

# 芸術作品に興味を促すキャプションの自動生成

永野里佳奈<sup>†</sup> 山本 祐輔<sup>†</sup>

<sup>†</sup> 静岡大学大学院総合科学技術研究科

〒 432-8011 静岡県浜松市中区城北 3-5-1

E-mail: <sup>†</sup>nagano@design.inf.shizuoka.ac.jp, <sup>††</sup>yamamoto@inf.shizuoka.ac.jp

**あらまし** 本稿では、芸術作品に対して、日常生活で起こりうる場面を想起させる注釈文を自動生成し、鑑賞者に提示する手法を提案する。提案手法では、まず日本語画像キャプションデータセット STAIR を用いて、画像内容を示す文章生成モデルを構築する。その後、Twitter ハッシュタグ「#名画で学ぶ」が付けられた画像ツイートを用いて、構築した文章生成モデルのファインチューニングを行う。提案手法によって生成された文章は、高度な背景知識を持ち合わせていなくても理解できるよう、日常生活と関わりのある内容となっている。芸術作品に対して日常生活と関連がある見方を示すことで、芸術を身近な存在として感じられるようにすることが期待される。

**キーワード** 美術鑑賞支援、鑑賞の個人化、画像キャプショニング

## 1 はじめに

芸術は、作品を介して作者の思考を表現したり、見る人の感情を引き出すものである。作品という一つのイメージから自由に発想を広げることや新しい価値観に触れることで、創造性や感性、豊かな人間性を育むことができる。このように芸術は、人生や教育面などの様々な側面で、重要な役割を担っている [1]。近年では、芸術に触れる新しい機会が広がり始めている。情報技術の発達により、直接美術館を訪れる必要なく、オンライン上で芸術作品を鑑賞できるようになった。例えば、Google Arts & Culture<sup>1</sup>や各美術館が提供する、オンラインビューリングが登場している。

芸術に触れる機会が広がっている一方、芸術作品に触れる人は限られている。文化庁の調査によると、過去 1 年間で文化芸術を「まったく・ほとんど鑑賞していない」と答えた人は 46.1% であった [2]。その理由は、「関心がないから」(35.4%)、「特にない・分からぬ」(22.8%) の順で多かった [2]。文化芸術を鑑賞しない具体的な理由ではなく、そもそも関心がないことが要因となっている。芸術に触れる機会を作らない人にとって、芸術作品は身近な存在でないため、興味・関心を持ちにくいと考えられる。

本稿では、芸術作品を敬遠している人々に興味を持たせることを目的として、芸術作品に対して、日常生活で起こりうる場面を想起させる注釈（キャプション）を自動生成し、鑑賞者に提示する手法を提案する。提案手法では、まず日本語画像キャプションデータセットの STAIR Captions [3] を用いて、画像内容を示す文章生成モデルを構築する。その後、Twitter ハッシュタグ「#名画で学ぶ」が付けられた画像ツイートを用いて、構築した文章生成モデルのファインチューニングを行う。キャプション生成には、画像認識に CNN、文章生成に LSTM を利用する。

1 : <https://artsandculture.google.com/?hl=ja>



図 1 美術館で鑑賞する場合（左）と提案手法（右）の比較

図 1 に、芸術作品「落穂拾い」<sup>2</sup>に対する典型的な解説文と提案手法の出力テキスト例を記す。美術館では、

「『落穂拾い』は農村の貧しい人々の姿を描いただけでなく、『旧約聖書』の『ルツ記』に基づいた作品である。」(from Wikipedia)

といったような解説文が作品とともに展示される。この種の解説文は、作品が描かれた時代背景等の知識があれば楽しむことができる。一方で、美術に関心がない一般人は、この解説文を読んでも作品に対して興味を持つことは難しいと考えられる。提案手法は同作品に対して、

「もう 2 度とベビースター食わせない」<sup>3</sup>

というキャプションを自動生成する（図 1 右）。左図の解説文

2 : [https://ja.wikipedia.org/wiki/落穂拾い#/media/ファイル:Jean-François\\_Millet\\_-\\_Gleaners\\_-\\_Google\\_Art\\_Project\\_2.jpg](https://ja.wikipedia.org/wiki/落穂拾い#/media/ファイル:Jean-François_Millet_-_Gleaners_-_Google_Art_Project_2.jpg)

