサムネイルの生成を行う．Text2Image モデルによる生成画像は，入力されたプロンプトと呼ばれるテキストに基づいた画像が生成される．そのため本研究では Web ページの記事テキストをもとにプロンプトを生成し，得られたプロンプトを用いてサムネイルの生成を行う．具体的には，まず記事テキストの要約文を複数生成する．その後，（1）記事の主題を表している，（2）画像として自然な出力が期待できる，という 2 点の観点からサムネイル生成に適したプロンプトを分類して取得する．このプロンプト分類モデルとして本研究ではランダムフォレストを構築する．そして得られたプロンプトを Text2Image モデルに入力することでサムネイルを生成する．

2　関連研究

2.1　サムネイルを用いた情報検索支援

Web ページのサムネイルを表したサムネイルの研究は多く行われている．Zhou らは Web ページを検索する時に使用したクエリを強調するサムネイルを提案している [23]．Teevan らは Web ページのタイトル，ロゴ画像，注目度が高い画像を用いたサムネイルを提案している [22]．また，稲垣らは Web ページに対する顕著性マップを用いたサムネイルを提案している [25]．顕著性マップとは，人が画像を認識する際に注視しやすい領域を画像で定量的に表現したものである．Web ページを閲覧する際に注視されやすい上位 10 の領域を 1 枚の画像に並べて配置した集約画像を生成することで，ページ内容の理解補助に繋げている．これらのサムネイルで使用している画像は，Web ページ内に含まれる画像を抽出し使用している．しかし内部画像が存在しない場合，サムネイルを作成することが困難であった．そこで Jiao らはインターネット上からその Web ページと関連性が高い画像を取得し，その画像を用いたサムネイルを提案している [9]．Web 検索時以外にも Web ページの概要を表したサムネイルの利用方法が提案されている．例えばサムネイルを Web ブラウザのタブ管理に使用することが提案されている [12]．Web ページの左右余白部分に閲覧した Web ページのサムネイルを配置することで，一度閲覧した Web ページへの

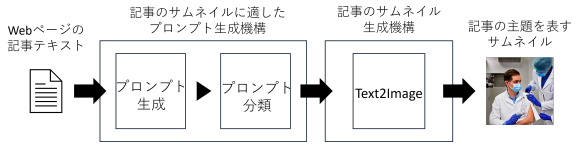

図 1　記事の主題を表したサムネイルの生成手法.

素早い切り替えを行うことができる.

2.2　画像生成モデル

画像生成モデルとしては，VAE（Variational AutoEncoder）[10] や GAN（Generative Adversarial Network）[6]，拡散モデル [8] などの手法が数多く提案されていた．VAE はエンコーダとデコーダで構成されており，エンコーダで入力データを潜在変数に落とし込み，デコーダで潜在変数から画像を生成する．GAN は敵対的生成ネットワークとも呼ばれ，生成器と識別器からなる 2 つのネットワーク構造を持つ．生成器が偽のデータを生成し，識別器が正解データと偽データを見分ける学習を行うことで生成器が本物に近く質の高い画像を生成することができる．拡散モデルは画像にノイズを付与する拡散過程と，ノイズを取り除いて画像を生成する逆拡散過程に分かれており，高い質の画像を生成することができる．また近年では，テキストから画像を生成する Text2Image モデルが多く提案されており，Stable Diffusion [20] や DALL-E2 [18] など質の高い画像を生成することができる．これらの画像生成するモデルでは，プロンプトと呼ばれるテキストを入力する必要があり，プロンプトの質が生成画像の質に大きく影響する．そのため，プロンプトの質を高めるプロンプトエンジニアリングの手法も多く提案されている.

2.3　サムネイル生成とプロンプトエンジニアリング

Text2Image モデルの登場により，テキストからサムネイル生成を行う研究が行われてきている．Liu らはニュース記事の挿絵を生成した [14]．Web 検索以外においてもサムネイル生成の研究が行われている．Shinomoto らは，ニュース記事に対してミーム画像を生成する手法を提案している [21]．また，プロンプトエンジニアリングによるサムネイル生成を行った研究もおこなわれている．Hao らは強化学習なども用いてプロンプトの最適化を行った上で画像の生成を行っている [7]．Liu らは，プロンプトの様々な条件のプロンプトで画像を生成することによってプロンプトと生成画像の関係を調査している [13].

3　提 案 手 法

本研究では Web ページの記事から，その記事の主題を表したサムネイルを生成する．提案手法の全体像を図 1 に示す．サムネイルの生成手法としては，テキストから画像を生成する Text2Image モデルを用いる．この Text2Image モデルには出力したい画像に対するプロンプトを入力する必要がある．そこで記事本文を要約することで Text2Image モデルに入力するプロンプトを複数生成する．複数生成されたプロンプトの内，サムネイル生成に適したプロンプトを分類し，最も適したプロンプトを Text2Image モデルに入力することで記事の主題を表したサムネイルを生成する．次節では Text2Image モデルに入力するプロンプトを生成するプロンプト生成機構と，記事のサムネイル生成機構の 2 つについて説明する.

