DEIM Forum 2025 XX-Y

タイトル重畳料理画像におけるフォントの魅力分析・推定

高木　七海†　　久徳　遙矢††,†　道満　恵介†††,†　駒水 孝裕†　　井手　一郎†

† 名古屋大学　〒 464-8601 愛知県名古屋市千種区不老町
†† 愛知工科大学　〒 443-0047 愛知県蒲郡市西迫町馬乗 50-2
††† 中京大学　〒 470-0393 愛知県豊田市貝津町床立 101
E-mail: † takagin@cs.is.i.nagoya-u.ac.jp, taka-coma@acm.org, ide@i.nagoya-u.ac.jp,
†† kyutoku-haruya@aut.ac.jp, ††† kdoman@sist.chukyo-u.ac.jp

あらまし　インターネットの普及やスマートフォンの利用機会の増加により，Web 上の料理レシピを閲覧する機会や，自身が考案した料理レシピを SNS（Social Networking Service）に投稿する人が増えている．料理レシピの投稿者は，多くの場合，自身の料理レシピを多くの人に閲覧・調理してもらうことを目的としている．そのためには，多数の投稿の中から自身の投稿がより多くの人に閲覧されるように，魅力的な料理レシピを投稿する必要がある．SNS 上の料理レシピ投稿においてサムネイル画像は人の目を惹くための重要な要素の 1 つであるため，我々はサムネイル画像の魅力の分析及び推定に取り組んでいる．料理レシピ投稿では，サムネイル画像にタイトルを重畳することで料理の見た目の魅力に加えてその内容を表現することが多い．本研究は，このような状況に限定し，魅力的なタイトル重畳料理画像の生成を目的に，サムネイル画像に対する印象に大きな影響を与え，人の目を惹くための重要な要素であると考えられるフォントに注目する．そして，入力した料理画像と料理タイトルに対して魅力的なフォントを選択するフォント選択モデルを提案する．また，そのために必要な学習・評価用データセットを構築する．様々なフォントで料理タイトルを重畳した料理画像の魅力度を決定する選好実験を元にデータセットを構築し，それを用いてモデルを構築し，定量的及び定性的に評価した．その結果，定量評価では，67%の正解率で 7 種類のフォントから最も魅力的なフォントを選択できることを示した．また，定性評価では，料理画像や料理タイトルの持つ印象的特徴とフォントのデザイン的特徴との間に強い関連性が見られ，提案モデルがそれらの関連性を適切に学習していることを確認した．

キーワード　料理レシピ，フォント選択，魅力度推定，SNS

1　はじめに

インターネットの普及やスマートフォンの利用機会の増加により，Web 上に多くの料理レシピが存在し，多様な料理レシピを容易に入手可能になった．特に，SNS（Social Networking Service）で料理レシピを手軽に閲覧できるようになり，料理レシピの入手先として SNS を利用する 20 代の割合は 38.1 % であると報告されている [12]．このように，SNS において料理レシピを閲覧する機会が増えている．例えば，Instagram[1]において，2024 年 11 月時点で，ハッシュタグ「#レシピ」を含む投稿は 239 万件あり，大量の料理レシピ投稿が存在している．

Instagram は画像や動画像から視覚情報を得やすいことに加え，ユーザの趣向に合いそうな投稿を自動的に表示する機能により，料理レシピ以外の情報を閲覧している時にも受動的に料理レシピを閲覧できるという特徴がある．また，近いシェア率をもつ SNS である X[2]に比べ，投稿文字数制限が緩く，料理レシピの材料や手順の情報を詳細に投稿するのに十分なテキストを確保できる [14] ．これらのことから，Instagram は料理レシピに高い親和性をもったプラットホームとして活用されている．

料理レシピ投稿者は，多くの場合，自身の料理レシピを多くの人に閲覧・調理してもらうことを目的としている．一方，料理レシピの閲覧者はハッシュタグ検索やおすすめ欄などに表示されたサムネイル群から目を惹いた投稿を閲覧する．この際に，サムネイル群から選択されるためには，目を惹くものであることが重要である．そのため，Instagram の料理レシピ投稿のサムネイル画像には，タイトルが重畳された料理画像（タイトル重畳料理画像）が投稿されることが多い．これは，サムネイル画像という限られた媒体で，視覚情報とテキスト情報を同時に効果的に提示できるためである．そのため，閲覧者は表示された投稿のうち料理画像やそれに重畳されたタイトル情報から直感的に美味しさや手軽さなどの魅力を感じられるものを閲覧すると考えられる．

これらのことから，料理レシピを閲覧してもらうためには，閲覧者の目を惹くタイトル重畳料理画像を作成する必要がある．タイトル重畳料理画像を魅力的にするための要素としては，料理画像の色調や視覚効果フィルタ，重畳する料理タイトルの配置やフォント，色，大きさなどが挙げられる．これらの要素の中でも，料理タイトルのフォントはその料理に対する印象を表現し，フォントの印象が料理特徴に合っているか否かで，サムネイル画像に対する印象に大きな影響を与え，適切なフォントの選択は人の目を惹くための重要な要素であると考えられる．

1：Meta, “Instagram, ” https://www.instagram.com/ ［2025/1/30 参照］.
2：X, “「いま」を見つけよう / X, ” https://x.com/ ［2025/1/30 参照］.

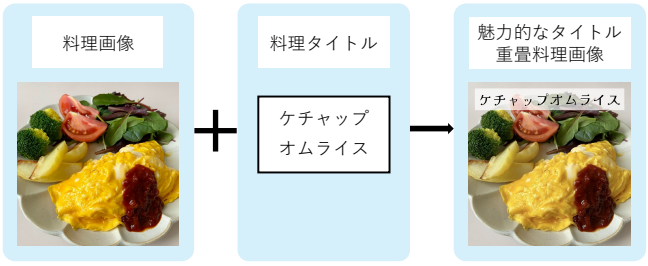


図 1: 魅力的なタイトル重畳料理画像の生成方法の概要.

