テキストタグを併用した 画像の品質評価モデルの提案

杉浦 敦之[1,a)]　山内 悠嗣[1,b)]

概要

デジタル画像の急増に伴い，画像の品質を自動的に評価する技術の需要が高まっている．画像の品質を回帰により推定する従来手法では画像の内容を考慮しないため，特定のテーマや意図を持った画像の評価ができない課題があった．そこで，本研究では画像とテキストタグを用いた品質評価手法を提案する．画像とテキストタグとの関係性を評価に取り入れ，視覚情報だけでなく意味的な要素も考慮した品質評価を実現する．

1. はじめに

スマートフォンや画像生成 AI の普及により，流通・生成される画像コンテンツの量は爆発的に増大している．これに伴い，大量の画像の中から高品質なものを自動で選別・評価する技術の需要が高まっている．画像の品質評価は，視覚的な美しさや魅力を定量的に評価するタスクであり，写真の整理，画像検索，コンテンツ推薦など幅広い分野で応用が期待される．

従来研究として，人間が付与した画像の品質スコアを回帰問題として直接予測する手法が提案されている [1]．このアプローチは，主に画像の視覚的特徴量のみを利用していた．しかし，この方法では人間が知覚するような複雑な美的要素や，画像の内容までを十分に捉えることが困難であるという課題があった．

そこで本研究では，画像の内容を記述したテキスト情報を活用することで，より人間の感性に近い品質評価を実現する新たなモデルを提案する．提案手法は，言語と画像の埋め込み空間を学習する Contrastive Language-Image Pre-training (CLIP) [2] を基盤として採用し，視覚的特徴とテキスト情報の双方を統合的に扱う．本モデルでは画像とテキスト間の意味的類似度の 2 つの指標に基づいたランキング学習を導入する．この学習により，単なる視覚的な美しさだけでなく，画像の内容との整合性も評価基準に組み込むことが可能となる．

本研究の貢献は以下に 2 点に集約される．

- 画像と画像の内容を記述したテキスト情報を明示的に活用するモデルを美的品質評価に導入することにより，画像のテーマや意図といった文脈を評価に取り込むことを可能にし，従来手法より人間の感性に近い評価精度を実現する．
- 人間の美的評価と意味的整合性の 2 つの基準に基づいた順位付けを行うランキング学習を導入する．このアプローチにより，特定のテーマや意図を持つ画像群の中から，より人間の感性に合致した画像の評価・選別を実現する．

2. 関連研究

2.1 画像の品質評価に関する先行研究

画像の品質評価は，構図，色彩，コントラストといった客観的要因と文化的背景，個人的な好みといった主観的な要因に影響される．近年，美的評価技術は，定量的で再現性のある評価が可能であるため注目を集めている．

従来手法 [1] では，CNN を使用し，画像から抽出された特徴を基に美的評価を表すスコアに変換する．モデルの出力を美的スコアとし，画像につけられた正解美的スコアと比較をすることによりモデルが学習される．しかし，視覚的特徴のみを用いて学習されるため，画像の内容は考慮されることなく，結果として画像にそぐわない，人間の評価とかけ離れた予測が行われる課題を持つ．

このような課題に対し，Celona et al.[4] は，CNN ベースの特徴抽出に加え，画像の構図やスタイルといった属性情報を活用する手法を提案した．同研究では，属性条件付きハイパーネットワークを導入し，事前に学習した属性情報を入力として受け取り，それに基づいて美的スコアを予測する評価ネットワークのパラメータを動的に生成する．この手法により，画像ごとに属性依存的に最適化された美的評価モデルが生成され，属性間の複雑な関係を考慮した柔軟な評価が可能となった．しかし，美的要因はこれ以上に多様であるため，未対応の要素が残されているほか，属性ごとにパラメータを生成する構造であるため計算コスト

