

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2023年5月19日(19.05.2023)



(10) 国際公開番号

WO 2023/084833 A1

(51) 国際特許分類:

G06T 7/00 (2017.01) G06F 16/583 (2019.01)
G06F 16/56 (2019.01)

(21) 国際出願番号: PCT/JP2022/025412

(22) 国際出願日: 2022年6月24日(24.06.2022)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:

PCT/JP2021/041801 2021年11月12日(12.11.2021) JP

(71) 出願人: 日本電信電話株式会社 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).

(72) 発明者: 西田 光甫 (NISHIDA, Kosuke); 〒1808585 東京都武蔵野市緑町3丁目9-11 N T T 知的財産センタ内 Tokyo (JP). 西田 京介

(NISHIDA, Kyosuke); 〒1808585 東京都武蔵野市緑町3丁目9-11 N T T 知的財産センタ内 Tokyo (JP). 西岡 秀一(NISHIOKA, Shuichi); 〒1808585 東京都武蔵野市緑町3丁目9-11 N T T 知的財産センタ内 Tokyo (JP).

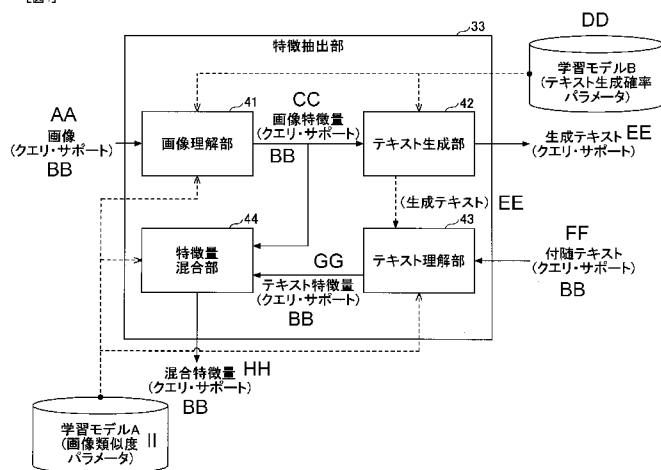
(74) 代理人: 伊東 忠重, 外(ITO, Tadashige et al.); 〒1000005 東京都千代田区丸の内二丁目1番1号 丸の内 M Y P L A Z A (明治安田生命ビル) 16階 Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE,

(54) Title: IMAGE PROCESSING DEVICE, IMAGE PROCESSING METHOD, AND PROGRAM

(54) 発明の名称: 画像処理装置、画像処理方法、及びプログラム

【図4】



33 Feature extraction unit
41 Image understanding unit
42 Text generation unit
43 Text understanding unit
44 Feature amount combining unit
AA Image
BB (Query support)
CC Image feature amount
DD Trained model B (text generation probability parameter)
EE Generated text
FF Accompanying text
GG Text feature amount
HH Combined feature amount
II Trained model A (image similarity parameter)

(57) Abstract: The purpose of the present disclosure is to extract a multimodal feature amount. To this end, an image processing device according to the present disclosure is for extracting a feature amount of image data and comprises: an image understanding unit 41 for vectorizing an image pattern in image data and extracting an image feature amount; a text understanding unit 43 for vectorizing a text pattern in accompanying text data that accompanies the image data and extracting a text feature amount; and a feature amount combining unit 44 for generating a combined feature amount as a feature amount by projecting the image feature amount extracted by the image understanding unit 41 and the text feature amount extracted by the text understanding unit 43 onto the same vector space and combining the image feature amount and the text feature amount.

PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD,
SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT,
TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

- (84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

- 国際調査報告(条約第21条(3))
-

(57) 要約: 本開示内容は、マルチモーダルな特徴量を抽出することを目的とする。そのため、本開示内容の画像処理装置は、画像データの特徴量を抽出する画像処理装置であって、画像データの画像パターンをベクトル化して画像特徴量を抽出する画像理解部41と、画像データに付随している付随テキストデータのテキストパターンをベクトル化してテキスト特徴量を抽出するテキスト理解部43と、画像理解部41によって抽出された画像特徴量とテキスト理解部43によって抽出されたテキスト特徴量を同じベクトル空間に射影して、画像特徴量とテキスト特徴量を混合することで、特徴量としての混合特徴量を生成する特徴量混合部44とを有する。

明 細 書

発明の名称： 画像処理装置、画像処理方法、及びプログラム

技術分野

[0001] 本開示内容は、画像処理装置、画像処理方法、及びプログラムに関する。

背景技術

[0002] 近年、AI(Artificial Intelligence)の進歩により、画像分類の精度が向上している。この画像分類は、例えば、何らかの画像（媒体）からその画像又は画像の中にある特定物が鳩なのか燕なのかをクラス分けすることである。

[0003] 従来、画像分類において、画像データ同士の特徴量の比較だけでなく、画像データに付随する又はユーザにより入力されたテキストデータ同士の比較結果を利用することで、画像分類の精度を向上させる技術が提案されている（非特許文献1 参照）。この場合、例えば、鳩の画像と、画像中に写っている鳩を描写した文であるテキストデータが用いられる。

先行技術文献

非特許文献

[0004] 非特許文献1：Shaping Visual Representations with Language for Few-Shot Classification

発明の概要

発明が解決しようとする課題

[0005] しかしながら、従来技術では、画像データ同士の特徴量の比較、及びテキストデータ同士の比較は独立して行っているに過ぎないため、マルチモーダルな特徴量を抽出することができないという課題が生じる。

[0006] 本発明は、上記の点に鑑みてなされたものであって、従来に比べて、マルチモーダルな特徴量を抽出することを目的とする。

課題を解決するための手段

[0007] 上記課題を解決するため、請求項1に係る発明は、画像データの特徴量を

抽出する画像処理装置であって、前記画像データの画像パターンをベクトル化して画像特徴量を抽出する画像理解部と、前記画像データに付随している付随テキストデータのテキストパターンをベクトル化してテキスト特徴量を抽出するテキスト理解部と、前記画像理解部によって抽出された前記画像特徴量と前記テキスト理解部によって抽出された前記テキスト特徴量を同じベクトル空間に射影して、前記画像特徴量と前記テキスト特徴量を混合することで、前記特徴量としての混合特徴量を生成する特徴量混合部と、を有する画像処理装置である。

発明の効果

[0008] 以上説明したように本発明によれば、従来に比べて、マルチモーダルな特徴量を抽出することができるという効果を奏する。

図面の簡単な説明

[0009] [図1]本実施形態の通信システムの概略図である。

[図2]画像分類装置及び通信端末のハードウェア構成図である。

[図3]本発明の実施形態に係る画像分類装置の機能構成図である。

[図4]画像分類装置における特徴抽出部の詳細な機能構成図である。

[図5]特徴抽出部におけるテキスト生成部の詳細な機能構成図である。

[図6]訓練（学習）フェーズにおいて画像分類装置が実行する処理を示すフローチャートである。

[図7]特徴抽出部が実行する詳細な処理を示すフローチャートである。

[図8]推論フェーズにおいて画像分類装置が実行する処理を示すフローチャートである。

[図9]図9は、実験結果を示す図である。

発明を実施するための形態

[0010] 以下、図面に基づいて本発明の実施形態を説明する。

[0011] [実施形態のシステム構成]

まず、図1を用いて、本実施形態の通信システム1の構成の概略について説明する。図1は、本発明の実施形態に係る通信システムの概略図である。

[0012] 図1に示されているように、本実施形態の通信システム1は、画像分類装置3、及び通信端末5によって構築されている。通信端末5は、ユーザYによって管理及び使用される。

[0013] また、画像分類装置3と通信端末5は、インターネット等の通信ネットワーク100を介して通信することができる。通信ネットワーク100の接続形態は、無線又は有線のいずれでも良い。

[0014] 画像分類装置3は、単数又は複数のコンピュータによって構成されている。画像分類装置3が複数のコンピュータによって構成されている場合には、「画像分類装置」と示しても良いし、「画像分類システム」と示しても良い。

[0015] 画像分類装置3は、AI(Artificial Intelligence)により、画像分類を行う装置である。この画像分類は、例えば、何らかの画像(媒体)からその画像又は画像の中にある特定物が鳩なのか燕なのかをクラス分けすることである。そして、画像分類装置3は、画像分類した結果である分類結果データを出力する。出力方法としては、通信端末5に分類結果データを送信することにより、通信端末5側で分類結果データに係るグラフ等を表示又は印刷したり、画像分類装置3に接続されたディスプレイで上記グラフ等を表示したり、画像分類装置3に接続されたプリンタ等で上記グラフ等を印刷したりすることが挙げられる。

[0016] 通信端末5は、コンピュータであり、図1では、一例としてノート型パソコンが示されているが、ノート型に限るものではなく、デスクトップパソコンであってもよい。また、通信端末は、スマートフォン、又はタブレット型端末であってもよい。図1では、ユーザYが、通信端末5を操作する。

[0017] [画像分類装置及び通信端末のハードウェア構成]

次に、図2を用いて、画像分類装置3及び通信端末5のハードウェア構成を説明する。図2は、画像分類装置及び通信端末のハードウェア構成図である。

[0018] 図2に示されているように、画像分類装置3は、プロセッサ301、メモ

リ３０２、補助記憶装置３０３、接続装置３０４、通信装置３０５、ドライブ装置３０６を有する。なお、画像分類装置３を構成する各ハードウェアは、バス３０７を介して相互に接続される。

[0019] プロセッサ３０１は、画像分類装置３全体の制御を行う制御部の役割を果たし、ＣＰＵ（Central Processing Unit）等の各種演算デバイスを有する。プロセッサ３０１は、各種プログラムをメモリ３０２上に読み出して実行する。なお、プロセッサ３０１には、ＧＰＧＰＵ（General-purpose computing on graphics processing units）が含まれていてもよい。