また，既存のものだけでなく，新たに作成されたフォントを利用することも考えられる．しかし，一般の投稿者が料理特徴を的確に捉えた魅力的なフォントを選択することは必ずしも容易ではなく，また，魅力的なタイトル重畳料理画像のためのフォント生成技術も確立していない．したがって，魅力的なタイトル重畳料理画像の作成を支援する仕組みが必要となる．

本発表では，魅力的なタイトル重畳料理画像の作成を支援する手段の 1 つとして，図 1 に示すように，料理画像とタイトルが与えられた状況において，魅力的なタイトル重畳料理画像の生成を目的とする．そのために，まず，フォントの選択がサムネイル画像の魅力に与える影響を明らかにする選好実験を行なう．具体的には，料理画像に対し，様々なフォントでタイトルを重畳したサムネイル画像を構築し，評価者が魅力的か否か判定し，タイトルのフォントと料理画像の組合せと魅力の関係性を分析する．なお，本研究において，タイトル重畳料理画像におけるタイトルのフォントの印象が料理画像及びタイトルの印象と合致している度合いを魅力度とする．

次に，選好実験の結果から，料理画像の視覚的特徴と料理の内容を表すタイトルの特徴の両方を考慮することで魅力的フォントを推定できることが予想されるため，画像とテキストの両特徴を考慮したマルチモーダルモデルを提案する．その際，選好実験の結果を利用して，データセットを構築する．このデータセットにより，入力した料理画像と料理タイトルに対して魅力的なフォントを自動選択するモデルを学習する．最後に，学習したモデルについて，フォント選択タスクにおける性能を検証し，その有効性を評価する．

本研究は，料理画像と料理タイトルの印象を考慮し，魅力的なフォントを選択するモデルを構築する初めての試みであり，その主な貢献は以下の通りである．

- 料理画像と料理タイトルの組合せの魅力にフォント選択が与える影響について，選好実験から関係性を明らかにする．
- 選好実験の結果から，料理画像と料理タイトルに対する各フォントの魅力度を算出することで，機械学習モデルの学習に適したデータセットを構築する．
- 料理画像と料理タイトルの特徴を統合し，魅力的なフォントを選択するマルチモーダルモデルを構築し，構築したデータセットを用いた実験により，魅力的なフォントを選択できることを示す．

2 関 連 研 究

関連研究として，料理画像とフォントの関係性に注目した研究や，サムネイル画像生成に関する研究がある．ここではまず，画像とフォントの関係性に注目した研究として， 2.1 節で料理画像とフォントの関係性分析に関する研究， 2.2 節で画像特徴を反映したフォント生成に関する研究を紹介する．また，サムネイル画像の関連研究として， 2.3 節で動画像コンテンツを対象としたサムネイル画像の生成に関する研究を紹介する．

2.1 料理画像とフォントの関係性分析に関する研究

料理画像とフォントの関係性分析に関する研究として，笠井ら [15] は料理画像のシズル感とフォントの形態的特性の関係性を分析している．シズル感とは食に対する「美味しそう」という印象のことを言い，この研究はフォントから得られる視覚情報によって美味しそうと人に感じさせるために有効な表現方法を明らかにすることを目的としている．この研究では被験者実験を通して，料理画像に対し，画像から得られる食品の印象に適しており，その美味しさを感じられるフォントを被験者実験を通して明らかにしている．また，Chen ら [3] は，食品の包装に注目し，フォントデザインが包装のデザインに与える美的影響を調査している．具体的には，食品の包装のフォントデザインに対する嗜好を調査し，異なる包装における漢字の視覚的および心理的印象を比較している．これらの研究では，料理画像あるいは包装に対する印象に合致するフォントを明らかにしているが，料理画像にタイトルを付与した状態での魅力度推定は行なっていない．

本研究は，料理画像にタイトルが重畳された状態を対象とし，料理画像とタイトルが視覚的に一体となったデザイン全体の魅力度を考慮した分析を行なう点で先行研究と異なる．また，これらの研究は，料理画像とフォントの関係性を分析するにとどまっており，フォントを選択する手法は提案していない．それに対して，本研究では，タイトル重畳料理画像に対する分析を行なうと共に，魅力的なフォントを選択する手法も提案する．

2.2 画像特徴を反映したフォントに関する研究

画像特徴を反映したフォントに関する研究として，Chen ら [2] は人の感情情報を反映したフォントを容易に生成することを目的とし，表情に対応するフォントを正確に分析し，その結果を元に，感情を表現したフォントの自動生成手法を提案している．この手法では，感情情報をフォントに反映するため，Generative Adversarial Network（GAN）[5] にフォント生成モジュールと感情情報を含むガイド信号を組み込んでいる．さらには，フォント生成モジュールに Earth Mover's Distance（EMD）[9] や勾配ペナルティを組み込み，複数のスタイルを組み合わせた新しいフォントを品質よく生成している．さらに，斉藤ら [16] はコミックにおいて，登場人物のセリフやナレーションなどで用いられている文字を適切なデザインで生成することを目的とし，既存のフォントをコミックの場面に対する印象に合うように組み合わせて，新たなフォントを生成する手法

表 1: 料理カテゴリと料理種類の一覧.