[1] 中部大学
[a)] tr25009-0170@sti.chubu.ac.jp
[b)] yuu@fsc.chubu.ac.jp

が高いのも課題である．

このような画像内容の課題を補う試みとして，美的批評キャプショニングも活発に研究されている．Vera Nieto et al.[5] が提案したデータセットでは，美的スコアに加えて，ユーザーによる批評コメントも収集しており，モデルが「なぜその画像が良いのか」を言語的に説明できる可能性を拓いている．画像に対する批評コメントの内容を直接的に美的評価に反映可能となり，より解釈性の高い評価が可能となっている．一方，感情極性が美的評価を適切に表現できない場合があることや，言語表現や評価基準が文化的背景や個人差に依存する点などの 課題も残されている．

画像の美的評価においては，視覚的特徴に依存するのではなく，他のモダリティを統合したマルチモーダルな評価モデルの構築が今後の重要な方向性となる．

2.2 画像とテキストのマルチモーダル学習

深層学習の分野では，画像とテキストといった複数情報を同時に扱うマルチモーダル学習が大きな注目を集めている．このアプローチの目的は，単一の情報だけでは捉えきれない，より複雑な情報をモデルに理解させることにある．特に画像とテキストの分野では，画像とその説明文との間の意味的な関連性を学習させ，両者を共通のコンテキスト空間上に表現することで，テキストによる画像検索や画像の内容要約など，多岐にわたる応用が可能となる．

画像とテキストのマルチモーダル学習を体現する代表的なモデルとして，OpenAI によって提案された Contrastive Language-Image Pre-training (CLIP) [2] が挙げられる．CLIP は，Vision Transformer (ViT) [3] などの画像エンコーダと Transformer ベースのテキストエンコーダを用いて，画像とテキストをそれぞれ共通のベクトル空間に埋め込む．その上で，インターネットから収集した膨大な画像とテキストのペアに対し，意味的に対応するペアの特徴量ベクトルが空間上で近接するように対照学習を行う．この強力な事前学習により，CLIP は特定のタスクのために再学習の必要性がなく，多様な下流タスクに対して高い汎化性能を発揮する．

CLIP の登場以降も，この分野では多様なモデルが提案されている．Flamingo[6] は，画像とテキストが混在したシーケンスを入力とし，視覚的な情報に基づいた質問応答や対話を得意とする Visual Language Model (VLM) を提案した．また、Salesforce Research が提案した BLIP (Bootstrapping Language-Image Pre-training)[7] は，キャプション生成とノイズ除去を繰り返す独自の手法で学習データの品質を高め，画像キャプション生成や VQA(Visual Question Answering) といった理解と生成の両タスクで高い性能を達成した．

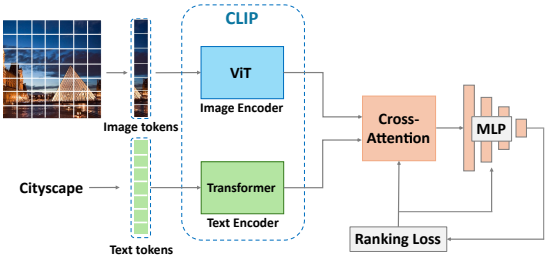

図 1 提案手法の流れ．

3. 提案手法

本稿では，画像およびその内容を記述するテキストタグを用いた画像評価手法を提案する．提案手法は，画像の評価においてテキストタグが有する情報を活用する点に特徴を持つ．さらに，画像と対応するタグの整合性に基づいたランキング学習を導入することにより，各テキストタグに関して事前に定義された指標に基づく画像品質の優劣を学習することが可能である．

提案手法の流れを図 1 に示す．本手法ではベースモデルに Constructive Language-Image Pre-training (CLIP)[2] モデルを採用する．CLIP ではテキストエンコーダに Transformer，画像エンコーダに Vision Transformer(ViT) ベースのアーキテクチャが採用されている．画像とテキストタグのペアを CLIP に入力し，それぞれの画像とテキストに対して各エンコーダから 512 次元の埋め込みベクトルを抽出する．次に，抽出された両埋め込みベクトルを Cross-Attention に入力し，両方の特徴量を統合した 512 次元の埋め込みベクトルを生成する．出力された埋め込みベクトルを Multi-Layer Perceptron(MLP) に入力することで入力画像の品質スコアを出力する．