3.1　記事本文の要約によるプロンプト生成

サムネイル生成を行うために必要であるプロンプトの生成について説明する．Web ページの記事本文には多くの情報が含まれているため，Text2Image モデルに全ての文を入力することによって記事の主題を表した画像を生成することは難しい．そこで本研究ではタイトルを除いた記事本文を T5 [17] を用いて要約することで，記事の要点を短くまとめたプロンプトを生成する．しかし，Web ページの記事は章立てられていることが多く要点が多く存在すると考えられる．そのため，全文をまとめて要約するのではなく全文をいくつかの文章に分け，各文章を要約することでプロンプトを生成する．本文を区切って入力する理由としては 2 つある．1 つ目は T5 の入力トークン数には上限があり，本文全文を入力することが困難な場合があるからである．2 つ目は Web ページの文章は章立てられていることが多く，文の位置によって述べられているトピックが異なるからである．あまりにも多くのトピックが含まれた文章の要約を行うと情報量が減少してしまう恐れがあるため，数文に区切って入力する.

記事本文 article を文 s_i の系列として以下のように表現する.

$$\text{article} = s_1 s_2 \dots s_n \tag{1}$$

ここで n は記事本文 article の文数を表す．ただし，1 文はピリオド，エクスクラメーションマーク，クエスチョンマークで区切る．また，記事本文から系列の一部を取り出す関数である窓関数を以下に定義する.

$$\text{window}(\text{article}, i, j) = s_i s_{i+1} \dots s_{i+j-1} \tag{2}$$

この窓関数を用いて j 文で構成される文章 sentence_i を以下に定義する.

$$\text{sentence}_\text{i} = \text{window}(\text{article}, i, j) \tag{3}$$

ただし，i は取り出したい部分の開始位置を表す.

こうして得られた各文章 sentence_i を要約することでプロンプトを生成する.

3.2　サムネイル生成に適したプロンプトの分類

3.2.1　問 題 定 義

Text2Image モデルによって生成される画像の質は，入力するプロンプトの質に大きく依存する．そのため，画像で表現することが難しいプロンプトを入力すると質の低い画像が生成される．例えば「方法」や「安い」といった単語を画像で表現することは難しい．そのため入力として適するプロンプトを選ぶ必要がある．また，3.1 節で生成された複数のプロンプト内には，記事の主題を含んでいるプロンプトと含んでいないものが

表 1　ランダムフォレストで使用する特徴量.

特徴量変数名	説明
interrogative_words	疑問詞を含むか
word_length	単語数
title_prompt_sim	タイトルとプロンプトの類似度
LexRank	LexRank
adjective_count	形容詞の数
noun_count	名詞の数
proper_noun_count	固有名詞の数
verb_count	動詞の数
proportion_adjective	形容詞の割合
proportion_noun	名詞の割合
proportion_proper_noun	固有名詞の割合
proportion_verb	動詞の割合
concreteness	各単語の具体度の平均

あると考えられる．本研究の目的は記事の主題を表したサムネイルの生成であるため，記事の主題を含んでいないものはサムネイル生成に適していないプロンプトである．そこで，プロンプトの良し悪しを分類する分類モデルを構築することで，記事の主題を表したサムネイル生成に適したプロンプトを選ぶ．ただし，ここでの良いプロンプトとは以下の 2 つを十分に満たすものである．

- 記事の主題を含んだプロンプトである．
- 画像で表現することが可能なプロンプトである．

本研究ではプロンプトの良さを 3 段階に分割し，以下の様にラベルを定義する．

$$\text{label} = \begin{cases} l_1 & \text{（サムネイル生成に適さないプロンプト）} \\ l_2 & \text{（どちらでもないプロンプト）} \\ l_3 & \text{（サムネイル生成に適したプロンプト）} \end{cases} \quad (4)$$