料理カテゴリ	料理種類
パスタ	和風パスタ，バジルパスタ，トマトパスタ，ミートソースパスタ，クリームパスタ，オイルパスタ，イカスミパスタ，スープパスタ
オムライス	和風オムライス，ハヤシオムライス，トマトソースオムライス，ケチャップオムライス，デミグラスオムライス，クリームオムライス，カレーオムライス，チーズオムライス
サラダ	和風サラダ，ビーンズサラダ，にんじんサラダ，トマトサラダ，シーザーサラダ，オーロラサラダ，マヨサラダ，ポテトサラダ
スープ	味噌汁，中華スープ，クリームスープ，コンソメスープ，コーンスープ，チゲスープ，トマトスープ
丼もの	海鮮丼，中華丼，そぼろ丼，天丼，カツ丼，牛丼，豚丼，鶏丼，ビビンバ丼，ローストビーフ丼

を提案している．これらのように，画像における特定の特徴をフォントに反映することに注目した研究は様々に行なわれている．しかし，料理画像に着目してフォントを選択する研究はこれまで行なわれていない．

2.3 サムネイル画像生成に関する研究

サムネイル画像の生成に関する研究として，Shen ら [10] は動画像から実用的なサムネイル画像を自動生成する手法を提案している．具体的には，動画像中の各フレームの意味情報を取得するため，外部画像データに対して教師付きで事前学習された最新の Convolutional Neural Network（CNN）によって生成された各フレームの埋め込みを活用し，代表的かつ高品質なフレームを選択し，魅力的なサムネイル画像を生成している．また，Apostolidis ら [1] らは，教師なしで動画像のサムネイルを選択する手法を提案している．学習においては GAN と強化学習の組み合わせ，映像コンテンツの代表性と美的品質という 2 つの基準に基づいてサムネイルを選択している．そして，Open Video Project（OVP）[3] と Youtube[4] のデータセットを用いた実験により，提案手法の有効性を示している．このように，動画像のサムネイル画像生成に関する研究は多く行なわれているが，SNS の料理レシピ投稿のサムネイル画像生成に関する研究は，これまで行なわれていない．

3 タイトル重畳料理画像と魅力度の関係分析

まず，フォントの選択がサムネイル画像の魅力に与える影響を調査するため，タイトル重畳料理画像と魅力度の関係分析を行なう．具体的には，料理画像に対して様々なフォントでタイトルを重畳した画像を人手で作成し，評価者が魅力度を付与する選好実験を行なった．

3：The Open Video Project, “The Open Video Project-Development Site, ” https://open-video.org/ ［2025/1/30 参照］.

4：Google, “Youtube, ” https://www.youtube.com/ ［2025/1/30 参照］.

(a) Noto Sans Japan (b) Noto Serif Japan (c) Dela Gothic One

(d) Hachi Maru Pop (e) Stick (f) Reggae One

(g) Kaisei Decol

図 2: 使用した Google Fonts[5] のフォント.

3.1 タイトル重畳料理画像の作成

フォントがタイトル重畳料理画像の魅力度に与える影響の要因や傾向を洞察するために，表 1 に示す種類の料理に対する料理画像を用意した．タイトル重畳料理画像を作成するための料理画像は，Instagram で検索して収集した．多様な料理を対象とする観点から，本実験では,「パスタ」,「オムライス」,「サラダ」,「スープ」,「丼もの」の 5 種類の料理カテゴリを設定した．さらに，各料理カテゴリから代表的な種類の料理を選択し，カテゴリ内で料理の種類が偏らないように表 1 の通りに選定した．画像を収集する際には，以下の 4 点に注意し，料理カテゴリごとに 103 枚，計 515 枚を人手で選択した．

1. タイトルが重畳されておらず，構造や明るさが適切である料理画像を選択した．
2. 料理画像とフォントの関係性のみをより精緻に分析するために，料理の位置や大きさが一定で同じ位置にタイトルが載るよう条件を可能な限り統一した．特に，皿の色や画像内の皿の位置，撮影角度を可能な限り統一した．
3. Instagram のサムネイルの仕様を参考に，画像の縦横比率が 1:1 で投稿されている料理画像を選択した．
4. 料理のタイトルから料理の特徴がわかりやすく，タイトルがひらがな，カタカナ，漢字のいずれかからなり，記号や数字などが含まれないものを選択した．

タイトルを描画するフォントは，図 2 に示す 7 種類を選定した．これらのフォントは，Google Fonts[5]から以下の「太さ」,「ひげ」,「丸み」の 3 つの特徴を基準として，これらの特徴が強いものと弱いものをそれぞれ偏りがないように選定した．

- 太さ：本研究では文字の線の太さを「太さ」と定義する．太いフォントは，目立つ特徴があり，強調された印象を与え，細いフォントは，洗練された外観を持ち，軽やかな印象に影響を与えると考えられる．
- ひげ：本研究では文字の端や曲がり角に現れる細かいデザインを「ひげ」と定義する．視覚的な装飾として機能し，読みやすさを向上させたり，文字を目立たせる役割も果たすと考えられる．

5：Google, “Google Fonts, ” https://fonts.google.com/ ［2025/1/30 参照］.

(a) Noto Sans Japan (b) Kaisei Decol (c) Dela Gothic One

図 3: タイトル重畳料理画像データセット中の画像例.