3.1 画像とテキストタグのマルチモーダル学習

本研究では，画像の品質，画像とテキストの意味的一致性の 2 つを同時に評価するための基盤として，CLIP を導入する．CLIP は画像とテキストをマルチモーダルに学習するモデルである．CLIP によって得られた画像とテキストの埋め込みベクトルは，視覚的特徴と言語的特徴の整合性を保ちつつ，美的スコア推定タスクにおける入力として有効に機能する．CLIP の事前学習では，画像とそのテキストタグのペアを入力し，画像とテキストの特徴量を同じ埋め込み空間にマッピングする．そして，画像とテキストの正しいペア間で類似度を最大化し，誤ったペアの類似度を最小化するよう学習する．

3.2 ランキング学習における画像の順位付けの定義

従来の回帰ベースのみの手法では，画像間の相対的な優劣を直接的に学習することが難しく，評価結果の一貫性や整合性に課題があった．そこで本手法では，同一のテキス

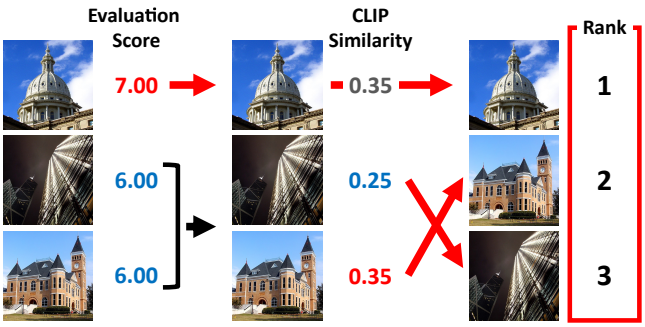

図 2 画像の順位付けの例 (プロンプト：”Architecture”).

トタグに属する画像群に対して，画像の品質スコアと CLIP 類似度の 2 指標を用いたランキング情報を付与し，その順位関係に基づいて学習 [9] を行う．このアプローチにより，画像の品質が高く，テキストと整合性の高い画像をモデルが高く評価できるようになる．

画像に対して順位付けするために，まず各画像に対し，人間による主観評価の平均である品質スコアを付与する．スコアが類似する場合は，画像とテキストの意味的関連度を示す CLIP 類似度に基づいて順位を決定する．これにより，品質の高い画像ほど上位に配置され，画像品質が同程度の場合は，テキストとより一致する画像が上位となる．このような定義を導入することで，品質が高く，かつ意味的に適合した画像が正当に評価されるようになり，評価の信頼性と一貫性を向上させることが可能となる．

提案手法における損失関数 $L(\theta)$ は式 (1) に示すように，ランキング情報が上位のテキスト T と画像 x のペア (T, x_i) と下位のテキストと画像のペア (T, x_j) に対して MLP から出力される予測スコア f_θ の差，及び回帰損失から計算される．

$$L(\theta) = -\mathbb{E}_{(T,x_i,x_j)\sim D}\left[\log\left(\sigma\left(f_\theta(T,x_i) - f_\theta(T,x_j)\right)\right)\right] + \lambda \cdot \frac{1}{D}\sum_{i=1}^{D}\left((f_\theta(T,x_i) - s_i)^2\right) \quad (1)$$

ここで，σ はシグモイド関数，D はテキストと画像のペア数，s は画像 x に対する正解の品質スコア，λ はバランスを調整する重みである．第 1 項はランキングに基づく損失，第 2 項は回帰に基づく損失を表す．提案手法では，同じテキストタグの画像同士で比較し，ランキング情報に基づいて画像 x_i が画像 x_j より品質の予測スコアが高くなるように最適化する．

4. 評価実験

提案手法の有効性を確認するために，2 つの評価実験を行う．1 つ目は，人間が目視で品質を評価したデータセットを用いて品質の予測結果を比較する．2 つ目は，各手法において生成画像に対する品質評価の精度を比較する．

4.1 画像の品質評価の比較実験

従来手法 [1] と提案手法，提案手法にテキストタグを入力しない手法の美的評価精度を比較する．従来手法は Efficient Net をベースモデルとして用い，画像特徴量から画像の品質スコアを回帰推定する手法である．モデルにテキストタグを入力しない方法は，画像を CLIP の画像エンコーダに入力して埋め込みベクトルを抽出し，得られた埋め込みベクトルを MLP に入力することで画像の品質を出力する手法である．テキストタグを入力しない手法と比較することで，画像を表すテキスト情報を品質評価に用いることによる有効性を評価する．