3 : #名画で学ぶ主婦業 [4] から引用

とは異なり、提案手法によって生成されたテキストは、高度な背景知識を持ち合わせていなくても理解できるよう、日常生活と関わりのある内容となっている。鑑賞者はこのテキストを見ることで、子どもが床に食べ散らかしたお菓子を片付ける主婦の姿と、落穂拾いをする様子が重ねることができる。

これまで、美術館での鑑賞体験を向上させるアプローチとして、訪問者側が行う芸術作品推薦[5][6]や、体験型の展示[7]などが提案されてきた。これらのアプローチは、ある程度美術に興味・関心をもっている人には有効である。しかし、そうでない人には、芸術作品への興味を喚起するような別の支援が必要となる。提案手法は芸術作品に対して日常生活と関連がある見方を示すことで、芸術を身近な存在として感じられるようになることが期待される。

## 2 関連研究

### 2.1 美術教育

一般的に多くの人が芸術に接する機会の一つに、学校での美術教育が存在する。Zimmermanは、美術教育による創造性育成の課題について述べた[8]。これまで創造性は、生まれつき持っているものであり、自然と涵養されるものと考えられていた。そのため、美術教育の現場では、教師は生徒の創作活動に直接介入せず、動機づけなど間接的なサポートが中心だった。しかし、このような教育アプローチでは、生徒が芸術の意味や価値を理解しようとする限り、創造性を育むことはできない。

金子は、日本の美術教育が、学習者の自己完結で行われている問題を指摘した[9]。日本では、創作活動や作品鑑賞による自己表現及び自己鑑賞が、教育そのものになると考えられてきた。そのため、美術教育は教育を受ける側の活動に委ねられ、教師や教育内容はほとんど意味を持たない問題が生じている。

以上から、現在の美術教育の問題点として、学習が個人に依存していることが挙げられる。そのため、他の教科と比べて、個人によって理解や意欲に差が生じやすいと考えられる。本研究では、このような美術教育の問題点を踏まえて、芸術に興味・関心を持たない人に焦点を当てた鑑賞支援を行う。

### 2.2 ICT技術を利用した美術鑑賞支援

近年、ICT技術の利点を生かした新しい鑑賞支援が行われている。例えば、作品や美術館での鑑賞体験を自分自身や近しい人と関連づけることで、新しい価値を示したものがある。

Spenceらは、美術館の作品を親しい人にギフトとして贈るアプリケーションを提案した[5]。アプリケーション利用者は、贈る相手のためにふさわしい作品を選ぶという新しい視点で作品に触れることができる。Foshらは、訪問者が同伴するパートナーに向けて、展示作品の鑑賞方法をデザインする手法を提案した[6]。評価実験の結果、同伴者は参加者がデザインした鑑賞方法を肯定的に受け入れ、通常の鑑賞とは異なる個人化された鑑賞を体験することが可能であることが明らかになった。Munteanらは無形の民族文化を理解し、体験する展示設計とシステムを提案した[10]。提案システムは、民族文化の展示物が

どのような使われ方をしたか、現代の物との関連性から訪問者に探索させることを企図している。これらの関連研究は、自分自身や自分と近しい人と作品の繋がりを考えさせることによって、作品に個人的な意味を持たせている。

また、訪問者を魅了しつつ、教育面の充実も考慮した展示内容も考案されている。Mallavarapuらは、訪問者の動きによって映像が変化する体験型展示に対して、視覚的なフィードバックを行う展示方法を提案した[7]。フィードバックによって、訪問者は展示作品が自然の生態系を表していることを理解し、変化の原因を考えるようになった。

このように、訪問者にとって美術館の体験がより魅力的になり、また学習効果が高くなるよう、ICTを活用した様々なアプローチが提案されている。しかし、これらの提案はある程度芸術に興味・関心のある人が対象となっている。一方、本研究では、美術館や博物館への来館意欲を高めるために、芸術に関心のない人に芸術に対する興味を持たせる方法を提案する。

### 2.3 画像キャプショニング

VinyalsらはCNNとRNNを組み合わせ、入力された画像に対して、その内容を表す文章を自動生成する手法を提案した[11]。この研究では、エンコーダのCNNから画像を認識し、デコーダのRNNによって機械翻訳を行っている。従来の手法と比較した結果、提案手法は画像内容を正確に表現し、文章の完成度としても精度が高いことが明らかになった。