- 丸み：本研究では文字の曲線の丸みや角の丸みを「丸み」と定義する．丸みがあるフォントは柔らかく，親しみやすい印象を与え，角が鋭いフォントは，文字の角が鋭く，直線的な特徴が目立ち，力強い印象を与えると考えられる．

それぞれの画像に 7 種類のフォントを適用したタイトルを重畳配置し，計 3,605 枚のタイトル重畳料理画像を作成した．このとき，Instagram の料理レシピ投稿を参考にし，図 3 に示すように，文字色は黒，背景色は白とし，料理画像を完全に隠さないよう透過させて画像上部の中央に配置した．

3.2 タイトル重畳料理画像への魅力度付与

本研究では，タイトル重畳料理画像データセットに対して，料理画像に対する各フォントの魅力度を定量化する方法として，順序尺度によるスコアを魅力度として採用した．順序尺度とは，評価対象に対する相対的な順位や順序をつける尺度で，対象の間で絶対値的な差を測ることなく，順番やランクを評価する尺度である．本研究では，評価者が魅力度に基づいて料理画像を分類し，その結果に基づいて順序尺度を算出した．

具体的には，次式のように，各料理画像 i に対するフォント f の魅力度 $\overline{S}_{i,f}$ を N 人の評価者が料理画像 i に対してフォント f に付与した魅力度 $S_{i,f,j}$ を平均することで算出した：

$$\overline{S}_{i,f} = \frac{1}{N}\sum_{j=1}^{N} S_{i,f,j} \qquad (1)$$

本実験では，各評価者が与える魅力度 S を $\{+5, +1, 0\}$（したがって，$0 \leqq \overline{S} \leqq 5$）とし，以下のように割り当てた．

- 魅力的であると判断された画像：$S = +5$
- 魅力的でないと判断された画像：$S = 0$
- どちらでもないと判断された画像：$S = +1$

この基準により 3,605 件のタイトル重畳料理画像に対して，30 人の評価者により魅力度を評価した．各評価者は，図 4 に示すインタフェースを用い，7 種類のフォントを適用した料理画像について「魅力的なタイトル重畳料理画像」と「魅力的でないタイトル重畳料理画像」をそれぞれ 2 枚選択する．なお，各料理画像は 3 人の評価者により評価した（$N = 3$）．1. で述べた通り，魅力的であるとは，タイトル重畳料理画像において，タイトルフォントの印象が料理画像及び料理タイトルの印象と合致していることであることを評価者に明示した．最後に，各評価者の魅力度をもとに，料理画像に対するフォントの魅力度を式 1 に基づき算出した．

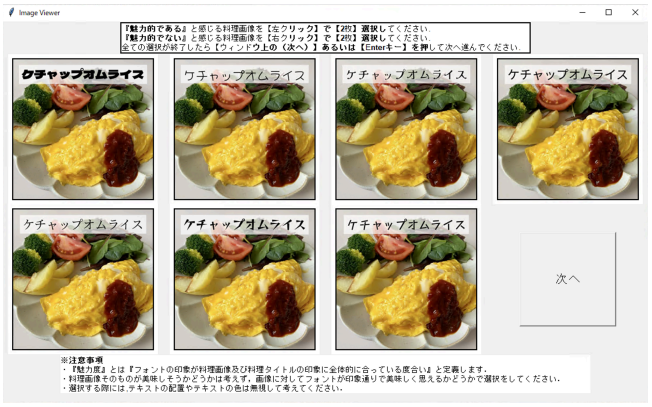


図 4: 選好実験のインタフェース.

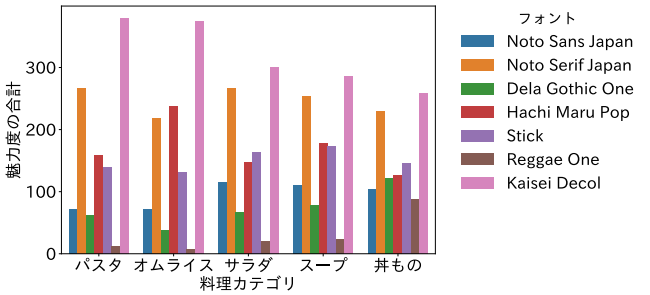


図 5: 各料理において魅力的なフォントを分析した結果.

3.3 タイトル重畳料理画像と魅力度の関係

まず，各料理カテゴリに対するフォントの魅力の傾向を調査した．図 5 に料理カテゴリごとにフォントの魅力度をフォントごとに合計した結果を示す．また，特に目立った特徴や興味深い傾向が見られた料理画像の具体例を図 6 に示す．

図 5 に示すように，「パスタ」や「オムライス」では，Kaisei Decol（図 2(c)）が高い魅力度を示した．Kaisei Decol は一般的にエレガント，優美，おしゃれなどの印象を抱くように設計されている [6]．そのため，「パスタ」や「オムライス」などの洋食に対するおしゃれな印象 [13] とフォントの印象が合致したと考えられる．その中でも「オムライス」では，Hachi Maru Pop（図 2(d)）が他の料理カテゴリよりも高い魅力度を示した．これは，「オムライス」の丸い形状がフォントの丸い印象と合致したためと考えられる．さらに，「丼もの」では，Reggae One（図 2(f)）が他の料理カテゴリよりも高い魅力度を示した．これ

6：いいフォント，“解星デコール（Kaisei Decol），” https://goodfreefonts.com/595/ ［2025/1/30 参照］.

(a) Noto Serif Japan.

(b) Hachi Maru Pop.

(c) Reggae One.

(d) Kaisei Decol.

図 6: 選好実験結果：各フォントが魅力的として選ばれたタイトル重畳料理画像の例.