[0020] メモリ３０２は、ＲＯＭ（Read Only Memory）、ＲＡＭ（Random Access Memory）等の主記憶デバイスを有する。プロセッサ３０１とメモリ３０２とは、いわゆるコンピュータを形成し、プロセッサ３０１が、メモリ３０２上に読み出した各種プログラムを実行することで、当該コンピュータは各種機能を実現する。

[0021] 補助記憶装置３０３は、各種プログラムや、各種プログラムがプロセッサ３０１によって実行される際に用いられる各種情報を格納する。

[0022] 接続装置３０４は、外部装置（例えば、表示装置３１０、操作装置３１１）と画像分類装置３とを接続する接続デバイスである。

[0023] 通信装置３０５は、他の装置との間で各種情報を送受信するための通信デバイスである。

[0024] ドライブ装置３０６は記録媒体３３０をセットするためのデバイスである。ここでいう記録媒体３３０には、ＣＤ－ＲＯＭ（Compact Disc Read-Only Memory）、フレキシブルディスク、光磁気ディスク等のように情報を光学的、電気的あるいは磁氣的に記録する媒体が含まれる。また、記録媒体３３０には、ＲＯＭ（Read Only Memory）、フラッシュメモリ等のように情報を電氣的に記録する半導体メモリ等が含まれていてもよい。

[0025] なお、補助記憶装置３０３にインストールされる各種プログラムは、例えば、配布された記録媒体３３０がドライブ装置３０６にセットされ、該記録媒体３３０に記録された各種プログラムがドライブ装置３０６により読み出

されることでインストールされる。あるいは、補助記憶装置 303 にインストールされる各種プログラムは、通信装置 305 を介してネットワークからダウンロードされることで、インストールされてもよい。

[0026] また、図 2 には、通信端末 5 のハードウェア構成が示されているが、符号が 300 番台から 500 番台に変わっただけで、各構成は同様であるため、これらの説明を省略する。

[0027] [画像分類装置の機能構成]

次に、図 3 を用いて、画像分類装置の機能構成について説明する。図 3 は、本発明の実施形態に係る画像分類装置の機能構成図である。

[0028] 図 3 において、画像分類装置 3 は、入力部 30、読出部 31、選択部 32、特徴抽出部 33、類似度計算部 34、損失計算部 35、パラメータ更新部 36、及び出力部 39 を有している。これら各部は、プログラムに基づき図 2 のプロセッサ 301 による命令によって実現される機能である。

[0029] 更に、図 2 のメモリ 302 又は補助記憶装置 303 には、学習モデル A、B が記憶されている。学習モデル A は、後述の多数の画像類似度パラメータによって構築されている。また、学習モデル B は、後述の多数のテキスト生成確率パラメータによって構築されている。更に、図 2 のメモリ 302 又は補助記憶装置 303 には、教師データとしてのサポートデータの候補群である多数の画像データが記憶されている。また、多数の画像データのそれぞれには、画像の内容を示すテキストデータが付随している。即ち、画像データと付随テキストデータによって 1 ペアのサポートデータが構成されており、図 2 のメモリ 302 又は補助記憶装置 303 には、大量のペアのサポートデータが記憶されている。例えば、1 ペアのサポートデータには、鳩の画像データと、この画像データに付随して画像中に写っている鳩を描写した文であるテキストデータが含まれている。以降、この画像データに付随するテキストデータを「付随テキストデータ」と示す。なお、「付随している」には、テキストデータが画像データに付加されている場合、テキストデータと画像データが別々に入力又は出力されて関連づいている場合が含まれる。なお、

画像データに付随するテキストデータが、画像分類装置3で画像データに基づいて生成され（生成テキストデータ）、この画像データに付加される場合もある。

[0030] 入力部30は、訓練用又は推論用の分類対象（評価対象）データとしてのクエリデータである画像データを入力する。例えば、入力部30は、ユーザYが通信端末5から画像分類装置3に対して送信したクエリデータを画像分類装置に入力する。このクエリデータである画像データには付随テキストデータが付随している。即ち、画像データと付随テキストデータによって1ペアのクエリデータが構成されている。なお、訓練フェーズの場合には、必ず付随テキストデータが付随されているが、推論フェーズの場合には、付随テキストデータが付随されていないことがある。付随テキストデータの付随の方法としては、画像データにキャプション(caption)されている場合と、ユーザYによる手入力の場合がある。多くの機械学習モデルでは、画像分類の推論に人間が介入することができないが、ユーザYによるテキストデータの入力を可能とすることで、ユーザYが画像分類の推論に介入することが可能となる。

[0031] 読出部31は、図2のメモリ302又は補助記憶装置303から、クエリデータと比較するためのサポートデータの候補群（M種類で、かつ各種類でjペア）を読み出す。例えば、Mが100でjが60である。この場合、合計で6000ペアが読み出されることになる。なお、Mが100でjが60は一例であり、Mが100超又は100未満であってもよく、jが60超又は60未満であってもよい。

[0032] 選択部32は、サポートデータの候補群から、クエリデータと比較するためのN種類kペアずつのサポートデータをランダムに選択する。ここでは例えばNが5種類で、kが1ペアずつ（合計5ペア）のサポートデータをランダムに選択したものとして以降の説明を行う。この5種類で1ペアずつのサポートデータを選択する方法は一般的に行われる処理であるが、選択部32は、必ずしも5種類で1ペアずつのサポートデータを選択する必要はない。

例えば、10種類で2ペアずつ（合計20ペア）でもよい。なお、訓練用のサポートデータには、画像データ及び付随テキストデータに加えて、画像データの画像中に写っている被写体の種類（「クラス」）ともいう）を示す情報が付与されている、クラスは、例えば、画像が鳥の画像なのであれば、「鳩」、「鷹」、「燕」等など、鳥の種類を示すものとする。

[0033] 特徴抽出部33は、1ペアの中の画像データから画像特徴量を抽出し、更に、同じペアの中のテキストデータからテキスト特徴量も抽出する。更に、特徴抽出部33は、画像特徴量とテキスト特徴量を混合させて混合特徴量を生成する。また、特徴抽出部33は、画像特徴量からテキストデータも生成する。以降、画像特徴量から生成されたテキストデータを「生成テキストデータ」と示す。即ち、生成テキストデータは、画像由来のテキストデータであり、テキスト由来の付随テキストデータとは種類が異なる。

[0034] <特徴抽出部>

ここで、図4を用いて、画像分類装置における特徴抽出部を詳細に説明する、図4は、画像分類装置における特徴抽出部の詳細な機能構成図である。

[0035] 図4に示されているように、特徴抽出部33は、画像理解部41、テキスト生成部42、テキスト理解部43、及び特徴量混合部44を有している。なお、画像理解部41、テキスト生成部42、特徴量混合部44、及び類似度計算部34には、任意のニューラルネットワークを使用することができる。例えば、画像理解部41に4層のCNN(Convolutional Neural Network)が使用されている。テキスト生成部42及びテキスト理解部43の事前学習を行うことで、テキスト生成能力及びテキスト理解能力を向上させる。

[0036] これらのうち、画像理解部41は、入力部30からクエリデータのうちの画像データ（第1の画像データの一例）を取得すると共に、選択部32から5種類1ペアのうちの特定の1ペアのサポートデータのうちの画像データ（第2の画像データの一例）を取得する。そして、画像理解部41は、クエリデータの画像データの画像パターンをベクトル化してクエリ用の画像特徴量を抽出すると共に、サポートデータの画像データの画像パターンをベクトル

化してサポート用の画像特徴量を抽出する。画像特徴量はベクトルであり、テキスト生成部42は任意のニューラルネットワークを利用でき、画像特徴量を初期値とするRNN(Recurrent neural network)やTransformerが一般的である。

[0037] テキスト生成部42は、画像理解部41が抽出したクエリ用の画像特徴量をテキストデータのベクトル空間に射影し、デコードすることで、画像由来のクエリ用の生成テキストデータを生成すると共に、画像理解部41が抽出したサポート用の画像特徴量をテキストデータのベクトル空間に射影し、デコードすることで、画像由来のサポート用の生成テキストデータを生成する。

[0038] (テキスト生成部)

ここで、図5を用いて、テキスト生成部42について、更に詳細に説明する。図5は、テキスト生成部の詳細な機能ブロック図である。

[0039] 図5に示されているように、テキスト生成部42は、線形変換層421、及びデコーダ422を有している。更に、線形変換層421は線形変換層用パラメータ421pを保持しており、デコーダ422はデコーダパラメータ422pを保持している。なお、線形変換層用パラメータ421p及びデコーダパラメータ422pは、図4に示されている学習モデルBに含まれている。

[0040] 線形変換層421は、線形変換層用パラメータ421pを用いることで、画像理解部41から取得した画像特徴量を付随テキストデータのベクトル空間に射影することで、画像由来の特徴量を抽出する。

[0041] デコーダ422は、デコーダパラメータ422pを用いることで、線形変換層421から取得した特徴量から、画像由来の生成テキストデータを生成する。

[0042] ここで、テキスト生成部42及びテキスト理解部43に、既存の事前学習済み言語モデルを流用することで、テキスト生成部42及びテキスト理解部43の事前学習を行ったとみなすことができる。但し、テキスト生成部42

については既存の言語モデルをそのまま利用することができない。これは、テキストを生成する能力を有する既存の言語モデルが、Encoder-Decoder型の構造を持つためである。

[0043] Encoder-Decoder型の構造を持つ言語モデルについては、例えば、参考文献1に開示されている。

＜参考文献1＞Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Encoder-Decoder型の構造とは、まずテキストを入力として与えられてEncoderによって特徴量に変換し、Decoderにその特徴量を入力し、Decoderがテキストを生成する構造のことをいう。本実施形態においては、テキスト生成部62に画像特徴量が入力されるため、参考文献1における既存の言語モデルのEncoderを使用せず、代わりに線形変換層などの任意のニューラルネットワークをDecoderの前に追加する。この構成によって、画像特徴量を言語モデルに適した特徴量に変換し、Decoderに入力し、テキストを生成することが可能になる。