一方で、画像内容を説明するだけでなく、様々な要素を取り入れることで、各目的に合わせたキャプション生成が考案されている。例えば、Yoshidaらは入力画像に対して、笑えるキャプションを自動生成する手法を提案した[12]。この手法は、大喜利投稿ウェブサイト「ボケて」に投稿されたお題画像とボケの文章、評価につけられた星の数を利用する。提案手法によるキャプションを評価したところ、単純な画像キャプションと比較して、より面白いと捉えられることが分かった。別の応用例では、Liuらが入力画像に対して、自由形式の詩を自動生成する手法を提案した[13]。手法では画像からシンボルを認識した後、データセットから関連する詩の特徴を抽出し、詩の生成を行った。生成された詩を客観的指標とチューリングテストで評価し、手法の有効性を示した。

本研究では、芸術作品に興味を促すため、芸術作品の画像から日常生活で起こりうる場面を想起させるキャプションを生成する。

## 3 提案手法

本章では、画像として表現された芸術作品から日常生活で起こりうる場面を想起させるキャプションを自動生成する手法について述べる。提案手法は、以下の手順で画像キャプショニングを行う。

(1) 芸術作品ではない一般的な画像とそのキャプションのペアから、画像キャプショニングモデルを構築する（当該モデルを「一般画像キャプショニングモデル」と呼ぶことにする）

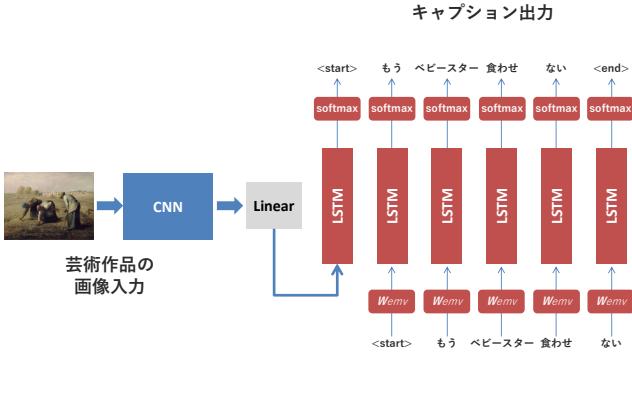


図 2 本稿で用いた画像キャプショニング用深層学習モデル

(2) 「#名画で学ぶ」画像つきツイートデータを用いて、(1)で作成した一般的な画像キャプショニングモデルをファインチューニングする

(3) 任意の芸術作品画像に対して(1)で構築したモデルを適用し、日常生活で起こりうる場面を想起させるキャプションを生成する。

以下、各手順の詳細な説明を述べる。

### 3.1 一般画像キャプショニングモデルの構築

手順1では、一般的な画像とそのキャプションのペアから、一般画像キャプショニングモデルを構築する。画像キャプショニングの生成モデルを構築するには、深層学習を用いた様々な手法が提案されているが、本稿では、画像認識に Convolutional Neural Networks (CNN)，文章作成に Long short-time memory (LSTM) を利用した、単純なネットワークを用いた画像キャプション生成モデルを構築する [11]。CNN と LSTM を組み合わせることで、入力された未知の画像から、出力として画像内容を示す新規の文章を生成することができる。図 2 は、構築したネットワークモデルを示している。

一般画像キャプショニングモデルの学習に、日本語画像キャプションデータセットの STAIR Captions [3] を利用した。STAIR Captions では、MS COCO [14] の各画像 164,062 枚に対して、内容を示す 5 つの日本語の文章がついている。合計で、820,310 件の画像とキャプションのペアがある。本稿では、STAIR Captions の全キャプションのうち、公開されている 616,435 件を利用した。

### 3.2 日常生活で起こりうる場面を想起させる画像キャプション生成モデル

手順2では、手順1で構築した一般画像キャプショニングモデルをファインチューニング [15] することで、日常生活で起こりうる場面を想起させる画像キャプション生成モデルを構築する。

ファインチューニングとは、大量のデータセットで学習済みの別モデルを、目的とするデータセットで再学習させる手法で



図 3 ハッシュタグ「#名画で学ぶ主婦業」がつけられたツイート例

ある。通常、重みを決定しモデルを作成するには、目的のデータセットを大量に用意し、学習させる必要がある。一方、ファインチューニングは、再学習する際、既に学習済みのモデルの特徴量抽出部分を引き継ぐ。そのため、目的のモデル作成に必要なデータセットが少量でも、一定の精度を保ったモデルが作成できる。