は,「丼もの」には他の料理カテゴリよりも辛い，がっつり，スタミナといった印象を持つ料理が多かったため，料理に対する印象とフォントの刺々しい印象が合致したためと考えられる．

また，料理カテゴリ内の各料理種類において魅力的なフォントを分析すると，図 6 に示すように特徴的な結果が見られた．まず「スープ」では,「チゲスープ」に対して，Reggae One の魅力度が高かった．これは，辛い「チゲスープ」と，フォントの刺々しい印象が合致したためと考えられる．一方で,「味噌汁」や「中華スープ」など，和風や中華風の印象を持つ料理画像に対して，明朝体のフォントである Noto Serif Japan（図 2(b)）が魅力的であると判断された．また,「パスタ」では,「スープ」の「味噌汁」や「中華スープ」と同様，和風の印象を持つ料理画像に対して，Noto Serif Japan が魅力的であると判断された．更に，Hachi Maru Pop が魅力的であると判断された料理画像に注目すると,「エビ」,「ホタテ」など丸い形状をした具材が確認できるものが多く見られた．以上のことから，特定の印象を持つ料理に合うフォントが存在することや，フォントの形状と料理画像の全容や具材の形状との関連が示唆された．

選好実験の結果から，フォントの選択には，料理画像や料理タイトルが持つ印象が大きく影響することを確認した．具体的には，料理の味や料理画像内の視覚的特徴がフォント選択に反映された．このような結果は，評価者の選好が料理の特性とフォントとの関連性を的確に反映しており，本研究で構築する学習用データセットが本実験結果は妥当であることも示している．

4　フォント選択モデル

料理画像の画像特徴と料理タイトルのテキスト特徴を同時に考慮するマルチモーダルモデルを提案し評価する．

4.1　提 案 手 法

提案モデルの概要を図 7 に示す．画像特徴抽出には CNN，テキスト特徴抽出には Bidirectional Encoder Representations from Transformers（BERT）[7] を使用して，これら異なるモダリティ

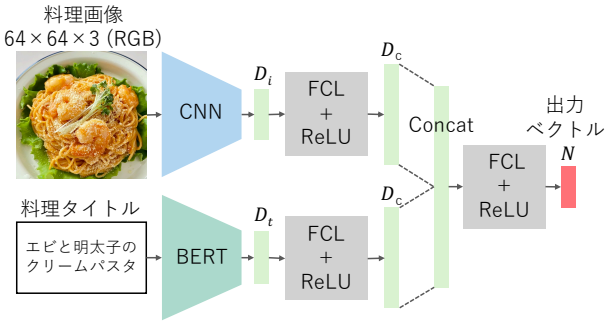


図 7: 提案モデルの概要.

の情報を結合して魅力度を算出する．料理画像には食材の質感や色など，魅力やカテゴリに関わる様々な視覚的特徴が含まれており，これらを抽出する能力が重要である．画像分類モデルとして事前学習された CNN は，学習により視覚的な特徴を高精度で捉えられるよう設計されているため，料理画像の特徴量抽出に適していると考えられる．また，料理タイトルは料理の特徴や印象を端的に伝える重要なテキスト情報であり，意味的特徴が料理全体の印象に影響を与える．BERT は，テキストの意味構造を深く理解するモデルであるため，料理タイトルの特徴量抽出に適していると考えられる．

画像特徴抽出においては，事前学習済みの CNN モデルを使用し，画像特徴を D_i 次元の特徴ベクトルとして抽出する．その後，この特徴ベクトルを，Rectified Linear Unit（ReLU）活性化関数を持つ全結合層（Fully-Connected Layer；FCL）に入力し，D_c 次元のベクトルに変換することで，テキスト特徴との結合に適した形式に調整する．

テキスト特徴抽出においては，入力テキストから得られる [CLS] トークンの出力を D_t 次元の特徴ベクトルとして抽出する．その後，ReLU 活性化関数を持つ全結合層に入力して D_c 次元のベクトルに変換することで，BERT が出力する汎用的な特徴から，画像特徴との結合を効果的に行なえるように，最適化された特徴に変換する．

次に，画像特徴とテキスト特徴を結合して $2D_c$ 次元の統合特徴ベクトルを構成する．この統合特徴ベクトルは，料理画像と料理タイトルの両モダリティの情報を効果的に反映し，これらに基づいた適切なフォント選択を可能にする．そして，このベクトルを ReLU 活性化関数を持つ全結合層に入力することで統合特徴ベクトルから N 次元の出力ベクトルを得て，カテゴリ分類を行なう．ここで，最後の全結合層は回帰ヘッドとして設計されており，各フォントに対する適合度スコアを予測する．

本モデルの損失関数においては，平均二乗誤差（Mean Squared Error；MSE）を使用し，モデルの予測値と目標値の誤差を最小化するように訓練を行なう．

4.2 評価実験

4.2.1 データセットと評価指標

タイトル重畳料理画像と魅力度の関係分析の結果を踏まえ，本研究における学習用データセットは，料理画像とその料理タイトルの組に対して，7 種類のフォントに対する魅力度をラベルとした 515 件のデータで構成されている．この魅力度は，3.2 節で述べた方法で算出した魅力度を使用した．また，データの分割には層化抽出法（Stratified Sampling）を採用し，学習データとテストデータを 7:3 の比率で分割した．この方法によりクラスごとのデータ数のバランスを保ち，クラスの偏りが学習結果に与える影響を軽減した．画像データの拡張には，無作為の水平反転に加え，縦横比を 1:1 に固定した状態で画像面積の 70%〜100% の範囲で無作為に切り出し，0 度〜60 度の範囲での無作為な回転を組み合わせた．さらに，ノイズの強度を平均値 0，標準偏差 0.05 に設定した Gaussian ノイズを導入すると共に，RGB 各チャネルに対してホワイトバランスのシフトを適用し，± 0.05 の範囲内で無作為に輝度変化を加えた．

本研究ではモデルの性能をより多面的に評価するため，3 つの評価指標を用いた．1 つ目に，MSE，2 つ目に，平均絶対誤差（Mean Absolute Error；MAE）を使用した．MSE は誤差を二乗するため，大きな誤差に対して敏感であり，外れ値に強く影響される．一方，MAE は誤差の平均的な大きさを重視し，外れ値による影響を受けにくい．これらを併用することで，誤差分布や外れ値の影響をより詳細に分析できる．3 つ目に，予測結果の中で最も魅力度が高いと評価されたフォントが，最も魅力的なラベル（Ground Truth；GT）と一致する割合を測定する魅力度の正解率（Accuracy）を指標とした．Accuracy により，モデルが最も魅力的と判断したフォントと実際のラベルの一致度合いを確認でき，モデルの精度を直接的に評価できる．本研究では，Top-1 accuracy および Top-3 accuracy を使用した．