[0044] 続いて、図4に戻り、テキスト理解部43は、入力部30からクエリデータのうちの付随テキストデータを取得すると共に、選択部32から5種類1ペアのうちの特定の1ペアのサポートデータのうちの付随テキストデータを取得する。そして、テキスト理解部43は、クエリデータの付随テキストデータのテキストパターンをベクトル化してクエリ用のテキスト特徴量を抽出すると共に、サポートデータの付随テキストデータのテキストパターンをベクトル化してサポート用のテキスト特徴量を抽出する。

[0045] 例えば、テキスト理解部43は、BERT(Bidirectional Encoder Representations from Transformers)などの既存の言語モデルによってテキストデータをベクトルに変換する。

[0046] なお、上述のように、訓練フェーズでは画像データに付随テキストデータが付随されているが、推論フェーズでは画像データに付随テキストデータが付随されていない場合がある。このような場合には、テキスト理解部43は

、テキスト生成部42が生成した画像由来のクエリ用の生成テキストデータを付随テキストデータとする（みなす）ことで、画像由来ではあるがクエリ用のテキストデータの特徴量を抽出する。

[0047] 次に、特徴量混合部44は、画像理解部41によって抽出されたクエリ用の画像特徴量と、テキスト理解部43によって抽出されたクエリ用のテキスト特徴量とを同じベクトル空間に射影して、クエリ用の画像特徴量とクエリ用のテキスト特徴量を混合することで、クエリ用の特徴量としての混合特徴量を生成する。同様に、特徴量混合部44は、画像理解部41によって抽出されたサポート用の画像特徴量とテキスト理解部43によって抽出されたサポート用のテキスト特徴量とを同じベクトル空間に射影して、サポート用の画像特徴量とサポート用のテキスト特徴量を混合することで、サポート用の特徴量としての混合特徴量を生成する。なお、画像特徴量とテキスト特徴量を混合する処理は、一方の特徴量のベクトル空間に他方の特徴量を射影する場合と、互いに異なる第三のベクトル空間に射影する場合がある。

[0048] 例えば、特徴量混合部44は、画像特徴量とテキスト特徴量の双方を類似度計算に反映することが可能となる。特徴量混合部44は画像特徴量とテキスト特徴量の双方を入力として受け付ける任意のニューラルネットワークが利用可能である。

[0049] ここで、特徴量混合部44について、更に詳細に説明する。

[0050] 特徴量混合部44として以下のモデルを利用する。画像特徴量を x_{image} 、テキスト理解部43が出力するテキスト特徴量を x_{Lang} とする。MLP (Multilayer perceptron)を3層ニューラルネットワークとする。Linearを2次元への線形変換層とする。 $[;]$ をベクトルを縦に繋げる操作とする。このとき、特徴量混合部44が出力するベクトル h は、（式1）、（式2）、（式3）により、以下のように示される。

[0051]

[数1]

$$z_{lang} = MLP(x_{lang}) \quad \dots (式1)$$

[0052] [数2]

$$\lambda_{image}, \lambda_{lang} = softmax(Linear([x_{image}; z_{lang}])) \quad \dots (式2)$$

[0053] [数3]

$$h = \lambda_{image}x_{image} + \lambda_{lang}z_{lang} \quad \dots (式3)$$

まず、特徴量混合部44は、(式1)を用いて、MLPによってBERTが出力するテキスト特徴量を画像特徴量と同じ空間に射影する(z_{Lang})。

[0054] 次に、特徴量混合部44は、(式2)を用いて、画像特徴量とテキスト特徴量の重要度を λ_{image} と λ_{Lang} によって動的に決定している。 λ_{image} 、 λ_{Lang} は、和が1になる非負の数であることがsoftmax演算によって保証されている。例えば、画像データの元の解像度が低い場合(画像中で目的の物体が極端に小さく、ぼやけて写っている場合)には、画像データに付随されている付随テ

キストデータが分類結果に与える度合いを大きくするように λ_{image} と λ_{Lang} が動的に決定される。また、ユーザ自身が λ_{image} と λ_{Lang} を0から1の範囲に調整することで、ユーザが入力したテキストが分類結果に反映される度合いを手動で変更することも可能である。なお、Linearは重み行列を左から掛け、バイアスベクトルを加える操作のことである。Linear操作中の重み行列とバイアスベクトルが、学習モデルAの画像類似度パラメータ及び学習モデルBのテキスト生成確率パラメータに含まれる。

[0055] 最後に、特徴量混合部44は、(式3)を用いて、出力する特徴量を重要度による重み付き和によって決定している。

[0056] また、図4に示されているように、学習モデルAの画像類似度パラメータは、画像理解部41、テキスト理解部43、及び特徴量混合部44が各処理を実行する場合に使用される。学習モデルBのテキスト生成確率パラメータは、画像理解部41及びテキスト生成部42が各処理を実行する場合に用いられる。但し、推論フェーズの場合で、画像データに付随テキストデータが付随されているときには、学習モデルBのテキスト生成確率パラメータは使用されない。

[0057] なお、訓練フェーズの場合には、画像データに付随テキストデータが付随されているときであっても、学習モデルBのテキスト生成確率パラメータは使用され、訓練(学習)による更新も行われる。このようにするのは、推論フェーズの場合に、画像データに付随テキストデータが付随されていないときでも、テキスト生成部42が生成テキストデータを生成することができるようにするためである。また、学習モデルBを訓練(学習)することで、テキスト生成確率パラメータを使用する画像理解部41の理解能力が向上するという良い影響を与えるためでもある。

[0058] 続いて、図3に戻り、類似度計算部34は、クエリ用の混合特徴量とサポート用の混合特徴量を比較して画像類似度を計算する。推論フェーズの場合、この画像類似度は出力部39に出力されて画像分類の分類結果データに用いられる。一方、訓練フェーズの場合、この画像類似度は損失計算部35に

出力される。

[0059] 例えば、類似度計算部34は、双線形層である。ここで、N-way k-shot image classificationを考える。類似度計算部34では、まず各クラスについてk個のサポート特徴量（ベクトル）が与えられる。これらを平均したベクトルをクラス特徴量とする。N個のクラス特徴量（ベクトル）を並べた行列をXとする。クエリデータの特徴量をy、学習可能なパラメータをWとする。このとき、クエリデータが各クラスであることの分類スコアは以下のように表される。

[0060] [数4]

$$X^T W y \in \mathbb{R}^N$$

このベクトルの各成分が、クエリデータが各クラスに属することの確からしさを示す。

[0061] 損失計算部35は、画像類似度から損失関数値を計算する。また、損失計算部35は、クエリデータ・サポートデータの生成テキストデータと、クエリデータ・サポートデータの生成確率分布と、クエリデータ・サポートデータの付随テキストデータから損失関数値を計算する。

[0062] 例えば、損失計算部35が計算する損失関数には、類似度計算部34の分類スコアやテキスト生成に関する任意の損失を利用することができる。標準的にはCross-Entropy Lossや負の対数尤度関数が用いられる。

[0063] パラメータ更新部36は、類似度計算部34によって計算された画像類似度から損失計算部35によって計算された損失関数値に基づいて、特徴抽出部33、類似度計算部34を構成するニューラルネットワークの学習モデルAの画像類似度パラメータを更新する。この場合、損失計算部35は、サポ

ートデータの画像データとクエリデータの画像データの類似度が縮まるように、更には、不正解画像との類似度は大きくなるように学習を行う。

[0064] また、パラメータ更新部36は、損失計算部35によって計算された損失関数値に基づいて、特徴抽出部33、類似度計算部34を構成するニューラルネットワークの学習モデルBのテキスト生成確率パラメータを更新する。この場合、損失計算部35は、生成テキストデータが付随テキストデータに類似する確率が高くなるように学習を行う。

[0065] 例えば、パラメータ更新部56は、損失計算部35によって計算された損失に基づいて、損失の勾配を計算し、パラメータを更新する。

[0066] [実施形態の処理又は動作]

続いて、図6乃至図8を用いて、本実施形態の処理又は動作について詳細に説明する。なお、訓練（学習）フェーズと推論フェーズに分けて説明する。

[0067] <訓練フェーズ>

まずは、図6及び図7を用いて、訓練フェーズについて説明する。図6は、訓練（学習）フェーズにおいて画像分類装置が実行する処理を示すフローチャートである。

[0068] まず、入力部30は、訓練用の教師データ（クエリデータ）を入力する（S10）。読出部31は、訓練用の教師データ（サポートデータ）の候補群を読み出す（S11）。選択部32は、候補群から教師データとしてランダムに5種類1ペア（画像データと付随テキストデータ）のサポートデータを選択する（S12）。また、選択部32は、同じ5種類から任意の数のペアをクエリデータとして選択する。この際、選択部32は、選択した各クエリデータに対して、同じ種類のサポートデータを、該クエリデータに対する正解であると定義し、異なる種類のサポートデータを、該クエリデータに対する不正解であると定義することで、サポートデータに正解又は不正解を定義したデータを付加する。例えば、クエリデータが「鳩」を示す場合、サポートデータの5種の内、「鳩」を示すサポートデータは正解、その他の種類（クラス）を示すサポートデータは不正解と定義される。なお、正解又は不正

解の定義は、読出部 31 によって行われてもよい。

[0069] 次に、特徴抽出部 33 は、入力部 30 から取得したクエリデータに基づいてクエリ用の混合特徴量を生成すると共に、選択部 32 が選択した 5 種類 1 ペア（合計 5 ペア）のサポートデータのうちの所定の 1 つのサポートデータに基づいてサポート用の混合特徴量を生成する（S13）。この際、特徴抽出部 33 は、正解又は不正解の定義済みのセットデータ（クエリデータ、サポートデータ、及び正解又は不正解の定義データ）を受け取り、このセットデータ内に含まれるクエリデータ、サポートデータの混合特徴量を算出して類似度計算部 34 へ出力する。なお、この際、サポートデータのペア数 k が 2 以上の場合は、各ペアの画像のデータの画像特徴量を平均したベクトルを、そのサポートデータの画像特徴量とすればよい。