プロンプト分類モデルに入力された複数のプロンプトから，最もサムネイル生成に適したプロンプト l_3 を 1 つ取得することが目標である．

3.2.2　プロンプト分類モデル

本研究では，プロンプトの分類モデルとしてランダムフォレストを用いる．ランダムフォレストは複数の決定木を用いたアンサンブル学習法の 1 つである．数値やカテゴリといった様々な種類の特徴量を扱うことができるため，良いプロンプトの条件を種類の異なる複数の特徴量で表現することができると考えられる．ランダムフォレストで使用する特徴量を表 1 に示す．特徴量の選択は主に 3 つの軸に基づき決定した．1 つ目は記事の主題を表すかである．特徴量として，LexRank [5] とタイトルとプロンプトの類似度を選択した．LexRank とは，文書から各文をノードとみなしたグラフ構造を作り出し，各ノード間の類似度を計算することでそれぞれの文の重要度を算出するアルゴリズムである．本研究では 1 つのノードに 1 文を割り当てるのではなく，プロンプト生成に用いた要約元の文章を 1 つのノードとして類似度を計算する．要約元の文章の重要度を計算することで，各プロンプトの文書内での重要度を算出する．これらの文章間の類似度を計算するには Sentence-BERT[1] [19] を用いてベクトル化を行い，コサイン類似度を計算する．Web ページのタイトルは文書の内容を簡潔に表していることがある．特にニュース記事などは読者に記事の情報を素早く理解してもらうために，タイトルが内容を簡潔に表現していると考えられる．そこでタイトルとプロンプトの類似度を特徴量に用いた．類似度は，タイトルとプロンプトのコサイン類似度を計算することで算出する．2 つ目は画像として自然な出力が期待できるかである．画像で表現することが容易な単語は具体度が高いと考えられるため，特徴量としてプロンプトに含まれる単語の具体度を用いる．具体度の計算には，様々な単語に具体度が付与された辞書 [2] を使用する．これは 40,000 単語に対して 1 から 5 の具体度を付与したものである．各プロンプトに対する具体度を以下の様に定義する．

$$\text{concreteness}_i = \frac{\sum_i^n \text{conc}(w_i)}{n} \quad (5)$$

ここで w_i はプロンプトを構成する単語を表し，$\text{conc}(w_i)$ は単語 w_i の具体度を返す関数である．ただし，具体度は 1 から 5 の間の実数であり，最も具体度が高い場合 5 を取る．3 つ目は，1 つ目と 2 つ目の両方を含むもしくはそれ以外に当てはまるものである．特徴量として，疑問詞の有無，プロンプトの単語数，品詞の数と割合を選択した．プロンプトに疑問詞が含まれる場合，例えばプロンプトに「何」,「なぜ」,「どこで」といった疑問詞を含む場合，表現することが困難であると考えられる．そのため疑問詞の有無を特徴量に用いた．またプロンプトの長さは生成画像の質に影響することが報告されている [24]．そのため特徴量にプロンプトの長さを用いた．ただしプロンプトの長さとはトークン数を表す．また，プロンプト中に名詞や固有名詞が含まれていなければ抽象度が高く，うまく画像を生成できないと考えられる．同様に形容詞や動詞が含まれていない場合でも抽象度が高まると考える．そのため，品詞の数とトークン数に対する割合を特徴量に選択した．

3.3　Web ページの主題を表したサムネイル生成

分類モデルによって得られたサムネイル生成に適したプロンプトを用いて，Web ページの内容に合ったサムネイルの生成を行う．サムネイルの生成手法としては，テキストから画像を生成する Text2Image モデルを用いる．

4　評 価 実 験

4.1　サムネイル生成に適したプロンプト分類の評価

構築したプロンプト分類モデルの性能評価を行う．分類モデルとしてはランダムフォレスト，XGBoost，BERT を用いる．

4.1.1　データセット

プロンプトを生成するためのデータセットとして CNN ニュース記事を用いる．CNN はアメリカ合衆国の国際的なケーブルテレビニュースネットワークであり，全世界にニュースを伝え

1：https://huggingface.co/sentence-transformers/all-mpnet-base-v2

表 2　ランダムフォレストの学習に用いたハイパーパラメータ.

パラメータ	値
決定木の深さ	20
葉に必要な最小サンプルサイズ	1
分割後ノードの最小サンプルサイズ	2
決定木の数	300

ている．そのため，記事に対するサムネイルやタイトルなどの質は高いと考えられる．CNN ニュースサイトからニュース記事 577 件を収集し，各ニュース記事のページから URL，タイトル，記事本文，サムネイルを取得した．これらを取得する際には Python のライブラリである Newspaper3k[2]を使用した．

4.1.2　実 験 条 件

収集したニュース記事の本文を用いてプロンプトのデータセット構築を行う．プロンプトの生成を行うための言語生成モデル T5 は，t5-one-line-summary[3]を使用する．T5 へ本文を 3 文ずつに区切って入力することで記事についてのプロンプトを複数生成する．生成された全プロンプト数は 7,255 である．生成された各プロンプトに対してプロンプトの良さを表すラベルを付与した．記事に掲載されているサムネイルとプロンプトから生成される画像との類似度を計算し，その値によってラベルを付与する．類似度が 0.0〜0.4 であればサムネイル生成に適さないプロンプト，0.4〜0.7 であればどちらでもないプロンプト，0.7〜1.0 であればサムネイル生成に適したプロンプトとラベルを付与する．ただし，プロンプトから画像を生成するモデルとして Stable Diffusion [20] を使用した．また，類似度を計算するときには 1 つのプロンプトから 5 枚画像を生成し，類似度の平均をとった．複数画像を生成することで，生成のたびに画像の質が異なることを考慮した．画像同士の類似度を計算するために，CLIP [16] を使用する．