本稿では、ファインチューニングを行うための画像テキストペアのデータセットとして、Twitter ハッシュタグ「#名画で学ぶ」がつけられた画像つきツイートに着目した。ハッシュタグ「#名画で学ぶ」は、芸術作品に描写された様子や人物の表情から、日常生活の出来事に結び付けたツイートを投稿するものである。このハッシュタグはシリーズ化されており、「#名画で学ぶ主婦業」、「#名画で学ぶ大学院」など様々な種類が存在する。例えば、ハッシュタグ「#名画で学ぶ主婦業」であれば、ツイートは主婦が遭遇しやすい場面を取り上げている（図 3 に実際のツイート例<sup>4</sup>を示す）。これら画像テキストペアは、芸術作品に付与される一般的な解説文とは異なり、美術の知識が一切ない一般人でも画像とそのテキストの内容が理解できるよう、美術のコンテキストを排除した内容となっている。その内容は、日常生活の出来事と結びついており、多くのツイッターユーザから関心を集めていることもあり、日常生活で起こりうる場面を想起させる画像テキストペアとして有用であると考えられる。

このハッシュタグは、様々なカテゴリのシリーズがある中で、カテゴリ内の専門用語が多く含まれるものも存在する。本稿では、多くの人が想像しやすいキャプションになるよう、「主婦業」「大学院」「飲食店」などの 11 つのテーマに着目して、「#名画で学ぶ」シリーズの画像付きツイートを収集した。該当ハッシュタグがつけられた画像ツイートは、過去のツイートをまとめたウェブサイト、Togetter から収集した。「#名画で学ぶ主婦業」に関しては、Twiiter アカウント「#名画で学ぶ主婦業」

<sup>4</sup>: <https://twitter.com/mochin22/status/992299695539023873>

表 1 画像ツイート数の内訳

ハッシュタグ名	ツイート件数
#名画で学ぶ主婦業 <sup>6</sup>	178
#名画で学ぶ大学院 <sup>8</sup>	274
#名画で学ぶ飲食業 <sup>9</sup>	83
#名画で学ぶ小説家 <sup>10</sup>	51
#名画で学ぶリモートワーク <sup>11</sup>	20
#名画で学ぶフリーランス <sup>12</sup>	27
#名画で学ぶ病院 <sup>13</sup>	24
#名画で学ぶドラッグストア <sup>14</sup>	12
#名画で学ぶ休園休校中の育児 <sup>15</sup>	39
#名画で学ぶ本屋 <sup>16</sup>	65
#名画で学ぶ銀行 <sup>17</sup>	186



図 4 実際に生成された、画像内容を示すキャプション（左）と提案手法のキャプション（右）

【厳選】<sup>5</sup>からも収集した。この Twitter アカウントでは、リツイートやいいねが多くされた過去の人気ツイートが投稿されている。この中で、Togetter で収集したツイートと重複するものは除いている。以下、利用した各ハッシュタグのツイート数を示す（表 1）。今回は、モデル作成のために合計 959 件の画像付きツイートを利用した。

収集した画像付きツイートのデータセットを用いて、手順 1 で構築した一般画像キャプショニングモデルの文章生成を行う LSTM 部分をファインチューニングした。美術作品「落穂拾

5 : [https://twitter.com/meiga\\_gensen?lang=ja](https://twitter.com/meiga_gensen?lang=ja)  
 6 : <https://togetter.com/li/1224086>  
 7 : <https://togetter.com/li/1225191>  
 8 : <https://togetter.com/li/1223161>  
 9 : <https://togetter.com/li/1224479>  
 10 : <https://togetter.com/li/1224218>  
 11 : <https://togetter.com/li/1502188>  
 12 : <https://togetter.com/li/1261150>  
 13 : <https://togetter.com/li/1496099>  
 コロナに関するツイートを集めた記事を利用する。  
 14 : <https://togetter.com/li/1466214>  
 15 : <https://togetter.com/li/1506931>  
 16 : <https://togetter.com/li/1590255>  
 17 : <https://togetter.com/li/1222452>

い」<sup>2</sup> の画像に対して、提案手法によって実際に生成されたキャプションと一般画像キャプショニングモデルで生成されたキャプションの例を図 4 に示す。

本来「落穂拾い」は、農民が収穫後に残った麦の穂を届んで拾う様子を描写した作品である。一般画像キャプショニングモデルでは、「落穂拾い」に対して「草原にいる象の群れの中に 1 匹の子象がいる」という文章が生成される（図 4 左）。腕を下に伸ばして拾う農民たちの姿が、複数の象として認識されたと考えられる。一方で、提案手法は「乳しぼりをしているところに遭遇した」という文章が生成される（図 4 右）。落ちた麦の穂を拾う農民の様子を乳しぼりをしていると認識し、その様子を目撃した状態を表している。一般画像キャプショニングモデルの文章と比較して、提案手法による文章は、日常生活と関連しており、共感を得やすいと考えられる。このように、芸術作品画像に対して、提案手法で生成されたキャプションを提示することで、ユーザは芸術作品を身近な存在として感じられる。