[0070] ここで、図 7 を用いて、特徴抽出部が実行する詳細な処理を説明する。図 7 は、特徴抽出部が実行する詳細な処理を示すフローチャートである。

[0071] 図 7 に示されているように、画像理解部 41 は、クエリデータ及びサポートデータの各画像データに基づいて、各画像特徴量（クエリ用の画像特徴量、サポート用の画像特徴量）を抽出する（S131）。テキスト生成部 42 は、各画像特徴量に基づいて、各生成テキストデータを生成する（S132）。訓練フェーズでは、後述のステップ S133、S135 は実行されず、続いて、テキスト理解部 43 は、クエリデータ及びサポートデータの各付随テキストデータに基づいて、各テキスト特徴量（クエリ用のテキスト特徴量、テキスト用のテキスト特徴量）を抽出する（S134）。特徴量混合部 44 は、クエリ用の画像特徴量とクエリ用のテキスト特徴量を混合してクエリ用の混合特徴量を生成すると共に、サポート用の画像特徴量とサポート用のテキスト特徴量を混合してサポート用の混合特徴量を生成する（S136）。

[0072] 続いて、図 6 に戻り、類似度計算部 34 は、クエリ用の混合特徴量（第 1 の混合特徴量の一例）とサポート用の混合特徴量（第 2 の混合特徴量の一例）を比較して、画像類似度を計算する（S14）。この際、類似度計算部 3

4は、セットデータ内に含まれるクエリデータとサポートデータ各ペアの類似度を計算して損失計算部に渡す。

[0073] そして、特徴抽出部33が、選択部32によって選択された5種類1ペア（合計5ペア）のサポートデータのうちの5ペア全ての類似度の計算が終了しているか否かを判断する（S15）。そして、特徴抽出部33が、5ペア全てのサポートデータに対する類似度の計算が終了していないと判断した場合には（S15；NO）、ステップS13に戻り、類似度の計算が終了していないサポートデータに対して、ステップS13以降が行われる。なお、入力部30から取得されたクエリデータに関しては、既に混合特徴量が生成されているため、ステップS13以降の再処理は行われない。

[0074] 一方、ステップS15において、特徴抽出部33が、5ペア全ての類似度の計算が終了していると判断した場合には（S15；YES）、損失計算部35は、損失を計算する（S16）。この際、損失計算部35は、各セットデータ内に含まれるクエリデータとサポートデータのペアの各類似度、クエリデータに対するサポートデータの各ペアの正解又は不正解の定義データを基に損失を計算する。なお、この類似度には、画像間の類似度と、付随テキスト間の類似度が含まれる。

[0075] <補足A>

なお、教師データ中の付随テキストデータが利用可能な訓練フェーズでは、教師データ中の付随テキストデータは、テキスト理解部43に入力されている。しかし、推論フェーズでは、テキスト生成部42が生成した生成テキストデータが入力される可能性があるため、訓練フェーズと推論フェーズで乖離が発生してしまう。そこで、本実施形態では、画像特徴量及び付随テキストデータから計算されたクロスエントロピー損失 $L_{\text{class, gold}}$ だけでなく、画像特徴量及びテキスト生成部42が生成した生成テキストデータとから計算されたクロスエントロピー損失 $L_{\text{class, gen}}$ も併用して学習する。この処理により、訓練フェーズと推論フェーズの乖離を抑えることが可能になる。

[0076] <補足B>

また、（式４）に示す対照学習の損失 L_{cntr} を利用することでモデルがテキストデータの微細な違いを捉えて特徴量を獲得することも可能である。

[0077] [数5]

$$L_{cntr} = \frac{1}{2N} \sum_c \log \frac{\exp[\cos(v_{gold}^c \text{ }^T v_{gen}^c)/\tau]}{\sum_{c'} \exp[\cos(v_{gold}^c \text{ }^T v_{gen}^{c'})/\tau]} + \frac{1}{2N} \sum_{c'} \log \frac{\exp[\cos(v_{gold}^{c'} \text{ }^T v_{gen}^{c'})/\tau]}{\sum_c \exp[\cos(v_{gold}^c \text{ }^T v_{gen}^{c'})/\tau]} \quad \dots \text{ (式4)}$$

ここで、

[0078] [数6]

$$v_{gold}^c$$

は、特徴量混合部４４が教師データ中の付随テキストデータを入力して作成したベクトル h を、類似度計算部３４がクラス c の k 個のサポート特徴量について平均したベクトルである。

[0079]

[数7]

$$v_{gen}^c$$

は、損失計算部35が、同じようにテキスト生成部42が生成した生成テキストデータの入力に基づいて計算したベクトルである。なお、対照学習 (contrastive learning) とは、正例と負例を区別し、入力と正例が近づくように、入力と負例が遠ざかるように行う学習である。ここでは、パラメータ更新部36が、 L_{ctr} を用いて、同じクラスの画像に付随する「教師データ中の付随テキストデータ」と「テキスト生成部が生成した生成テキストデータ」の各テキスト特徴量

[0080] [数8]

$$v_{gold}^c, v_{gen}^c$$

が近づくようにテキスト生成確率パラメータを更新すると共に、異なるクラスの画像に付随する「教師データ中の付随テキストデータ」と「テキスト生成部が生成した生成テキストデータ」の特徴量

[0081]

[数9]

$$v_{gold}^c, v_{gen}^c$$

は遠ざかるようにテキスト生成確率パラメータを更新する。このように、対照学習を行うことで、テキスト理解部43が両テキストデータの微細な違いを捉えた特徴量を出力するように学習を進めることができるという効果が生じる。

[0082] 以上、損失関数Lは（式5）により上述の4つのLを合計した値となる。

[0083] [数10]

$$L = L_{class,gold} + L_{class,gen} + \lambda_{text}L_{text} + \lambda_{cntr}L_{cntr} \quad \cdots (式5)$$

なお、

[0084] [数11]

$$\tau, \lambda_{text}, \lambda_{cntr}$$

はハイパーパラメータである。

[0085] 次に、パラメータ更新部36が、損失の勾配を計算し、学習モデルAの画像類似度パラメータ、及び学習モデルBのテキスト生成確率パラメータを更新（訓練）する（S17）。この際、パラメータ更新部36は、上記損失を最小化するように、パラメータを更新する。

[0086] 次に、選択部32が、規定回数（例えば、20回）の選択を終了したかを判断する（S18）。例えば、選択部32が、規定回数として20回の選択をする場合、1回の選択で5ペアのサポートデータを選択するため、合計で100ペアのサポートデータを選択することになる。但し、選択部32は、候補群からランダムに5種類1ペア（合計5ペア）のサポートデータを選択するため、同じサポートデータが複数回選択される場合が生じる。

[0087] そして、上記ステップS18において、選択部32が、規定回数の選択が終了していないと判断した場合には（S18；NO）、ステップS12の処理に戻り、選択部32は新たに候補群からランダムに5種類1ペア（合計5ペア）のサポートデータを選択し、その後、ステップS13以降の処理が行われる。

[0088] 一方、上記ステップS18において、選択部32が、規定回数の選択が終了していると判断した場合には（S18；YES）、図6に示されている訓練フェーズの処理が終了する。

[0089] <推論フェーズ>

次に、図7及び図8を用いて、訓練フェーズについて説明する。図8は、推論フェーズにおいて画像分類装置が実行する処理を示すフローチャートである。

[0090] まず、入力部30は、推論用の分類対象データであるクエリデータを入力する（S30）。読出部31は、推論用のサポートデータを読み出す（S31）。

[0091] 次に、特徴抽出部33は、入力部30から取得した分類対象データであるクエリデータに基づいてクエリ用の混合特徴量を生成すると共に、選択部32が選択した5種類1ペア（合計5ペア）のサポートデータのうちの所定の

1つのサポートデータに基づいてサポート用の混合特徴量を生成する（S 3 2）。ここで、図 7 を用いて、特徴抽出部が実行する詳細な処理を説明する。図 7 は、特徴抽出部が実行する詳細な処理を示すフローチャートである。

[0092] 図 7 に示されているように、画像理解部 4 1 は、クエリデータ及びサポートデータの各画像データに基づいて、各画像特徴量（クエリ用の画像特徴量、サポート用の画像特徴量）を抽出する（S 1 3 1）。テキスト生成部 4 2 は、各画像特徴量に基づいて、各生成テキストデータを生成する（S 1 3 2）。推論フェーズでは、後述のステップ S 1 3 3, S 1 3 5 は実行される。

[0093] <補足 C>

なお、推論フェーズでは、beam-searchを用いて生成テキストデータが生成される。そのため、例えばbeam幅を10とすると、各時刻の上位10個のトークンを生成候補とすることができる。しかし、beam-searchは計算量が重いいため、訓練フェーズに用いることができない。そこで、訓練フェーズと推論フェーズの乖離を抑えるために、テキスト生成部 4 2 は、「言語生成」の処理を行う。

[0094] 「言語生成」の処理は、訓練フェーズでは貪欲法とランダムサンプリングの併用によって生成テキストデータを生成する処理である。貪欲法時には、テキスト生成部 4 2 は、各時刻で最も上位のトークンを選び、生成テキストデータを生成する。ランダムサンプリング時には、テキスト生成部 4 2 は、各時刻で上位の所定数（例えば、top20）のトークンからサンプリングによってトークンを選ぶことで、生成テキストデータを生成する。また、テスト時には、入力部 3 0 から付随テキストデータが与えられない場合は、テキスト生成部 4 2 が生成テキストデータをテキスト理解部 4 3 に渡す。この際、テキスト生成部 4 2 は、length penalty を 0. 5、ビーム幅を 5 としたビームサーチを行って、クエリデータが各クラスであることの分類スコアが最大となる生成テキストデータを生成してテキスト理解部 4 3 に渡す。