4.1.3　ランダムフォレスト

プロンプトを分類する 1 つ目のモデルは，ランダムフォレストである．使用する特徴量として，表 1 を用いた．学習データ，テストデータを 8:2 の比率で分割し，さらに学習データを 8:2 の比率で学習データと検証データに分割した．学習データと検証データを用いてグリッドサーチと交差検証を行い，ハイパーパラメータのチューニングを行った．求まったハイパーパラメータを表 2 に示す．チューニングを行った後，学習データと検証データを用いて再学習を行った．

4.1.4　XGBoost

2 つ目のモデルは XGBoost である．使用した特徴量はランダムフォレストと同様に表 1 を用いた．データについてもランダムフォレストと同様に学習データ，テストデータを 8:2 の比率で分割し，さらに学習データを 8:2 の比率で学習データと検証データに分割した．学習データと検証データを用いてグリッドサーチと交差検証を行い，ハイパーパラメータのチューニングを行った．求まったハイパーパラメータを表 3 に示す．チューニングを行った後，学習データと検証データを用いて再学習を行った．

表 3　XGBoost の学習に用いたハイパーパラメータ.

パラメータ	値
学習率	0.2
決定木の深さ	5
決定木の数	300
各決定木のサンプル抽出比	0.9

表 4　BERT のファインチューニングに用いたハイパーパラメータ.

パラメータ	値
バッチサイズ	16
学習率	2×10^{-5}
早期終了	3 epoch
最適化手法	Adam
損失関数	交差エントロピー

4.1.5　BERT

3 つ目のモデルは BERT である．事前学習済みモデルとして，インターネット上の大規模なテキストコーパスを用いて学習された bert-base-uncased[4]を使用し，この事前学習済みモデルをファインチューニングして使用する．ファインチューニングのデータとしては，4.1.1 節で述べたデータセットを用いる．ファインチューニングに用いたハイパーパラメータを表 4 に示す．BERT への入力は，プロンプトである．先頭には特殊トークンの [CLS] を付与し，プロンプトの末尾には [SEP] を付与して入力する．ただし入力トークン長が最大入力長に満たない場合，特殊トークンの [PAD] で埋める．

4.1.6　評 価 尺 度

プロンプトをクラス l_k と分類するタスクにおいて，実際にクラス l_k である場合を正例，そうでない場合を負例とする．またクラスが l_k であるプロンプトを実際に正例とモデルが予測できたときを真陽性（TP），負例と予測したときを偽陰性（FN）とし，クラスが l_k でないプロンプトを正例とモデルが予測した時を偽陽性（FP），負例と予測したときを真陰性（TN）とする．評価尺度としては適合率（Precision），再現率（Recall），F 値のマクロ平均を用いる．適合率 P，再現率 R，F 値 F は以下の式で求められる．

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$F = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}} \tag{8}$$

マクロ平均とは，各クラスの評価指標の平均である．マクロ平均適合率 P_{macro}，マクロ平均再現率 R_{macro}，マクロ平均 F 値 F_{macro} は以下の式で求められる．

2：https://newspaper.readthedocs.io/en/latest/

3：https://huggingface.co/snrspeaks/t5-one-line-summary

4：https://huggingface.co/bert-base-uncased

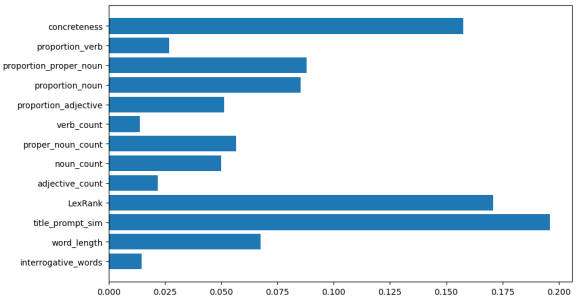

図 2　ランダムフォレストにおける各特徴量の重要度.

$$P_{\text{macro}} = \frac{1}{K}\sum_{i}^{K} P_i \tag{9}$$

$$R_{\text{macro}} = \frac{1}{K}\sum_{i}^{K} R_i \tag{10}$$

$$F_{\text{macro}} = \frac{1}{K}\sum_{i}^{K} F_i \tag{11}$$

ここで K は分類するクラス数であり，$K = 3$ である.