## 4 実行例

結果として、一般画像キャプショニングモデルとファインチューニングを行った提案手法による文章では、大きな違いがみられなかった。また、入力する画像によっては、未知語を示す<unk>が入った未完成の文章が出力された。以下、提案キャプショニングモデルの適用例について記す。まず、入力した芸術作品の画像は次の 9 作品である。

- (1) 春一連作「四季」より<sup>18</sup>
- (2) 糸杉と星の見える道<sup>19</sup>
- (3) 草上の昼食<sup>20</sup>
- (4) 巨人<sup>21</sup>
- (5) ファンゴッホの寝室<sup>22</sup>
- (6) 舟遊びをする人々の昼食<sup>23</sup>
- (7) 積みわら<sup>24</sup>

18 : <https://upload.wikimedia.org/wikipedia/commons/thumb/e/ed/Renoir23.jpg/600px-Renoir23.jpg>

19 : [https://upload.wikimedia.org/wikipedia/commons/thumb/1/10/Van\\_Gogh\\_-\\_Country\\_road\\_in\\_Provence\\_by\\_night.jpg/600px-Van\\_Gogh\\_-\\_Country\\_road\\_in\\_Provence\\_by\\_night.jpg](https://upload.wikimedia.org/wikipedia/commons/thumb/1/10/Van_Gogh_-_Country_road_in_Provence_by_night.jpg/600px-Van_Gogh_-_Country_road_in_Provence_by_night.jpg)

20 : [https://upload.wikimedia.org/wikipedia/commons/thumb/7/74/Monet\\_dejeunersurlherbe.jpg/600px-Monet\\_dejeunersurlherbe.jpg](https://upload.wikimedia.org/wikipedia/commons/thumb/7/74/Monet_dejeunersurlherbe.jpg/600px-Monet_dejeunersurlherbe.jpg)

21 : [https://upload.wikimedia.org/wikipedia/commons/thumb/b/be/El\\_coloso.jpg/640px-El\\_coloso.jpg](https://upload.wikimedia.org/wikipedia/commons/thumb/b/be/El_coloso.jpg/640px-El_coloso.jpg)

22 : [https://upload.wikimedia.org/wikipedia/commons/thumb/c/c8/Vincent\\_Willem\\_van\\_Gogh\\_137.jpg/440px-Vincent\\_Willem\\_van\\_Gogh\\_137.jpg](https://upload.wikimedia.org/wikipedia/commons/thumb/c/c8/Vincent_Willem_van_Gogh_137.jpg/440px-Vincent_Willem_van_Gogh_137.jpg)

23 : [https://upload.wikimedia.org/wikipedia/commons/thumb/8/8d/Pierre-Auguste\\_Renoir\\_-\\_Luncheon\\_of\\_the\\_Boating\\_Party\\_-\\_Google\\_Art\\_Project.jpg/700px-Pierre-Auguste\\_Renoir\\_-\\_Luncheon\\_of\\_the\\_Boating\\_Party\\_-\\_Google\\_Art\\_Project.jpg](https://upload.wikimedia.org/wikipedia/commons/thumb/8/8d/Pierre-Auguste_Renoir_-_Luncheon_of_the_Boating_Party_-_Google_Art_Project.jpg/700px-Pierre-Auguste_Renoir_-_Luncheon_of_the_Boating_Party_-_Google_Art_Project.jpg)

24 : [https://upload.wikimedia.org/wikipedia/commons/thumb/b/b8/Wheatstacks\\_%28End\\_of\\_Summer%29%2C\\_1890-91\\_%28190\\_Kb%29%3B\\_Oil\\_on\\_canvas%2C\\_60\\_x\\_100\\_cm\\_%2823\\_5-8\\_x\\_39\\_3-8\\_in%29](https://upload.wikimedia.org/wikipedia/commons/thumb/b/b8/Wheatstacks_%28End_of_Summer%29%2C_1890-91_%28190_Kb%29%3B_Oil_on_canvas%2C_60_x_100_cm_%2823_5-8_x_39_3-8_in%29%2C_The_Art_Institute_of_Chicago.jpg/600px-Wheatstacks_%28End_of_Summer%29%2C_1890-91_%28190_Kb%29%3B_Oil_on_canvas%2C_60_x_100_cm_%2823_5-8_x_39_3-8_in%29)