4.2.2 フォント選択モデルの学習条件

本研究では，画像特徴抽出には，事前学習済み CNN モデルとして ResNet18 [6] または VGG16 [11] を用い，テキスト特徴抽出には，事前学習済み日本語 BERT モデル BERT base Japanese（IPA dictionary, whole word masking enabled）[7] [4] を使用した．

ResNet18 の実装では，事前学習済みモデル [8]を使用し，画像特徴を $D_i = 512$ 次元の特徴ベクトルとして抽出した．そして，全結合層を適用して，$D_c = 4,096$ 次元に変換した．同様に，VGG16 についても事前学習済みのモデル [9]を使用し，画像特徴を $D_i = 4,096$ 次元の特徴ベクトルとして抽出し，全結合層を通じて $D_c = 4,096$ 次元に調整した．テキスト特徴抽出においては，[CLS] トークンの出力を $D_t = 768$ 次元とし，全結合層を適用して $D_c = 4,096$ 次元に変換した．これらの画像特徴とテキスト特徴を結合し，$2D_c = 8,192$ 次元の統合特徴ベクトルを構成した．そして，全結合層を通じて $N = 7$ 次元の出力ベクトルを得た．ここでは，ResNet18 と VGG16 を用いたマルチモーダルモデルをそれぞれ Image-ResNet18-Text-BERT, Image-VGG16-Text-BERT と呼ぶ．

また，提案手法をシングルモーダルのモデルと比較するため，画像特徴のみ及びテキスト特徴のみで学習させたモデルも作成した．まず，画像特徴のみのモデルでは，ResNet18 から画像特徴を 512 次元の特徴ベクトルとして抽出し，全結合層を適用して，4,096 次元に変換した．同様に，VGG16 を使用したモデルでは，画像特徴を 4,096 次元の特徴ベクトルとして抽出し，全結合層を通じて 4,096 次元に変換した．最後に，全結合層を通じて $N = 7$ 次元の出力ベクトルを得た．また，テキスト特徴のみのモデルにおいては，BERT からテキスト特徴を 768 次元の特徴ベクトルを抽出し，全結合層を適用して 4,096 次元に変換し，全結合層を通じて $N = 7$ 次元の出力ベクトルを得た．ここで，ResNet18 のみを用いたモデル，VGG16 のみを用いたモデル，BERT のみを用いたモデルをそれぞれ，Image-ResNet18, Image-VGG16，Text-BERT と呼ぶ．

本実験の学習過程では，モデルのパラメータ最適化に Adaptive Moment Estimation（Adam）[8] を使用した．ここでは，学習率スケジューラとして ReduceLROnPlateau を導入することで，必要に応じて学習率を動的に調整し，検証データにおける損失（Validation Loss）の改善が停滞した場合，学習率を減少させるように設計した．具体的には，本研究では初期学習率 $\alpha = 0.001$ から開始し，検証損失の改善が 5 エポック以上続かなかった場合，学習率が 10 分の 1 に減少する設定を適用した．また，バッチサイズは 32 に設定した．

4.2.3 定量的評価

提案モデルを用いたフォント選択の評価結果を表 2 に示す．評価指標として，ResNet18 および VGG16 を使用したときの MSE，MAE，Top-1 accuracy，Top-3 accuracy を示す．画像とテキストの特徴を統合したマルチモーダルモデルである，Img_{Res}-Txt_{BERT} と Img_{VGG}-Txt_{BERT} は，全ての評価指標でシングルモーダルモデルを上回った．特に，Img_{VGG}-Txt_{BERT} モデルは，Top-1 accuracy および Top-3 accuracy において最も高いテスト性能（Top-1 accuracy = 67.33%，Top-3 accuracy = 92.08%）を示し，安定した結果を示した．一方で，Img_{Res}-Txt_{BERT} モデ

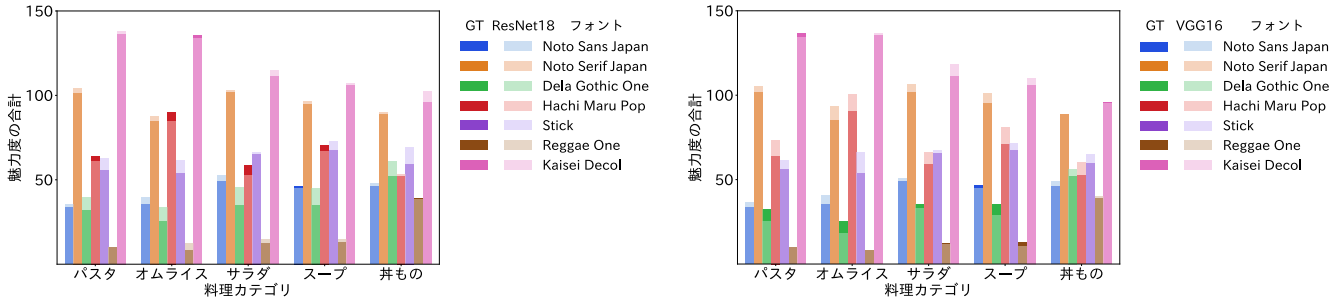
7：東北大学自然言語処理研究グループ, “BERT base Japanese（IPA dictionary, whole word masking enabled）,” https://huggingface.co/tohoku-nlp/bert-base-japanese-whole-word-masking［2025/1/30 参照］.