4.2　サムネイル生成に適したプロンプトの分類実験結果

各モデルによるプロンプトの分類結果を表 5 に示す．実験の結果，ランダムフォレストはマクロ平均適合率，マクロ平均再現率，マクロ平均 F 値全てにおいて最も高い値であった．また，良いプロンプトであるラベル l_3 の適合率，再現率，F 値も最も高い値であった．反対に自然言語処理タスクにおいて高い性能を示す BERT は最も低い値であった．深層学習に基づく分類手法よりも特徴量による分類の方が高い精度で分類を行えたことがわかる．また，最も分類精度が高かったランダムフォレストにおける各特徴量の重要度を図 2 に示す．図 2 より，タイトルとプロンプト類似度，LexRank，各単語の具体度の平均が特に重要な特徴量であったことがわかる.

4.3　Web ページの主題を表したサムネイル生成の評価

構築したプロンプト分類モデルを用いて選ばれたプロンプトを使用し，記事に対するサムネイルの生成を行う．使用するプロンプト分類モデルとしては，分類精度が最も高かったランダムフォレストを使用する．モデルがサムネイル生成に適したプロンプトと最も高い確率で予測した 1 つのプロンプトを画像生成モデルに入力する．もしプロンプトの中にサムネイル生成に適したプロンプトと予測したものが無い場合，最も高い確率でどちらでもないプロンプトと予測した 1 つのプロンプトを画像生成モデルに入力することで記事に対する画像を生成する．また比較手法として記事のタイトルをプロンプトとし，タイトルを画像生成モデルに入力することで記事に対する画像を生成する．本評価実験では，提案手法で生成された画像と比較手法で生成された画像について評価を行う.

4.3.1　データセット

CNN ニュース記事と旅行記事が掲載されているブログなどのサイトを用いる．CNN ニュース記事からは，Business，Entertainment，Health，Sport，Travel のジャンルから各 5 記事ずつの計 25 記事を使用する．ただし，CNN ニュース記事は分類モデルの学習，評価に使用していないデータを用いる．また，ブログ等の旅行記事サイトは 14 サイトからランダムに計 25 記事収集し使用する．ただし，ページのタイトルが「人気観光地 10 選」のようなまとめ記事は使用しない．CNN ニュース記事からは Newspaper3k ライブラリを用いて本文を取得し，旅行記事サイトからは手作業で本文を取得した.

4.3.2　実験条件

画像生成モデルとして Stable Diffusion と DALL-E2 を使用する．DALL-E2 は Stable Diffusion と同様にテキストから画像を生成することができるモデルである．1 つの Web ページ記事に対し，Stable Diffusion と DALL-E2 を用いて 1 枚ずつ画像を生成する．すなわち評価者は，1 つの記事に対して（提案手法とベースライン）×（Stable Diffusion と DALL-E2）による生成の 4 枚の画像を評価する．生成された画像について，同研究室の学生 2 名と著者によるアンケートを通じて評価を行う．評価アンケートでは以下の点について評価を行ってもらう.

（1）　画像が記事の主題を表していますか.

（2）　画像として自然ですか.

（3）　記事のサムネイルとしてどちらの画像が適していますか.

評価 1 では，各サムネイルについて記事の主題を表しているかを評価する．評価者には 5 段階評価で回答を求めた．ここでは 5 が最高点とした．評価 2 では，各サムネイルについて，画像としての自然であるかを評価する．不自然な画像とは，画像が表している風景，状態を読み取ることができない画像のことを指す．評価 1 と同様に評価者には 5 段階評価で回答を求めた．ここでも最高点は 5 点とした．評価 3 では，提案手法とベースラインどちらで生成された画像が記事のサムネイルとして適しているかを回答してもらう．なお，評価者にはどの画像がどの手法によって生成されたかなどの情報を公開せずに評価を行う.

4.3.3　評価尺度

評価 1 と評価 2 では，各生成モデルについて評価者の点数の平均を用いて手法の比較評価を行う．Web ページのジャンル g について，評価で用いる点数 Score_g は各記事 Article_i の評価点の平均であり，以下の式で求められる.

$$\text{Score}_g = \frac{\sum_i^n \text{Article}_i}{n} \tag{12}$$

ただし，n はジャンル g に含まれる記事数である.

評価 3 では，各生成モデルについて提案手法とベースラインで好まれた記事数の平均を用いて評価を行う．Web ページのジャンル g について，評価で用いる点数 Score_g は以下の式で求められる.

$$\text{Score}_g = \frac{\sum_i^n f(\text{Article}_i)}{n} \tag{13}$$

$$f = \begin{cases} 1 & (\text{Article}_i\text{のサムネイルとして適している}) \\ 0 & (\text{Article}_i\text{のサムネイルとして適していない}) \end{cases} \tag{14}$$

ここで f はサムネイルが Article_i のサムネイルとして適していれば 1 を返し，適していなければ 0 を返す関数である.