実験には Aesthetic Visual Analysis(AVA) Dataset[8] を使用する．AVA Dataset は，画像に対し複数の人間による画像の品質評価の平均スコアと画像内容を表すテキストタグが付与されている．学習では 135,000 枚の画像とテキストタグのペア，評価用画像としては 1,350 枚とテキストタグのペアを使用する．AVA Dataset には 66 種類のテキストタグが画像ごとに最大 2 個付与されている．本実験では 1 つのタグにつき 1,000 枚以上の画像となるものを対象として使用する．AVA Dataset の平均美的スコアを正解スコアとし，正解スコアと各手法により予測したスコアで評価する．

評価には，スコア誤差と誤差分散，Linear Correlation Coefficient（LCC），Spearman’s Rank Correlation Coefficient（SRCC），KL Divergence，Earth Mother’s Distance を用いる．LCC は，予測スコアと正解スコアとの間で計算された相関係数である．SRCC は，予測スコアとランキング情報とで計算された相関係数である．KL Divergence は，予測スコアの確率分布が，正解スコアの確率分布とどれだけ異なるかを非対称的に測る指標である．EMD は，予測スコア分布と正解スコア分布の間の「形の違い」を評価する指標である．

表 1 に品質推定の比較結果を示す．表 1 より，提案手法は画像の品質の誤差が従来手法と同程度，その他の全ての指標において上回る結果を示した．提案手法では，品質の低い画像から高い画像まで一貫性をもった予測が可能であることを確認できる．また，表 1 の KL Divergence と EMD により求めた正解スコアと予測スコアの分布の類似度に着目すると，どちらの指標においても提案手法が上回る結果を示した．データセットの全体的な傾向に影響されることなく，より人間の評価に近い予測ができていることがわかる．

図 3 に評価用画像の品質の正解スコア分布図と予測スコア分布図を重ね合わせたものを示す．図 3 より，従来手法では，予測スコアが最頻値付近に集中していたが，提案手法では真の分布に近しい予測を実現していることが分かる．

図 4 に，画像のタグ及び画像の品質スコアの例を示す．画像の視覚的特徴のみを考慮する従来手法では人間の評価

表 1　各手法における美的評価精度結果.

Method	Error ↓	LCC ↑ (mean)	SRCC ↑ (mean)	KL Divergence ↓	EMD ↓
Previous	0.478(±0.36)	0.596	0.556	0.295	0.106
Ours(w/o Text)	0.442(±0.136)	0.650	0.581	0.185	0.116
Ours	0.482(±0.147)	0.684	0.620	0.138	0.094

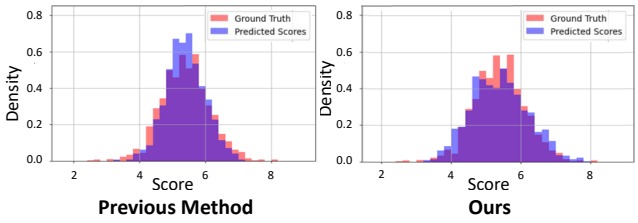

図 3　各手法における品質予測スコアの分布.

TextTag	Birds
Previous	5.36
Ours	6.62
Human Evaluation	6.84

図 4　画像に対する真のスコアと，予測スコアの例.

との間にずれが生じている．一方，視覚的特徴に加え，テキスト情報を考慮する提案手法では人間の評価により近い予測ができている．

4.2　生成画像に対する美的品質評価精度

Stable Diffusion では，生成する際のパラメータを変更することで，意図的に違いも持った画像データを作成することができる．提案手法の有効性，汎化性を検証するためにプロンプトとの整合性を調整した生成画像に対して評価する．実験には，画像生成モデルとして realistic vision[*1]を用い，異なる Classifier-Free Guidance（CFG）スケールに基づいて生成した画像データセットを使用する．CFG スケールはプロンプトに対する画像生成の忠実度を調整するパラメータである．CFG スケールは数値が高いほど，よりプロンプトに忠実な画像を生成できる．本評価実験では，画像サイズ 512 × 512[px]，Sampling Steps は 30 に設定した．CFG スケールは 1，3，5，7，9 の 5 種類を用意し，各 CFG スケールにつき 100 枚の画像を生成した．各モデルを用いて，各生成画像に対する品質スコアを推定し，CFG スケールによる画像品質の傾向を分析，比較する．