[0095] よって、テキスト理解部 4 3 は、クエリデータとサポートデータの両方にそれぞれ付随テキストデータが含まれているか、即ち、クエリデータの画像

データとサポートデータの画像データの両方にそれぞれ付随テキストデータが付随されているかを判断する（S 1 3 3）。そして、テキスト理解部 4 3 は、クエリデータとサポートデータの両方にそれぞれ付随テキストデータが含まれている、即ち、クエリデータの画像データとサポートデータの画像データの両方にそれぞれ付随テキストデータが付随されていると判断した場合には（S 1 3 3 ; Y E S）、テキスト理解部 4 3 は、クエリデータ及びサポートデータの各付随テキストデータに基づいて、各テキスト特徴量（クエリ用のテキスト特徴量、テキスト用のテキスト特徴量）を抽出する（S 1 3 4）。

[0096] 一方、ステップ S 1 3 3 において、テキスト理解部 4 3 が、クエリデータとサポートデータの両方にそれぞれ付随テキストデータが含まれていないと判断した場合、即ち、クエリデータの画像データとサポートデータの画像データの両方にそれぞれ付随テキストデータが付随されていないと判断した場合には（S 1 3 3 ; N O）、テキスト理解部 4 3 は、以下の処理を行う。

[0097] 即ち、上記の場合（S 1 3 3 ; N O）、サポートデータに付随データが含まれていないときには、テキスト理解部 4 3 は、クエリデータの付随テキストに基づいてテキスト特徴量を抽出し、サポートデータの生成テキストに基づいてテキスト特徴量を抽出する（S 1 3 5）。同様の場合であって（S 1 3 3 ; N O）、クエリデータに付随データが含まれていないときには、テキスト理解部 4 3 は、サポートデータの付随テキストに基づいてテキスト特徴量を抽出し、クエリデータの生成テキストに基づいてテキスト特徴量を抽出する（S 1 3 5）。同様の場合であって（S 1 3 3 ; N O）、クエリデータとサポートデータの両方にそれぞれ付随テキストデータが含まれていないときには、テキスト理解部 4 3 は、クエリデータ及びサポートデータのそれぞれの生成テキストに基づいて、それぞれのテキスト特徴量を抽出する（S 1 3 5）。

[0098] <補足 D>

なお、テキスト生成部 4 2 が、生成テキストデータの生成の損失として、t

teacher-forcingとクロスエントロピー損失で計算された L_{text} のみで学習すると、「教師データ中の付随テキストデータを再現する」ことが学習の目的となり、「画像分類に資する生成テキストデータを生成する」ことは、考慮できていない。これは、テキスト生成部42の離散的処理により勾配グラフが壊れることが原因である。つまり、画像分類によって得られる損失($L_{\text{class, gold}}, L_{\text{class, gen}}$)から得られる勾配が、誤差逆伝播法によりテキスト生成部42にまで伝播しないことが原因である。これを改善するために、テキスト理解部43が以下に示す処理を行なった。

[0099] 即ち、本実施形態では、テキスト理解部43が、(式6)により、まず x_{image} をテキスト特徴量と同じベクトル空間に写像(射影)する。

[0100] [数12]

$$f_{\text{I2T}}(x_{\text{image}}) = \text{Linear}(\text{LayerNorm}(x_{\text{image}})) \quad \dots (\text{式6})$$

ここで、LayerNorm はLayer Normalizationを示す(参考文献2)。

<参考文献2> Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

テキスト理解部43は、得られた表現をテキスト生成部42に長さ1の系列として渡す。そして、テキスト生成部42は、自己回帰的にj番目のトークン t_j を以下の(式7)で示す確率 p_j に従い生成する。即ち、確率 p_j は、テキスト生成部42によって生成された生成テキストデータに係る所定番目(j番目)のトークン t_j が、どのくらいの確率で正しそうか(確からしさ)を示す

。

[0101] [数13]

$$p_j = Pr(t_j; f_{I2T}(x_{image}), t_{0:j-1}) \quad \cdots \text{(式7)}$$

また、テキスト理解部43は、(式8)で示すテキスト特徴量 x_{Lang} を抽出するため、生成テキストデータに係る所定番目(j番目)のトークン t_j の確からしさを示す確率 p_j を用いた重み付きプーリングを行う。

[0102] ここで、テキスト理解部43の最終層の隠れ状態の系列を H_{BERT} とする。トークンがストップワードの場合 $w_j=0$ 、それ以外は $w_j=1$ とする。また、テキスト生成部42の代わりに入力部30によって付随テキストデータが入力された場合は、全トークンで $p_j=1$ とする。

[0103] [数14]

$$x_{lang} = \frac{1}{\sum p_j w_j} \sum p_j w_j h_{BERT,j} \quad \cdots \text{(式8)}$$

このように、（式 8）から定義できるテキスト特徴量 x_{Lang} を採用することで、画像分類から得られる損失の勾配が、確率 p_j を利用してテキスト生成部 42 に伝播することができる。

[0104] そして、ステップ S 134 又は S 135 の後、特徴量混合部 44 は、クエリ用の画像特徴量とクエリ用のテキスト特徴量を混合してクエリ用の混合特徴量を生成すると共に、サポート用の画像特徴量とサポート用のテキスト特徴量を混合してサポート用の混合特徴量を生成する（S 136）。

[0105] 続いて、図 8 に戻り、類似度計算部 34 は、クエリ用の混合特徴量（第 1 の混合特徴量の一例）とサポート用の混合特徴量（第 2 の混合特徴量の一例）を比較して、画像類似度を計算する（S 33）。

[0106] 次に、特徴抽出部 33 が、選択部 32 によって選択された 5 種類 1 ペア（合計 5 ペア）のサポートデータのうちの 5 ペア全てのサポートデータに対する比較が終了しているか否かを判断する（S 34）。そして、特徴抽出部 33 が、5 ペア全てのサポートデータに対する比較が終了していないと判断した場合には（S 35；NO）、ステップ S 32 に戻り、5 種類 1 ペア（合計 5 ペア）のサポートデータの比較が終了していないサポートデータに対して、ステップ S 32 以降が行われる。なお、入力部 30 から取得された分類対象データであるクエリデータに関しては、既に混合特徴量が生成されているため、ステップ S 32 以降の再処理は行われない。

[0107] 一方、ステップ S 34 において、特徴抽出部 33 が、5 ペア全てのサポートデータに対する比較が終了していると判断した場合には（S 34；YES）、出力部 39 は、これまでの比較結果に基づいて、分類結果を示す分類結果データを出力する（S 35）。分類結果データには、例えば、分類対象データに係る画像は、鳩の画像である、鳩の画像である可能性が 90% であり、他の鳥の画像である可能性が 10% である等が示されている。

[0108] [実験結果]

続いて、実験設定及び実験結果について説明する。

[0109] <実験設定>

データセットCaltech-UCSD Birds (CUB) (参考文献3, 4) を5-way 1-shot 分類問題として用いた。本データは200種の鳥の品種をクラスとしており、1種に対して40-60枚の画像がある。200種の画像のうち、100種が訓練用、50種が開発用、50種がテスト用である。

<参考文献3> Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

<参考文献4> Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In CVPR, pp. 49-58, 2016.

<実験結果>

図9は、実験結果を示す図である。図9において、本実施形態による提案手法を全て実行した結果が1段目1行目のLIDEの値である。そこから、上述の<補足A>、<補足B>、<補足C>、及び<補足D>をそれぞれ行わない場合の実験結果が、2段目1行目、2段目2行目、3段目4行目、3段目3行目に示されている。また、上述の<補足A>、<補足B>、<補足C>、及び<補足D>の全てを行わない場合の実験結果が、2段目5行目に示されている。これにより、<補足A>、<補足B>、<補足C>、及び<補足D>による提案手法の4要素全てが性能向上に貢献することが確認できた。

[0110] [実施形態の主な効果]

以上説明したように本実施形態によれば、画像分類装置3は、画像データの画像特徴量と、画像データに付随している付随テキストデータのテキスト特徴量を混合することで、混合特徴量を生成する。これにより、画像分類装置3は、特徴抽出装置として、単に、画像データ同士の特徴量の比較、及びテキストデータ同士の比較の場合に比べて、マルチモーダルな特徴量を抽出することができるという効果を奏する。また、画像分類装置3は、より高精度な画像データに関する特徴量を抽出することで、より高精度な画像分類を行うことができるという効果を奏する。

[0111] また、本実施形態によれば、以下に示す処理により、精度良く画像を分類することができ、推論フェーズでユーザが付随テキストデータの入力によって分類結果に介入することが可能となるという効果を奏する。

[0112] <補足A>によって、少数事例画像分類タスクにおいてテキストデータを使って情報の補足を行う場合に、画像分類の損失関数の併用によって訓練フェーズと推論フェーズで利用可能なテキストデータに関する乖離を抑える。

[0113] また、<補足B>によって、対照学習によりテキストデータの微細な違いを捉えた特徴量を出力するようにテキスト理解部43の学習が進む。

[0114] 更に、<補足C>によって、ランダムサンプリングによる生成テキストデータの生成によって訓練フェーズと推論フェーズのテキスト生成手法に関する乖離を抑える。

[0115] また、<補足D>によって、テキスト生成スコア（トークン毎の生成スコア）を利用したプーリングによってテキスト生成部42が画像分類の性能向上を考慮した学習を行う。

[0116] [全体の補足]

本発明は上述の実施形態に限定されるものではなく、以下に示すような構成又は処理（動作）であってもよい。

[0117] 画像分類装置3はコンピュータとプログラムによって実現できるが、このプログラムを（非一時的）記録媒体に記録することも、通信ネットワーク100を介して提供することも可能である。