表 5　プロンプト分類モデルの評価結果.

分類モデル	ラベル l_3 の適合率	ラベル l_3 の再現率	ラベル l_3 の F 値	マクロ平均適合率	マクロ平均再現率	マクロ平均 F 値
ランダムフォレスト	**0.756**	**0.523**	**0.617**	**0.796**	**0.683**	**0.725**
XGBoost	0.704	0.478	0.570	0.760	0.659	0.696
BERT	0.593	0.467	0.523	0.670	0.648	0.656

4.4　Web ページの主題を表したサムネイルの評価結果

評価 1 のサムネイルが記事の主題を表しているかの評価結果を表 6，評価 2 のサムネイルが画像として自然であるかの評価結果を表 7，評価 3 の記事に適したサムネイルはどちらであるかの評価結果を表 8 に示す．表 6 より，全体的にはベースラインの方が記事の主題を表すことができていた．しかし Entertainment と Health のジャンルでは，提案手法のサムネイルの方が記事の主題を表すことができていたことがわかる．また Health のジャンルでは，Stable Diffusion と DALL-E2 両方において提案手法のサムネイルが記事の主題を表すことができた．表 7 より，全体的には提案手法とベースラインで画像としての自然さに違いはなかった．しかし Health のジャンルでは Stable Diffusion と DALL-E2 両方において提案手法の方がスコアが高い結果となった．表 8 より，全体的にはベースラインの方が記事のサムネイルとして選ばれる数が多かった．しかし Health のジャンルでは Stable Diffusion と DALL-E2 の両方において提案手法の方が高いスコアになった．また評価 1，評価 2，評価 3 について一致度を算出するために Krippendorff の α 係数 [11] を計算した．求まった α 係数の値を表 9 に示す．

5　議　　論

5.1　サムネイル生成に適したプロンプト分類

サムネイル生成に適したプロンプトの分類では，自然言語処理タスクで高い性能を示す BERT よりも，決定木をベースとしたアンサンブル学習手法であるランダムフォレストや XGBoost の方が高い分類精度を示した．その理由として，プロンプト分類に使用した情報量の差ではないかと考える．BERT の入力はプロンプトのみである．そのため BERT はプロンプトの単語のみの情報を分類に使用している．一方でランダムフォレストや XGBoost はプロンプトのみではなく，プロンプトの元となった文章の LexRank やタイトルといった記事中の情報も用いている．サムネイル生成に適したプロンプトとは，記事の主題を含んでいる且つ画像で表現することが可能なプロンプトである．プロンプトの単語情報だけでなく記事中の情報も用いていることから，サムネイル生成に適した 2 つの条件を考慮することができたため，ランダムフォレストの分類精度の方が高かったのではないかと考える．

5.2　Web ページの主題を表したサムネイル生成

プロンプトの分類を行うにあたり，良いプロンプトとして（1）記事の主題を表している,（2）画像として自然な出力が期待できる，の 2 つの観点を考慮した．そのため（1）を満たす特徴量として LexRank，タイトルとプロンプトの類似度,（2）の特徴量としてプロンプトの具体度が影響すると考えた．サムネイルに対する評価の点数とプロンプトの特徴量について相関係数を計算することで，特徴量の影響を分析した．相関係数の値を表 10 と表 11 に示す．記事の全ジャンルについて相関係数を計算した結果，記事の主題を表すかの点数と LexRank，タイトルとプロンプトの類似度にそれぞれ正の相関がみられた．画像として自然であるかの点数とプロンプトの具体度にも正の相関がみられた．また，Business と Entertainment 以外のジャンルに絞ると全ジャンルに比べて強い相関がみられた．以上のことから，LexRank とタイトルとプロンプトの 2 つの特徴量は類似度は記事の主題を表すかに影響を与え，プロンプトの具体度は画像の自然さに影響を与えることがわかった．

評価 1 と評価 2 において，Entertainment と Health のジャンルでは提案手法が好まれた．この 2 つのジャンルの中でも特に Health ジャンルはベースラインとの差が大きかった．その理由として，プロンプトの具体度が影響していると考える．Text2Image モデルに入力したベースラインの記事タイトルに比べ，提案手法によるプロンプトの方が具体度の平均が約 0.5 高く，これは他ジャンルには見られない傾向であった．以上の理由から Health ジャンルは他のジャンルと異なり，ベースラインよりも提案手法の方がスコアが高くなったと考える．

