7T-01

大規模マルチモーダルモデルを用いた広告画像レイアウトの評価と説明

砂田 達巳† 塩原 楓† 劉 岳松‡ 丹治 直人‡ 勢〆 弘幸‡ 山崎 俊彦†
東京大学† Septeni Japan株式会社‡

1 はじめに

近年の Web 広告の重要性と AI 技術の発展に伴って, AI によって広告を評価することで広告作成を支援する研究が広く行われている. Web 広告画像評価の研究では, クリック率を予測することで評価を行う場合が多い. Xia ら [1] は Web 広告の画像, テキスト, メタデータ (ジャンル, 媒体など) をもとにクリック率を予測した. また, Park ら [2] は Web 広告の画像, メタデータからクリック率を予測し, 画像中の重要度の可視化を行った. これらの先行研究ではどのように評価しているかが不明瞭という問題点がある. 実際に広告作成の場面でこれらの予測モデルを使う場合, クリック率を予測してもどこがどのように評価されているかの説明がなければ修正箇所がデザイナーに伝わらない. そこで, 本研究では図 1 に示すように自然言語によって広告の評価を行い, 評価基準と評価箇所を明確に示すことを目指す.

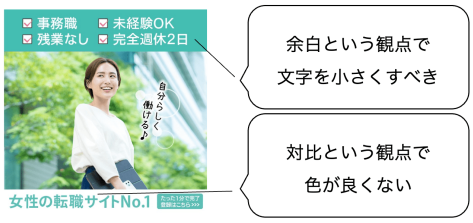

図 1 本研究の目標

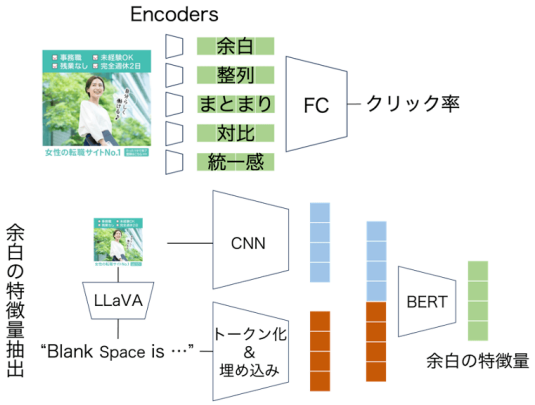

図 2 クリック率の予測モデルアーキテクチャ

2 提案手法

本章では, 説明可能な広告評価のための評価基準を含むクリック率予測モデルと, 摂動による評価モデルの説明手法の 2 つを提案する. まず, 本提案モデルでは広告画像を評価する基準を予め設定する. 高橋ら [3] はレイアウトの 5 つの法則として, (1) 余白を十分に取る, (2) 揃えて配置する, (3) 関連するものをグループ化する, (4) 強弱をつける, (5) 同じパターンを繰り返す, を挙げている. 本研究ではこの 5 つを評価基準としたモデルを提案する. まず, 大規模マルチモーダルモデル LLaVA [4] に広告画像に対して各基準の評価を行うようプロンプトを与える. LLaVA が出力したテキストを BERT [5] の tokenizer, embedding 層を用いてベクトル空間に埋め込み, テキスト特徴量とする. また, 広告画像から CNN を用いて画像特徴量を抽出する. これらの特徴量を結合し, 訓練済み BERT モデルに入力することで, 各評価基準に合わせた特徴量を得る. こうして得られた 5 つの特徴量ベクトルを結合し, 全結合層に入力することでクリック率を予測する. このように, 評価基準ごとに特徴量を分けることで, それぞれの評価基準の特徴量の全結合層における係数からクリック率に対する各評価基準の寄与が推測できると考えられる.

次に, 提案した評価モデルが広告画像の注目箇所の説明手法を提案する. 広告画像は背景, ロゴ, テキストなどの各レイヤーで構成されている. 本提案手法ではこれらのレイヤーに対して画像編集を行い, レイアウトを変更する. レイアウト変更で生成された様々なパターンの画像をそれぞれ評価モデルに入力し, 予測されるクリック率の分布を大規模言語モデルに分析させることで各レイヤーをどうすれば改善できるかがわかる.

Assessment and Explanation of Banner Image Layout with Large Multimodal Model
†Tatsumi Sunada, Kaede Shiohara, Toshihiko Yamasaki, The University of Tokyo
‡Yuesong Liu, Naoto Tanji, Hiroyuki Seshime, Septeni Japan, Inc.