[0118] 上記実施形態では、画像分類装置3が示されているが、特徴抽出部33に特化した場合には、特徴抽出装置と表現することが可能である。また、画像分類装置3及び特徴抽出装置は、共に画像処理装置と表現することが可能である。

[0119] また、上記実施形態以外に、ニューラルネットワークの学習で用いられる任意の処理を、上記実施形態に追加することができる。例えば、入力する付随テキストデータについてルールベースによる言い換えを行うことでデータの数を水増しすることができる。言い換え例としては、「This bird is big

」の「big」を「large」に言い換えて、「This bird is large」とする言い換えがある。

[0120] 〔付記項〕

上述の実施形態には、以下に示す発明としても表すことができる。

[0121] 〔付記項 1〕

画像データの特徴量を抽出する画像処理装置であって、
前記画像データの画像パターンをベクトル化して画像特徴量を抽出する画像理解ステップと、
前記画像データに付随している付随テキストデータのテキストパターンをベクトル化してテキスト特徴量を抽出するテキスト理解ステップと、
前記画像理解ステップによって抽出された前記画像特徴量と前記テキスト理解ステップによって抽出された前記テキスト特徴量を同じベクトル空間に射影して、前記画像特徴量と前記テキスト特徴量を混合することで、前記特徴量としての混合特徴量を生成する特徴量混合ステップと、
を実行する画像処理装置。

[0122] 〔付記項 2〕

前記画像理解ステップ、前記テキスト理解ステップ、及び前記特徴量混合ステップは、それぞれニューラルネットワークで実現され、前記画像理解ステップ、前記テキスト理解ステップ、及び前記特徴量混合ステップは前記ニューラルネットワークのモデルパラメータに基づいて処理を行う、付記項 1 に記載の画像処理装置。

[0123] 〔付記項 3〕

付記項 2 に記載の画像処理装置であって、
前記プロセッサは、
前記特徴量混合ステップによって生成された第 1 の画像データに係る第 1 の混合特徴量、及び前記特徴量混合ステップによって生成された第 2 の画像データに係る第 2 の混合特徴量の画像類似度を計算する類似度計算ステップと、

前記類似度計算ステップによって計算された前記画像類似度に基づいて、前記モデルパラメータに含まれる画像類似度パラメータを更新するパラメータ更新ステップと、
を実行する画像処理装置。

[0124] 〔付記項 4〕

付記項 2 に記載の画像処理装置であって、
前記プロセッサは、
前記画像理解ステップによって抽出された前記画像特徴量を前記付随テキストデータのベクトル空間に射影することで生成テキストデータを生成するテキスト生成ステップと、
前記テキスト生成ステップによって生成された前記生成テキストデータと前記付随テキストデータに基づいて、前記モデルパラメータに含まれるテキスト生成確率パラメータを更新するパラメータ更新ステップと、
を実行する画像処理装置。

[0125] 〔付記項 5〕

付記項 1 に記載の画像処理装置であって、
前記プロセッサは、
前記画像理解ステップによって抽出された前記画像特徴量を前記付随テキストデータのベクトル空間に射影することで生成テキストデータを生成するテキスト生成ステップを実行し、
前記画像データに前記付随テキストデータが付随していない場合には、前記テキスト理解ステップは、前記テキスト生成ステップによって生成された前記生成テキストデータを前記付随テキストデータとすることで、前記テキスト特徴量を抽出する処理を含む、画像処理装置。

[0126] 〔付記項 6〕

前記パラメータ更新ステップは、前記画像特徴量及び前記付随テキストデータに基づく損失、並びに、前記画像特徴量及び前記生成テキストデータに基づく損失に基づいて、前記テキスト生成確率パラメータを更新する処理を

含む、付記項 4 に記載の画像処理装置。

[0127] 〔付記項 7〕

前記パラメータ更新ステップは、同じクラスの画像データに関する前記付随テキストデータのテキスト特徴量及び前記生成テキストデータのテキスト特徴量が近づくように前記テキスト生成確率パラメータを更新すると共に、異なるクラスの画像データに関する前記付随テキストデータのテキスト特徴量及び前記生成テキストデータのテキスト特徴量が遠ざかるように前記テキスト生成確率パラメータを更新する処理を含む、付記項 4 に記載の画像処理装置。

[0128] 〔付記項 8〕

前記テキスト生成ステップは、各時刻で上位の所定数のトークンからのランダムサンプリングによる前記生成テキストデータの生成と、通常時の前記生成テキストデータの生成とを併用する処理を含む、付記項 4 に記載の画像処理装置。

[0129] 〔付記項 9〕

前記テキスト理解ステップは、前記生成テキストデータに係る所定番目のトークンの確からしさを示す確率を用いた重み付きプーリングを行うことで、前記テキスト特徴量を抽出する処理を含む、
付記項 4 に記載の画像処理装置。

[0130] 〔付記項 10〕

画像データの特徴量を抽出する画像処理装置が実行する画像処理方法であって、

前記画像処理装置は、

前記画像データの画像パターンをベクトル化して画像特徴量を抽出する画像理解ステップと、

前記画像データに付随している付随テキストデータのテキストパターンをベクトル化してテキスト特徴量を抽出するテキスト理解ステップと、

前記画像理解ステップによって抽出された前記画像特徴量と前記テキスト

理解ステップによって抽出された前記テキスト特徴量を同じベクトル空間に射影して、前記画像特徴量と前記テキスト特徴量を混合することで、前記特徴量としての混合特徴量を生成する特徴量混合ステップと、

を実行する画像処理方法。

[0131] 〔付記項 1 1〕

コンピュータに、付記項 1 0 に記載の方法を実行させるプログラムを記録した非一時的記録媒体。

[0132] 〔その他〕

本特許出願は 2 0 2 1 年 1 1 月 1 2 日に出願した国際出願 P C T / J P 2 0 2 1 / 0 4 1 8 0 1 に基づきその優先権を主張するものであり、国際出願 P C T / J P 2 0 2 1 / 0 4 1 8 0 1 の全内容を本願に援用する。

符号の説明

[0133] 1 通信システム

3 画像分類装置（画像処理装置の一例）

5 通信端末

3 0 入力部

3 1 読出部

3 2 選択部

3 3 特徴抽出部

3 4 類似度計算部

3 5 損失計算部

3 6 パラメータ更新部

3 9 出力部

4 1 画像理解部

4 2 テキスト生成部

4 3 テキスト理解部

4 4 特徴量混合部

4 2 2 デコーダ

請求の範囲

- [請求項1] 画像データの特徴量を抽出する画像処理装置であって、
前記画像データの画像パターンをベクトル化して画像特徴量を抽出する画像理解部と、
前記画像データに付随している付随テキストデータのテキストパターンをベクトル化してテキスト特徴量を抽出するテキスト理解部と、
前記画像理解部によって抽出された前記画像特徴量と前記テキスト理解部によって抽出された前記テキスト特徴量を同じベクトル空間に射影して、前記画像特徴量と前記テキスト特徴量を混合することで、前記特徴量としての混合特徴量を生成する特徴量混合部と、
を有する画像処理装置。
- [請求項2] 前記画像理解部、前記テキスト理解部、及び前記特徴量混合部は、それぞれニューラルネットワークで構成され、前記画像理解部、前記テキスト理解部、及び前記特徴量混合部は前記ニューラルネットワークのモデルパラメータに基づいて処理を行う、請求項1に記載の画像処理装置。
- [請求項3] 請求項2に記載の画像処理装置であって、
前記特徴量混合部によって生成された第1の画像データに係る第1の混合特徴量、及び前記特徴量混合部によって生成された第2の画像データに係る第2の混合特徴量の画像類似度を計算する類似度計算部と、
前記類似度計算部によって計算された前記画像類似度に基づいて、前記モデルパラメータに含まれる画像類似度パラメータを更新するパラメータ更新部と、
を有する画像処理装置。
- [請求項4] 請求項2に記載の画像処理装置であって、
前記画像理解部によって抽出された前記画像特徴量を前記付随テキストデータのベクトル空間に射影することで生成テキストデータを生

成するテキスト生成部を有し、

前記テキスト生成部によって生成された前記生成テキストデータと前記付随テキストデータに基づいて、前記モデルパラメータに含まれるテキスト生成確率パラメータを更新するパラメータ更新部と、
を有する画像処理装置。

[請求項5] 請求項1に記載の画像処理装置であって、

前記画像理解部によって抽出された前記画像特徴量を前記付随テキストデータのベクトル空間に射影することで生成テキストデータを生成するテキスト生成部を有し、

前記画像データに前記付随テキストデータが付随していない場合には、前記テキスト理解部は、前記テキスト生成部によって生成された前記生成テキストデータを前記付随テキストデータとすることで、前記テキスト特徴量を抽出する画像処理装置。

[請求項6] 前記パラメータ更新部は、前記画像特徴量及び前記付随テキストデータに基づく損失、並びに、前記画像特徴量及び前記生成テキストデータに基づく損失に基づいて、前記テキスト生成確率パラメータを更新する、請求項4に記載の画像処理装置。