評価 3 においては，提案手法によるサムネイルよりもベースラインによるサムネイルの方が好まれる結果であった．CNN ニュース記事では大きく差はないが，CNN ニュースの Travel と旅行記事サイトにおいてはベースラインが大きく好まれる傾向がみられた．旅行系記事には様々な風景についての情報などが含まれていた．そのため主題とは少し異なる内容を含むプロンプトも多く生成されたためにプロンプトの選択で誤ったものが選ばれ，ベースラインに勝ることができなかったのではないかと考える．また，Stable Diffusion では提案手法によるサムネイルが多く好まれ，DALL-E2 ではベースラインによるサムネイルが多く好まれた．その理由として，プロンプトのラベル付けを行ったときに用いた画像生成モデルが Stable Diffusion であったためだと考える．プロンプトに対するラベルは，プロンプトから生成された画像と記事に掲載されているサムネイルとの類似度を計算することで 3 段階のラベルを付与した．このときの画像生成モデルとして本研究では Stable Diffuision を使用した．そのため各プロンプトに付与されたラベルは Stable Diffusion を用いて画像生成を行うときに対するプロンプトの良さであると考えられるため，画像生成モデルによって提案手法と比較手法の評価に差が生じたと考える．

5.3　限　界　点

本研究ではサムネイルを生成するにあたり，T5 によって生

表 6　評価 1：サムネイルが記事の主題を表しているか.

生成モデル	手法	Business	Entertainment	Health	Sport	Travel	旅行記事サイト	全体の平均
Stable Diffusion	提案手法	2.87	**4.13**	**3.20**	4.53	2.20	3.93	3.47
	ベースライン	**3.47**	3.93	2.47	**4.60**	**3.93**	**4.10**	**3.89**
DALL-E2	提案手法	**2.6**	3.07	**3.40**	3.87	3.80	3.65	3.02
	ベースライン	2.4	**4.00**	2.87	**4.13**	**4.07**	**3.80**	**3.65**

表 7　評価 2：サムネイルが画像として自然であるか.

生成モデル	手法	Business	Entertainment	Health	Sport	Travel	旅行記事サイト	全体の平均
Stable Diffusion	提案手法	3.67	**4.93**	**4.27**	3.57	4.53	4.51	4.41
	ベースライン	3.67	4.73	3.53	**4.73**	**4.73**	**4.53**	4.41
DALL-E2	提案手法	**3.53**	4.27	**3.93**	3.38	4.33	3.84	4.17
	ベースライン	3.20	**4.73**	3.60	**4.40**	**4.80**	**4.20**	4.17

表 8　評価 3：記事に適したサムネイルはどちらか.

生成モデル	手法	Business	Entertainment	Health	Sport	Travel	旅行記事サイト	全体の平均
Stable Diffusion	提案手法	**2.67**	**2.67**	**2.67**	**2.67**	0.67	8.33	19.7
	ベースライン	2.33	2.33	2.33	2.33	**4.33**	**16.7**	**30.3**
DALL-E2	提案手法	2.33	1.33	**3.33**	1.67	0.67	8.00	17.3
	ベースライン	**2.67**	**3.67**	1.67	**3.33**	**4.33**	**17.0**	**32.6**

表 9　各評価における Krippendorff の α 係数.

生成モデル	評価 1	評価 2	評価 3
Stable Diffusion	0.59	0.24	0.44
DALL-E2	0.50	0.28	0.33

表 10　評価 1 についてのサムネイルに対する評価点と各特徴量間の相関係数.

記事のジャンル	LexRank	タイトルとプロンプトの類似度
全ジャンル	0.306	0.749
Business と Entertainment 以外	0.921	0.929

表 11　評価 2 についてのサムネイルに対する評価点と各特徴量間の相関係数.

記事のジャンル	プロンプトの具体度
全ジャンル	0.310
Business と Entertainment 以外	0.373

成されたプロンプトの中から記事の主題を表している，画像で表現が可能なプロンプトを分類モデルを用いて選ぶという手法を提案した．そのため T5 によって生成されたプロンプトの中に，サムネイルのための優れたプロンプトが存在しているということが前提条件の手法である．したがって T5 によるプロンプトの生成がうまく行われなかった場合，提案手法によるサムネイルの質は悪くなる．本研究の実験と評価では，T5 によって生成された複数のプロンプトの中に優れたプロンプトが存在するという仮定の下で行った．そのため T5 の性能向上，あるいは他の言語モデルの性能向上によって提案手法によるサムネイルの評価結果が改善される可能性がある．