8：PyTorch, “resnet18,” https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html［2025/1/30 参照］.

9：PyTorch, “vgg16,” https://pytorch.org/vision/main/models/generated/torchvision.models.vgg16.html［2025/1/30 参照］.

表 2: モデルの評価結果.

モデル	モダリティ	MSE ↓		MAE ↓		Top-1 Acc [%] ↑		Top-3 Acc [%] ↑	
		Train	Test	Train	Test	Train	Test	Train	Test
Img_{Res}	画像	1.012	1.330	0.793	0.884	68.32	63.34	90.83	87.13
Img_{VGG}	画像	1.306	1.478	0.879	0.927	68.61	62.34	89.97	87.13
Txt_{BERT}	テキスト	1.026	1.202	0.811	0.865	70.60	64.34	90.54	88.12
Img_{Res}-Txt_{BERT}	画像＋テキスト	**0.743**	1.284	**0.667**	0.855	71.48	64.36	**93.98**	87.13
Img_{VGG}-Txt_{BERT}	画像＋テキスト	0.837	**1.201**	0.704	**0.848**	**72.19**	**67.33**	93.41	**92.08**



(a) ResNet18 を用いたモデル.

(b) VGG16 を用いたモデル.

図 8: モデルごとのフォント魅力度の予測値分布.

ルは，訓練データにおける MSE および MAE が最も低く，訓練データに対して優れたフィッティング性能を示した．しかし，テストデータにおける MSE および MAE は Img_{VGG}-Txt_{BERT} より高く，過学習の可能性が示唆された．

シングルモーダルモデルでは，Img_{Res} および Img_{VGG} が，画像から捉えられる部分的な情報に依存するため，精度に限界が生じたと考えられる．特に，画像のみを用いた場合の MSE と MAE が全体的に高い値を示しており，画像情報だけではフォント選好を十分に捉えることが困難であることがわかった．一方で，Txt_{BERT} は，比較的優れた性能を示した．これは，料理タイトルに含まれるテキスト情報がフォント選好において重要な手がかりとなることを示唆している．しかし，テキスト情報のみでは情報が限定され，画像情報を含むマルチモーダルモデルと比較して精度は低かった．

これらの結果から，提案モデルが料理画像と料理タイトルに基づくフォント選好の傾向を的確に捉え，適切な予測を行なっていることが確認された．特に，画像特徴とテキスト特徴を統合することで情報不足を補い，より高精度で安定した結果を得られることが示された．従って，マルチモーダルモデルを用いた提案手法の有効性が定量的に示された．

さらに，提案モデルの評価において，選好実験で得られた魅力度の分布（GT の分布）と Img_{Res}-Txt_{BERT} 及び Img_{VGG}-Txt_{BERT} による魅力度の予測値の分布をそれぞれ比較した．GT と Img_{Res}-Txt_{BERT} の分布結果と，GT の魅力度と予測値の一致度を示す重なり率を計算した結果を図 8(a) に，GT と Img_{VGG}-Txt_{BERT} の分布結果と，重なり率を計算した結果を図 8(b) にそれぞれ示す．これらの結果から，両者は非常に類似しており，モデルが各料理に対するフォント選好の傾向を的確に捉えていることが確認された．このことから，提案モデルが，フォント選好の傾向を的確に捉え，学習用データセットにおける料理画像や料理タイトルとフォント印象の関連性を十分に学習できたことがわかる．

4.2.4 定性的評価による実験結果の分析

提案モデルの性能を直感的に評価するため，学習結果を視覚的に確認し，定性的評価を行なった．料理画像と料理タイトルに対して，提案モデルが魅力的であると予測したフォントでタイトルを重畳した料理画像を図 9 に示す．その結果，3.3 節で示したような,「和風パスタ」や「和風オムライス」,「和風サラダ」,「味噌汁」などの和風という印象を持つ料理画像に対しては明朝体の Noto Serif Japan（図 2(b)）が魅力的であることや，辛い，がっつり，スタミナなどの印象を持つ料理画像に対しては，Reggae One（図 2(f)）が魅力的であるという特徴，さらには,「きのこ」など丸い形状をした具材が確認できる料理画像に対しては，Hachi Maru Pop（図 2(d)）が魅力的であることなどを的確に予測することができた．これらの結果は，料理画像や料理タイトルの持つ印象的特徴とフォントのデザイン的特徴との間に強い関連性が存在しており，提案モデルがその関係性を適切に学習し，予測する能力を持つことを示している．

一方,「白菜のクリームスープ」や「醤油バターコーンのポタージュ」といった料理画像において，GT は Kaisei Decol（図 2(c)）であるのに対し，提案モデルは Noto Serif Japan を魅力的であると推定する結果が多く見られた．これは，選好実験における評価では「クリームスープ」や「コーンスープ」という西洋料理の要素が強調され，おしゃれや優美といった印象を持つ Kaisei Decol が魅力的と評価される傾向があったが，提案モデルでは，料理名に含まれる「白菜」や「醤油」といった和食に関連する

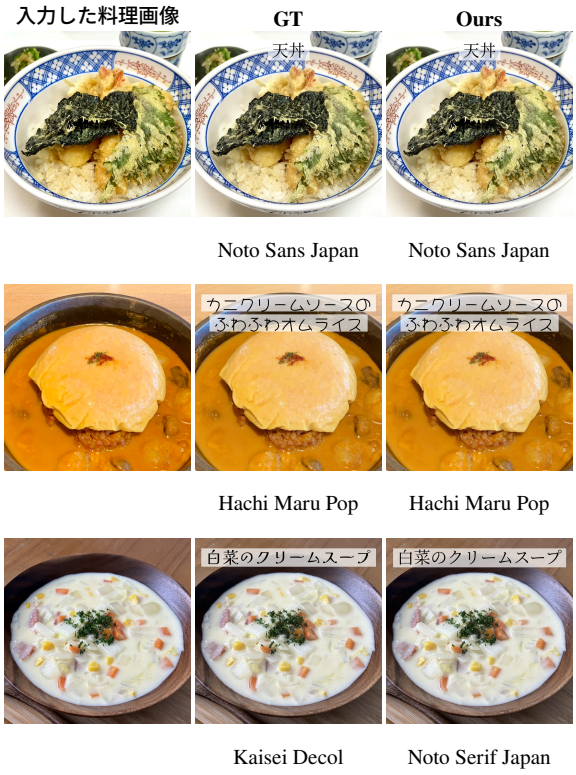


図 9: 魅力的なフォント推定結果の例.