表 1　クリック率予測の精度の比較

	相関係数 (↑)	MSE (↓)
ResNet [6]	0.548	0.365
ViT [7]	0.509	0.345
Ours	0.506	0.201

表 2　画像編集の条件

項目	条件
色	白, 黒, 赤, 橙, 黄, 緑, 青, 藍, 紫, 変化なし
サイズ	0.8, 1.0, 1.2 (倍)
x 座標	-20, -10, 0, +10, +20 (px)
y 座標	-20, -10, 0, +10, +20 (px)

3　実験

データセットは Septeni Japan株式会社で作成し, Yahoo! JAPAN に出稿したバナー画像広告を用いた．このデータセットは学習データが 21546 枚, テストデータが 5987 枚で構成される．提案したモデルからクリック率を予測し, 他モデルと比較する．比較対象としては, ResNet [6] および Vision Transformer (ViT) [7] を用いた．予測の損失関数は最小二乗誤差 (MSE), モデルの精度評価はクリック率の予測値と正解値の相関係数と MSE である．表 1 に結果を示す．提案モデルは ViT や ResNet と相関係数はほぼ変わらず, MSE は他モデルより優れていることがわかる．

また, 提案したモデルの注目箇所を説明するには, 入力画像のレイアウトを変更した様々な例をモデルに入力し, その出力との関係を大規模言語モデルに分析させる．レイアウトを変更するための画像データは Adobe Photoshop の PSD ファイル形式のものを用いる．PSD ファイルでは広告画像の各レイヤーの情報が保存されているので, レイヤーごとの編集が可能である．画像編集は表 2 に示すような条件で行う．色の変換については Palette-based Photo Recoloring [8] を用い, k 平均法でクラスタリングした色のうち, 最も多い色を指定の色に変換した．図 3 は画像編集の例である．本章では一例として図 3 に示した「AI × ヒトの共創で, 高効率のバナーを配信!」と書かれたテキストのレイヤーに対して実験を行う．画像の編集条件, 予測されるクリック率, 各評価基準のネットワークにおける重みを ChatGPT にプロンプトとして与える．ここで, どの編集条件の時にクリック率が改善され, それはどの評価基準による寄与が

図 3　画像編集の一例

1. **色**:
   - 赤とオレンジ色のケースでは、スコアが比較的高くなる傾向があります。これは、特に対比や統一感の要素で顕著です。
   - 青や黒のケースでは、スコアが低めですが、これは統一感の要素が影響している可能性があります。
2. **サイズ**:
   - サイズが1.2のケースでは、スコアが高くなる傾向が見られます。これは、まとまりや統一感によるものである可能性が高いです。
3. **x座標とy座標**:
   - x座標とy座標が0付近の場合（中心に近い場合）、スコアが高い傾向が見られます。これは、整列やまとまりの要素が影響している可能性があります。

図 4　編集条件と出力の関係を ChatGPT が分析した結果

大きいのかを ChatGPT に分析させる．例えば図 4 の結果では, 統一感とまとまりの観点から, サイズを大きくするとクリック率が改善される傾向があるとわかる．

4　まとめ

本研究では, 広告画像のレイアウト評価にあたって評価基準を明確にしたモデルと評価箇所ごとのモデルの反応を分析する手法の 2 つを提案し, 評価箇所と評価基準が明らかな広告画像の評価を目指した．今後は本評価モデルとその分析手法を実際の広告作成に使用するユーザー検証を行い, 広告を改善する支援として有効かどうかを調査することを考えている．

参考文献

[1] B. Xia, X. Wang, T. Yamasaki, K. Aizawa, and H. Seshime. Deep neural network-based click-through rate prediction using multimodal features of online banners. In *BigMM*, 2019.
[2] K.-W. Park, J.-W. Ha, J. Lee, S. Kwon, K.-M. Kim, and B.-T. Zhang. M2fn: Multi-step modality fusion for advertisement image assessment. *Applied Soft Computing*, 2021.
[3] 高橋佑磨, 片山なつ. 伝わるデザインの基本 増補改訂版 良い資料を作るためのレイアウトのルール. 技術評論社, 2016.
[4] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *NeurIPS*, 2023.
[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.
[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.
[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
[8] H. Chang, O. Fried, Y. Liu, S. DiVerdi, and A. Finkelstein. Palette-based photo recoloring. *ACM ToG*, 2015.