6　まとめと今後の課題

本研究では，サムネイル生成に適したプロンプトの分類を行うことによって画像生成モデルへの入力を選別し，Text2Image モデルを用いて Web ページの主題を表したサムネイル生成に取り組んだ．提案手法のプロンプト分類モデルとして，ランダムフォレストを構築した．プロンプト分類の結果，提案手法の分類精度が最も高く，自然言語処理タスクで高い性能を示す BERT よりも高い分類性能を示した．また，この構築したプロンプト分類モデルによって選ばれたプロンプトを用いて，Web ページのサムネイル生成を行った．CNN ニュースと旅行記事サイトに対して生成をし評価を行った結果，CNN ニュースの内 Travel のジャンル以外では提案手法によるサムネイルが好まれ，Travel と旅行記事サイトではベースラインであるタイトルをプロンプトとしたサムネイルが好まれた．ベースラインが勝った理由として，タイトルが記事の内容を端的に表現しており強力であったことや，プロンプトの誤分類が提案手法で起きていたことなどが考えられる．今後は，生成するプロンプトの質を高める必要があることやプロンプト分類の精度をさらに向上させる必要がある．

謝辞　本研究は JSPS 科学研究費助成事業 JP21H03774, JP21H03775, JP22H03905, による助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Eytan Adar, Jaime Teevan, and Susan T Dumais. Large scale analysis of web revisitation patterns. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 1197–1206, 2008.

[2] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, Vol. 46,

pp. 904–911, 2014.

[3] Andy Cockburn and Bruce McKenzie. What do web users do? an empirical analysis of web use. *International Journal of human-computer studies*, Vol. 54, No. 6, pp. 903–922, 2001.

[4] Susan Dziadosz and Raman Chandrasekar. Do thumbnail previews help users make better relevance decisions about web search results? In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 365–366, 2002.

[5] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, Vol. 22, pp. 457–479, 2004.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, Vol. 27, , 2014.

[7] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611*, 2022.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, Vol. 33, pp. 6840–6851, 2020.

[9] Binxing Jiao, Linjun Yang, Jizheng Xu, and Feng Wu. Visual summarization of web pages. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 499–506, 2010.

[10] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[11] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[12] Shenwei Liu and Keishi Tajima. Wildthumb: a web browser supporting efficient task management on wide displays. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pp. 159–168, 2010.

[13] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–23, 2022.

[14] Vivian Liu, Han Qiao, and Lydia Chilton. Opal: Multimodal image generation for news illustration. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–17, 2022.

[15] A Porselvi and S Gunasundari. Survey on web page visual summarization. *International Journal of Emerging Technology and Advanced Engineering*, Vol. 3, No. 1, pp. 26–32, 2013.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 5485–5551, 2020.

[18] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, Vol. 1, No. 2, p. 3, 2022.

[19] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[21] Erica K Shimomoto, Lincon S Souza, Bernardo B Gatto, and Kazuhiro Fukui. News2meme: An automatic content generator from news based on word subspaces from text and image. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pp. 1–6. IEEE, 2019.

[22] Jaime Teevan, Edward Cutrell, Danyel Fisher, Steven M Drucker, Gonzalo Ramos, Paul André, and Chang Hu. Visual snippets: summarizing web pages for search and revisitation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2023–2032, 2009.

[23] Allison Woodruff, Andrew Faulring, Ruth Rosenholtz, Julie Morrsion, and Peter Pirolli. Using thumbnails to search the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 198–205, 2001.

[24] Yutong Xie, Zhaoying Pan, Jinge Ma, Luo Jie, and Qiaozhu Mei. A prompt log analysis of text-to-image generation systems. In *Proceedings of the ACM Web Conference 2023*, pp. 3892–3902, 2023.

[25] 稲垣有哉, 岩田一, 白銀純子, 深澤良彰. ウェブページの構造と顕著性マップの組み合わせによる重要領域の視覚化手法. Web インテリジェンスとインタラクション研究会 予稿集 第 14 回研究会, pp. 45–50. Web インテリジェンスとインタラクション研究会, 2019.

表 12　ベースラインとして用いた記事タイトル例.

記事のジャンル	記事タイトル
Business	These popular fast food drive-thrus are getting faster
Entertainment	Late-night hosts return to air post-writers' strike
Health	mRNA vaccine: 5 things to know
Sport	Solheim Cup: Europe retains title after stunning comeback against USA
Travel	The top superyachts at Monaco Yacht Show 2023
旅行記事サイト	Hershey ' s Chocolate World Announces New Experience That Will Immerse Visitors In A Delicious Train Ride

表 13　提案手法によって選ばれたプロンプト例.

記事のジャンル	選ばれたプロンプト
Business	Artificial Intelligence Drive-Thru and Walk-Up Locations at Chick-fil-A
Entertainment	Late-Night Network Talks Revisited
Health	Essence RNA-Based Vaccine Technology
Sport	First blood vs. Ireland in the Solheim Cup
Travel	The peak of the super yachting sector
旅行記事サイト	The Chocolate World Experience

表 14　提案手法が好まれたサムネイル例.

記事のジャンル	ベースライン	提案手法
Business		
Entertainment		
Health		
Sport		
Travel		
旅行記事サイト