(8) 第九の波<sup>25</sup>

(9) 泣く女<sup>26</sup>

次に、各芸術作品に対する提案キャプションと、比較のための一般画像キャプションを示す（図5）。一般画像キャプションは、3.1節で構築した一般画像キャプショニングモデルから生成されたものを利用した。

## 5 考 察

### 5.1 提案手法によるキャプショニングモデル

本研究の提案手法は、入力された芸術作品画像に対して、日常生活に関連した見方を示すキャプションを生成することを目的としている。提案手法の実行例（図5）では、日常生活に関連した見方を示すことができなかった。一方、いくつかの実行例では、作品の象徴的な部分に着目しやすくなるキャプションが生成できたと考えられる。

例えば「糸杉と星の見える道」（図5画像2）では、提案手法は「クジャクが羽ばたいている」というキャプションを生成した。題名にある糸杉は、作品が描かれたヨーロッパと違い、日本ではありません身近な植物ではない。そのため、題名と作品画像のみよりも「クジャクが羽ばたいている」というキャプションがある方が、作品中央に描かれた植物に注目しやすい。また、作品の象徴である糸杉をクジャクの羽ととらえることで、ゴッホ特有の筆使いを表現できていると考えられる。このように提案キャプションによっては、題名だけでは想像しづらい、作品の新しい見方を提示できていると思われる。

### 5.2 一般画像キャプショニングモデル

画像内容を示す一般画像キャプショニングモデルに利用された画像データは、一般的な写真である。しかし、実物を写した写真と異なり、芸術作品は描写の仕方が幅広く存在している。例えば、写実主義のように対象物を忠実に再現した作品から、ピカソのキュビズム作品「泣く女」（図5画像9）のような抽象画や、ムンクの「叫び」などがある。そのため、一般画像キャプショニングモデルに抽象画などを入力すると、正確に内容を表した文章でないことが分かる。例えば「泣く女」を入力すると、「スケートボードに乗っている人がいる」という文章が出力される。

このように、画像キャプション生成自体がうまく行われていない場合が、芸術作品を入力画像とした際、生じやすいと考えられる。画像キャプションモデル作成では、芸術作品のような実際の写真ではない画像に対応する必要がある。一方で、人の認識とは異なるキャプションは、その差異から鑑賞者のインパクトに残りやすい場合がある。また、5.1節に記したように、作品の新しい着眼点を示す可能性も期待される。

%2C\_The\_Art\_Institute\_of\_Chicago.jpg

25 : [https://upload.wikimedia.org/wikipedia/commons/thumb/4/4a/Hovhannes\\_Aivazovsky\\_-\\_The\\_Ninth\\_Wave\\_-\\_Google\\_Art\\_Project.jpg/](https://upload.wikimedia.org/wikipedia/commons/thumb/4/4a/Hovhannes_Aivazovsky_-_The_Ninth_Wave_-_Google_Art_Project.jpg/)

26 : [https://upload.wikimedia.org/wikipedia/en/1/14/Picasso-The\\_Weeping\\_Woman\\_Tate\\_identifier\\_T05010\\_10.jpg](https://upload.wikimedia.org/wikipedia/en/1/14/Picasso-The_Weeping_Woman_Tate_identifier_T05010_10.jpg)

### 5.3 データ量の追加

今回本研究では、「#名画で学ぶ」画像付きツイートを合計959件しか利用することができなかった。そのため、日常生活に関連した文章を生成するためのデータ数が不足していた。結果として、入力する画像によっては、未知語を表すが文章中に現れてしまうことがあった。例えば、芸術作品「泣く女」の画像（図5画像9）を生成したモデルに入力すると、「競馬の接戦、<unk>、<unk>、<unk>、」という文章が出力される。

このような不足を補う方法として、データの水増しと別の大喜利データ利用の2点が考えられる。データの水増しでは、元データの一部に変更を加えることで、新しいデータを作成する。例えば、キャプションの再翻訳や類語への変換、文中の単語の入れ替えなどが挙げられる。一方データの水増しだけでは、目的となるキャプション生成に対してデータ数が少ない。そのため、「#名画で学ぶ」シリーズと類似した別の大喜利データの利用も必要である。具体的には、画像大喜利サイト「ボケて」<sup>27</sup>やYahoo!知恵袋<sup>28</sup>の大喜利カテゴリなどが考えられる。