図 5 に生成画像と各 CFG スケールにおける予測スコアの分布を示す．図 5 より，従来手法ではパラメータが異なる画像においても類似したスコアを出力するのに対し，提案手法では CFG スケールの値と強い相関関係を持つことが分かる．従来手法では画像の視覚的特徴のみで評価しており，類似した画像に対して細かな品質の判断が難しくなっている．一方，提案手法では，視覚的特徴の他にテキストタグとの整合性を評価に反映させているため，より多角的な評価をしていることを示している．

[*1] https://civitai.com/models/4201/realistic-vision-v60-b1

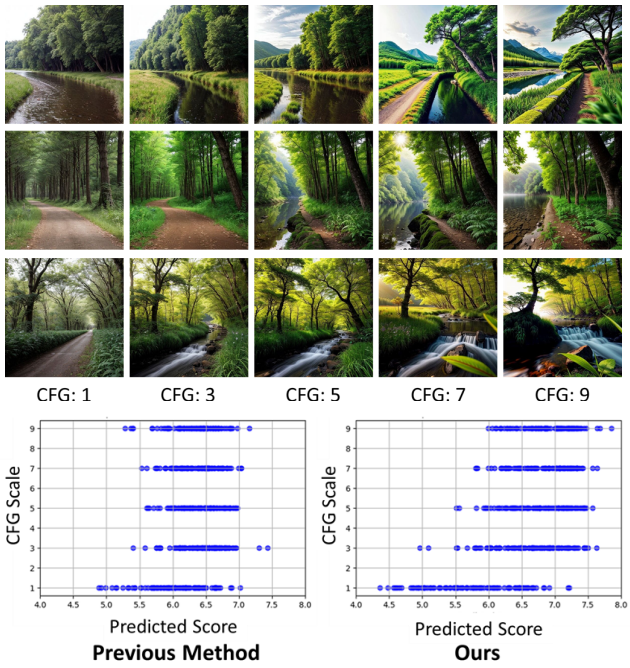

図 5　生成画像と生成画像に対する各手法の予測スコア分布 (プロンプト：”Nature”).

5.　おわりに

本研究では，テキストタグを用いた画像の品質評価モデルを提案した．提案手法を用いることで，画像の単調な美しさだけでなく，画像内容との関連性を評価に反映させることができ，人間により近い品質評価を行うことができる．今後は，品質評価の説明性の出力を実現する．

参考文献

[1] Talebi, em et al. ”NIMA: Neural image assessment.” IEEE transactions on image processing, 2018.
[2] Radford, em et al. ”Learning transferable visual models from natural language supervision.” ICML, 2021.
[3] Dosovitskiy, Alexey, et al. ”An image is worth 16x16 words: Transformers for image recognition at scale.” arXiv preprint arXiv:2010.11929 (2020).
[4] Celona, Luigi, et al. ”Composition and style attributes guided image aesthetic assessment.” IEEE Transactions on Image Processing 31 (2022): 5009-5024.
[5] Vera Nieto, Daniel, Luigi Celona, and Clara Fernandez Labrador. ”Understanding aesthetics with language: A photo critique dataset for aesthetic assessment.” Advances in Neural Information Processing Systems 35 (2022): 34148-34161.
[6] Alayrac, Jean-Baptiste, et al. ”Flamingo: a visual language model for few-shot learning.” Advances in neural information processing systems 35 (2022).
[7] Li, J., et al. ”BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation.” (2022).

[8] Murray, em et al. and Florent Perronnin. "AVA: A large-scale database for aesthetic visual analysis." CVPR, 2012.
[9] Burges, Chris, et al. "Learning to rank using gradient descent." Proceedings of the 22nd international conference on Machine learning. 2005.