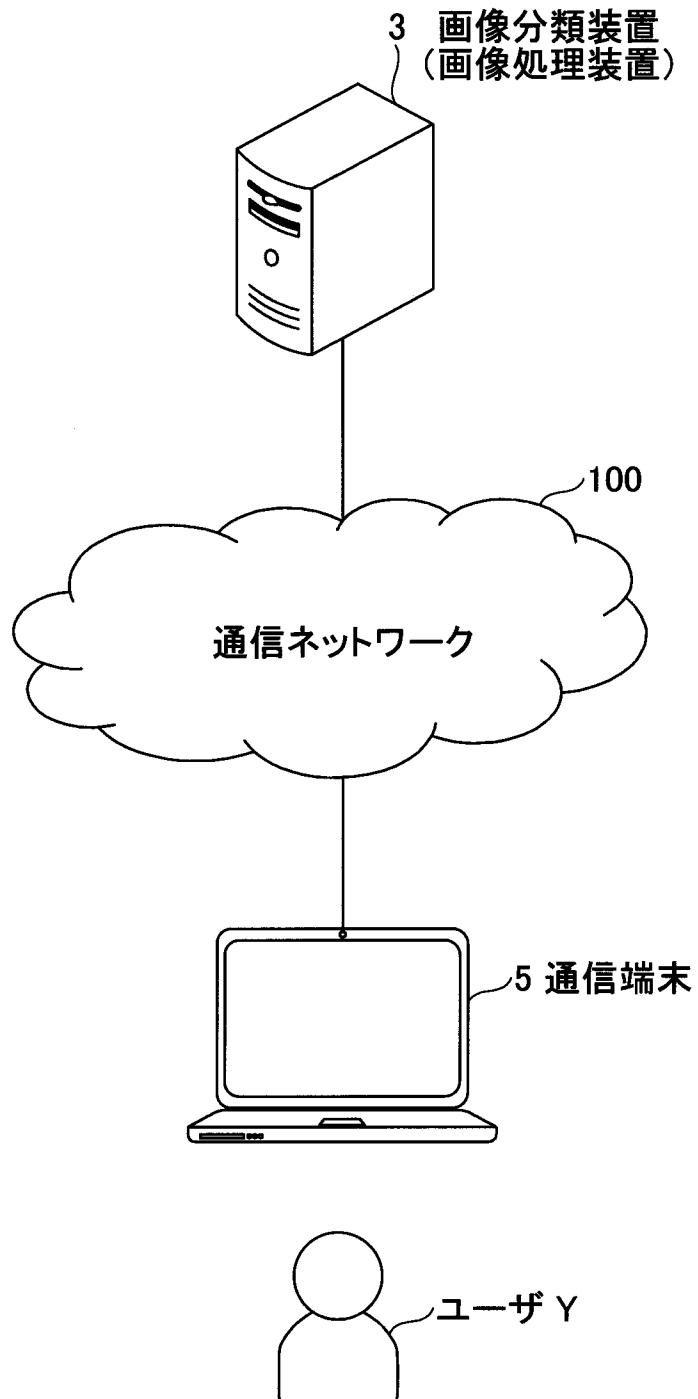
[請求項7] 前記パラメータ更新部は、同じクラスの画像データに関する前記付随テキストデータのテキスト特徴量及び前記生成テキストデータのテキスト特徴量が近づくように前記テキスト生成確率パラメータを更新すると共に、異なるクラスの画像データに関する前記付随テキストデータのテキスト特徴量及び前記生成テキストデータのテキスト特徴量が遠ざかるように前記テキスト生成確率パラメータを更新する、請求項4に記載の画像処理装置。

[請求項8] 前記テキスト生成部は、各時刻で上位の所定数のトークンからのランダムサンプリングによる前記生成テキストデータの生成と、通常時の前記生成テキストデータの生成とを併用する、請求項4に記載の画像処理装置。

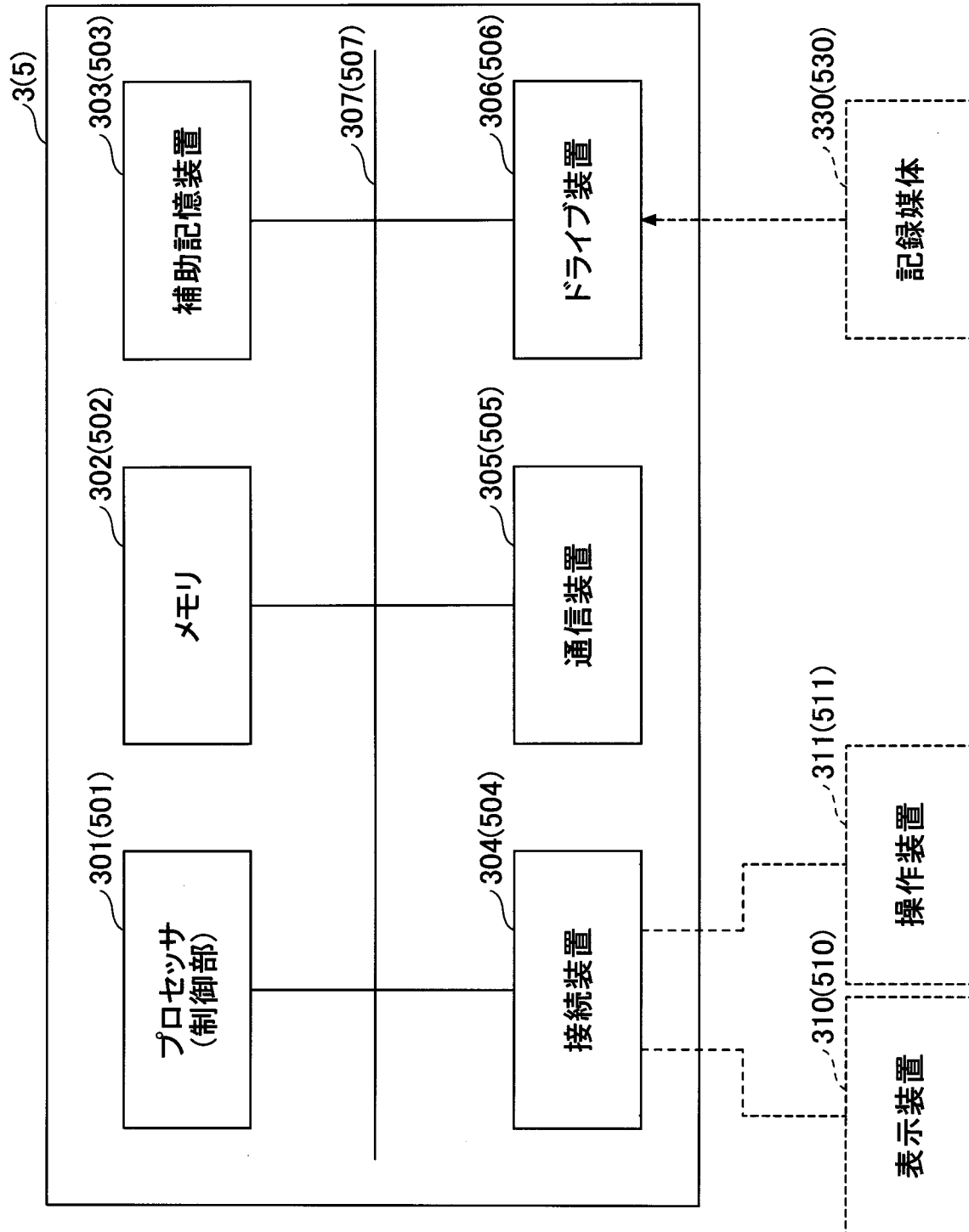
- [請求項9] 前記テキスト理解部は、前記生成テキストデータに係る所定番目のトークンの確からしさを示す確率を用いた重み付きプーリングを行うことで、前記テキスト特徴量を抽出する、
請求項4に記載の画像処理装置。
- [請求項10] 請求項2乃至9のいずれか一項に記載の画像処理装置と、
通信ネットワークを介して前記画像処理装置に前記画像データを送信し、前記通信ネットワークを介して前記画像処理装置から前画像類似度に基づく画像の分類結果データを受信する通信端末と、
を有する通信システム。
- [請求項11] 画像データの特徴量を抽出する画像処理装置が実行する画像処理方法であって、
前記画像処理装置は、
前記画像データの画像パターンをベクトル化して画像特徴量を抽出する画像理解ステップと、
前記画像データに付随している付随テキストデータのテキストパターンをベクトル化してテキスト特徴量を抽出するテキスト理解ステップと、
前記画像理解ステップによって抽出された前記画像特徴量と前記テキスト理解ステップによって抽出された前記テキスト特徴量を同じベクトル空間に射影して、前記画像特徴量と前記テキスト特徴量を混合することで、前記特徴量としての混合特徴量を生成する特徴量混合ステップと、
を実行する画像処理方法。
- [請求項12] コンピュータに、請求項11に記載の方法を実行させるプログラム。

[図1]