単語を強い特徴として捉えたためと考えられる．これは，人的評価が料理画像全体の見た目や印象に基づいて行なわれるのに対し，提案モデルが料理タイトルに含まれるテキスト情報に偏った判断をしている可能性を示唆している．そのため，提案モデルが料理画像と料理名の両方をバランスよく活用し，統合的な判断を下せるよう，画像とテキスト間の相互作用を深めるアーキテクチャを導入する必要がある．

5 むすび

本研究では，膨大な数の料理レシピ投稿が存在する Instagram に注目し，料理レシピ投稿のサムネイル画像が人の目を惹くための重要な要素であると考え，料理タイトルを画像上に配置する際に，魅力的なタイトル重畳料理画像を生成することを目的とした．この目的のため，料理画像において料理タイトルテキストを魅力的なフォントで重畳することを目指し，フォント選択モデルを提案した．

モデル構築に当たり，フォントの選択がサムネイル画像の魅力に与える影響を調査するため，多様なフォントでタイトルを装飾したタイトル重畳料理画像を構築して選好実験を行ない，各タイトル重畳料理画像の魅力度を評価した．その結果を踏まえ，料理画像と料理タイトルに対してフォントの魅力度合いを含むデータセットを構築した．フォント選択モデルには，画像特徴抽出に ResNet18 [6] または VGG16 [11] を，テキスト特徴抽出には BERT [7] を利用したモデルを提案した．このモデルを上記データセットで学習することで，料理画像と料理タイトルにおいて各フォントの魅力度を予測した．提案モデルの評価実験を行なった結果，67%の正解率で 7 種類のフォントから最も魅力的なフォントを選択できることを示した．また，ResNet18 に比べて VGG16 は汎化能力がやや高いことが分かった．今後の課題としては，フォント選択モデルの学習に必要なデータ量の増加と，過学習を防ぐための正則化技術の適用，また，画像とテキスト間の相互作用を深めるアーキテクチャの導入が挙げられる．

謝辞 本研究の一部は JSPS 科研費 JP20K12038，JP22H00548 の支援による．選好実験に協力していただいた中京大学の学生の皆様に感謝する．

文　献

[1] E. Apostolidis, E. Adamantidou, V. Mezaris, and I. Patras. Combining adversarial and reinforcement learning for video thumbnail selection. *Commun. ACM*, Vol. 27, No. 9, pp. 1–9, Sep. 2021.
[2] L. Chen, F. Lee, H. Chen, W. Yao, J. Cai, and Q. Chen. Automatic Chinese font generation system reflecting emotions based on generative adversarial network. *Appl. Sci.*, Vol. 10, No. 17, pp. 5976_1–17, Aug. 2020.
[3] X. Chen and W. Tang. Study on the influence of font type on the aesthetics evaluation of food packaging. In *Proc. 15th Int. Conf. Human System Interact.*, 5 pages, July 2022.
[4] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Trans. Audio Speech Lang. Process.*, Vol. 29, No. 11, p. 3504–3514, Nov. 2021.
[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.*, Vol. 27, pp. 2672–2680, Dec. 2014.
[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, Dec. 2016.
[7] D. Jacob, C. Ming-Wei, L. Kenton, and T. Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conf. N. Am. Chapt. Assoc. Comput. Linguist.*, pp. 4171–4186, May. 2019.
[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *Comput. Res. Reposit. arXiv Preprint*, No. arXiv:1412.6980, Dec. 2015.
[9] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.*, Vol. 40, No. 2, pp. 99–121, Nov. 2000.
[10] B. Shen, N. Pancha, A. Zhai, and C. Rosenberg. Practical automatic thumbnail generation for short videos. *Sympo. Electron. Imaging*, Vol. 33, No. 8, pp. 281_1–283_7, Jan. 2021.
[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. 3rd Int. Conf. Learn. Represent.*, 14 pages, May 2015.
[12] （株）マルハニチロホールディングス. 料理レシピに関する調査 2020. `https://www.maruha-nichiro.co.jp/corporate/news_center/news_topics/20200818_research_recipe2020_1.pdf`, Aug. 2020.
[13] マイスボイスコム（株）. 食のジャンルのアンケート調査（第 5 回）. `https://www.myvoice.co.jp/biz/surveys/27802/index.html`, Sep. 2021.
[14] （株）NTT ドコモモバイル社会研究所（編）. データで読み解くモバイル利用トレンド 2024–2025 モバイル社会白書. エヌ・ティ・ティ出版, 東京, Nov. 2024.
[15] 笠井ゆきひ, 佐藤弘喜. フォントの持つシズル感のデザイン化手法の検討. 日本デザイン学第 64 回春季研究発表大, No. B5-06, June 2017.
[16] 斉藤絢基, 中村聡史. コミック創作のためのフォント融合による文字デザイン手法. 第 32 回人工知能学全大, No. 3Z2-05, Jan. 2019.