### 5.4 利用したキャプションとツイートの性質の違い

ファインチューニングがうまくいかなかつた原因に、事前学習に用いた STAIR Captions と、「#名画で学ぶ」シリーズの画像付きツイートの性質が違うことが考えられる。STAIR Captions では、キャプションのデータセットを作成する際、クラウドワーカーに最低文字数や文調、單文に限定するなどの条件を課した[3]。一方、画像付きツイートでは鉤括弧で括られた会話形式の文章「『ママー！』『はいはい、何？』『ママ、見てー！』『見てる、見てる。』」<sup>29</sup>や、キヤッチコピーのような短文「深夜2時」<sup>30</sup>などの文章形式が多く見られた。ツイートの性質上、画像を説明するというよりは、見た人にインパクトを与える文章にする傾向がみられる。

このように、一般画像キャプショニングモデルと提案キャプションモデルで用いたデータは、同じ文章でも性質や使われる単語が大きく異なる。よって、ファインチューニングで生成された文章が、目的とするキャプションにならなかつたと考えられる。

### 5.5 評価実験

今回、提案手法によるキャプションが実際にどの程度興味を促すかどうか示すことができなかつた。そのため、今後は評価実験を行う必要がある。本研究では、次の2項目を評価するよう依頼する。

- (1) 実験協力者が、元々芸術にどの程度興味・関心があるか
- (2) 各キャプションから、対応する芸術作品にどの程度興味を持ったか

また、提案手法を評価するための比較キャプションとして、以下の3種類を用いる予定である。

27 : <https://bokete.jp/>

28 : <https://chiebukuro.yahoo.co.jp/>

29 : [https://twitter.com/mossan\\_open/status/992224748758122497](https://twitter.com/mossan_open/status/992224748758122497)

30 : [https://twitter.com/meiga\\_gensen/status/994587218294419456](https://twitter.com/meiga_gensen/status/994587218294419456)

入力した 芸術作品画像	提案手法による キャプション	一般画像 キャプション
	ツリーのキャラクターの キャラクターの マスコットが飾られている	花柄の布の上に、 赤い花が置かれている
	クジャクが 羽ばたいている	青い花の絵が描かれた 青い花瓶がある
	雪山でキャンプを する夫婦	雪山でスノーボードを 持っている人が3人いる
	サーフィンをしている人が 波にもまれている	波に乗っている人がいる
	オレンジの壁に絵が 描かれている	落書きのある壁の前にある 黄色い壁に落書きが されている
	屋外で家族みんなで ランチをしている	大きなケーキを囲んで パーティをしている
	草原に牛が放し飼いに されている	草原に牛がたくさんいて、 そのうちの1頭は 草を食べている
	夜のゲレンデで スキーヤーがジャンプ している	薄暗い空の下に、 海の上を飛ぶ鳥
	競馬の接戦、<unk>、 <unk>、<unk>、	スケートボードに 乗っている人がいる

図 5 提案キャプショニングモデル実行例

- 人間考案キャプション
- 一般画像キャプション
- 解説文

人間考案キャプション及び解説文は、書籍「#名画で学ぶ主婦業」シリーズ[4][16]に掲載されたものを利用する。  
実験協力者へのアンケート内容は、次のような手順で行う。

まず、実験協力者にこれまでの平均的な美術館の来館頻度を回答してもらう。次に、ある芸術作品の画像に対して、作者や背景をどの程度さらに知りたくなったかを、キャプションごとに選択形式で回答してもらう。評価指標には、リッカート尺度を用いる。評価実験から、対象である芸術作品に興味・関心がない人にとって、どの程度提案キャプションは効果があるのか、客観的に測定する。

## 6 まとめ

本稿では、芸術作品に興味を持たせるために、芸術作品の画像に対して、日常生活で起こりうる場面を想起させるキャプションを自動生成する。提案するキャプションによって、学校の授業や、美術館とは異なる切り口で芸術に触れる機会を作ることを目的とする。提案手法では、まず画像内容を示す日本語文章生成モデルを構築する。次に、Twitter ハッシュタグ「#名画で学ぶ」がついた画像付きツイートを用いて、構築した文章生成モデルをファインチューニングする。提案手法によるキャプションにより日常生活に関連した見方を示すことで、芸術を身近に感じ、芸術に対する背景知識がなくとも興味を持つようになることを期待する。そのため、提案手法によって生成されたキャプションが、実際に興味を持たせるかどうか客観的に確かめる評価実験を行う必要がある。

今回生成されたキャプションでは、情報量が少なく内容の意図が伝わりにくい場合を考えられる。そのため、芸術を身近に感じるようになるには、本研究で生成したキャプションだけでは不十分な点が存在する。今後は生成されたキャプションを用いた、新しい芸術作品の鑑賞支援を提案する必要がある。

具体的には、「複数の芸術作品によるストーリー生成」「属性に合わせたキャプション生成」の二点が挙げられる。一つ目の「複数の芸術作品によるストーリー生成」は、芸術作品及び生成したキャプションのペアを複数組み合わせることで、一つのストーリーが生まれるようする。単体の芸術作品とキャプションのペアよりも情報量が多いため、より興味を持つことができると考えられる。二つ目の「属性に合わせたキャプション生成」では、用いる Twitter ハッシュタグの種類を変えることで、より個人に合わせたキャプションを生成する。また、ターゲットにしたい美術館訪問者の属性に向けて、キャプションを生成することも考えられる。

## 謝 辞

本研究は JSPS 科研費 JP18KT0097, JP18H03243, JP18H03494, C18H032440, 課題設定による先導的人文学・社会科学研究推進事業、および 2020 年度国立情報学研究所共同研究「個人の興味に合わせた文化財間コンテキストの発見」の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] 文化審議会. 文化芸術立国の実現を加速する文化政策(答申)ー「新・文化庁」を目指す機能強化と 2020 年以降への遺産(レガシー)創出に向けた緊急提言ー, 2016. [https://www.bunka.go.jp/seisaku/bunkashikingikai/sokai/sokai\\_16/pdf/bunkageijutsu\\_rikkokutoshin.pdf](https://www.bunka.go.jp/seisaku/bunkashikingikai/sokai/sokai_16/pdf/bunkageijutsu_rikkokutoshin.pdf).
- [2] 文化庁. 文化に関する世論調査 報告書, 2019. [https://www.bunka.go.jp/tokei\\_hakusho\\_shuppan/tokeichosa/pdf/r1393020\\_01.pdf](https://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/pdf/r1393020_01.pdf).
- [3] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. Stair captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017.
- [4] 田中久美子. #名画で学ぶ主婦業. 宝島社, 2018.
- [5] Jocelyn Spence, Benjamin Bedwell, Michelle Coleman, Steve Benford, Boriana N. Koleva, Matt Adams, and Ju Row Farr. Seeing with new eyes: Designing for in-the-wild museum gifting. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, No. 5, pp. 1–13, 2019.
- [6] Lesley Fosh, Steve Benford, Stuart Reeves, and Boriana Koleva. Gifting personal interpretations in galleries. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pp. 625–634, 2014.
- [7] Aditi Mallavarapu, Leilah Lyons, Stephen Uzzo, Wren Thompson, Rinat Levy-Cohen, and Brian Slattery. Connect-to-connected worlds: Piloting a mobile, data-driven reflection tool for an open-ended simulation at a museum. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, No. 7, pp. 1–14, 2019.
- [8] Enid Zimmerman. Reconceptualizing the role of creativity in art education theory and practice. *Studies in Art Education*, Vol. 4, No. 50, pp. 382–399, 2009.
- [9] 金子一夫. 現代美術教育学研究の問題点とその解決—贈与交換論による美術教育の再定義を通して—. 美術教育学, No. 38, pp. 179–191, 2017.
- [10] Reese Muntean, Alissa N. Antle, Brendan Matkin, Kate Hennessy, and Jordan Wilson. Designing cultural values into interaction. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, No. 623, pp. 6062–6074, 2017.
- [11] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 2015.
- [12] Kota Yoshida, Munetaka Minoguchi1, Kenichiro Wani, Akio Nakamura, and Hirokatsu Kataoka. Neural joking machine: Humorous image captioning. arXiv:1805.11850, 2018.
- [13] Bei Liu, Jianlong Fu, Makoto P. Kato, and Masatoshi Yoshikawa. Beyond narrative description: Generating poetry from images by multi-adversarial training. In *Proceedings of the 26th ACM international conference on Multimedia*, MM'18, pp. 783–791, 2018.
- [14] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. arXiv:1504.00325, 2015.
- [15] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, Vol. 27, pp. 3320–3328. Curran Associates, Inc., 2014.
- [16] 田中久美子. #名画で学ぶ主婦業—主婦は再びつぶやく—. 宝島社, 2019.