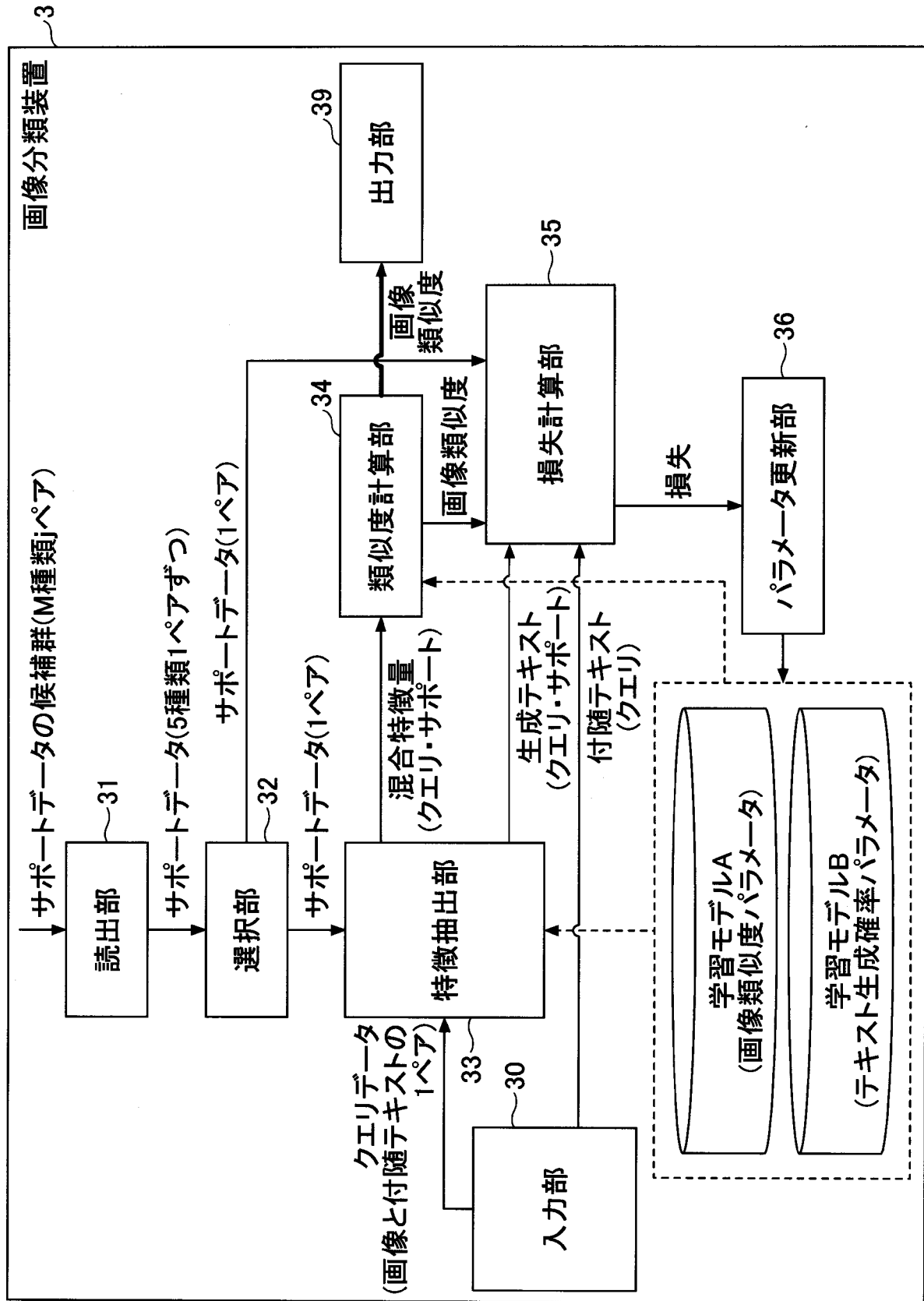
通信システム 1



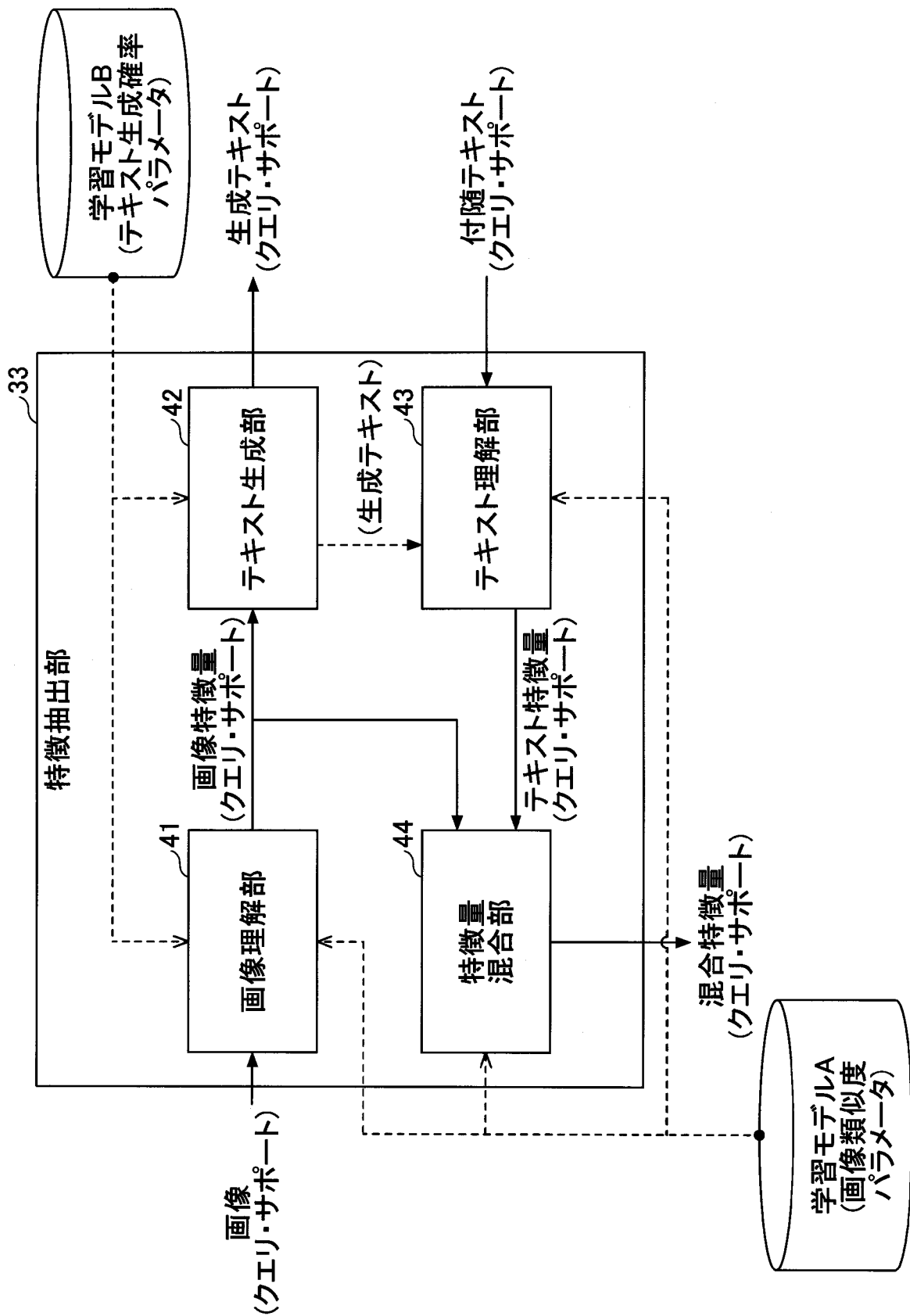
[図2]



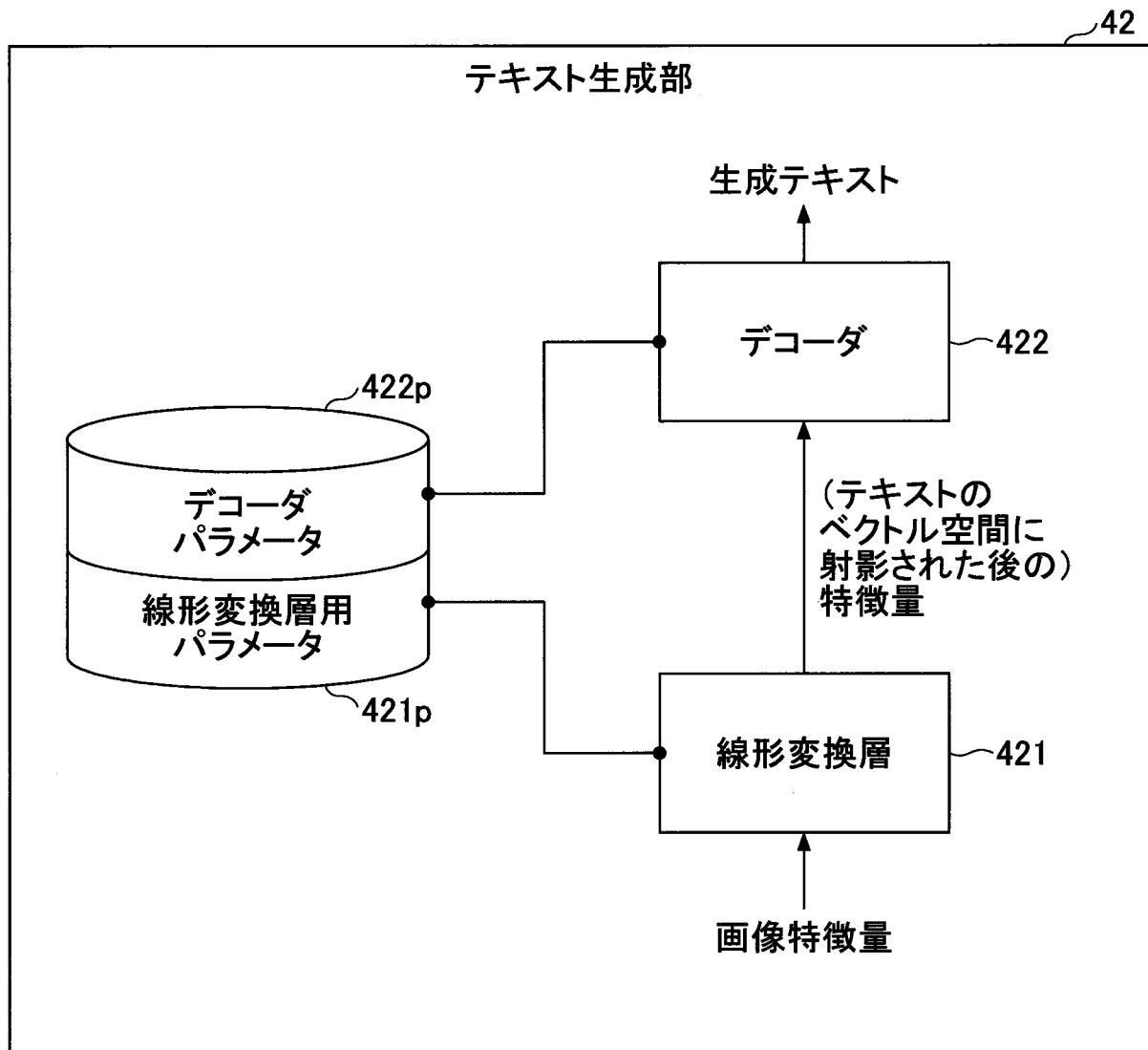
[図3]



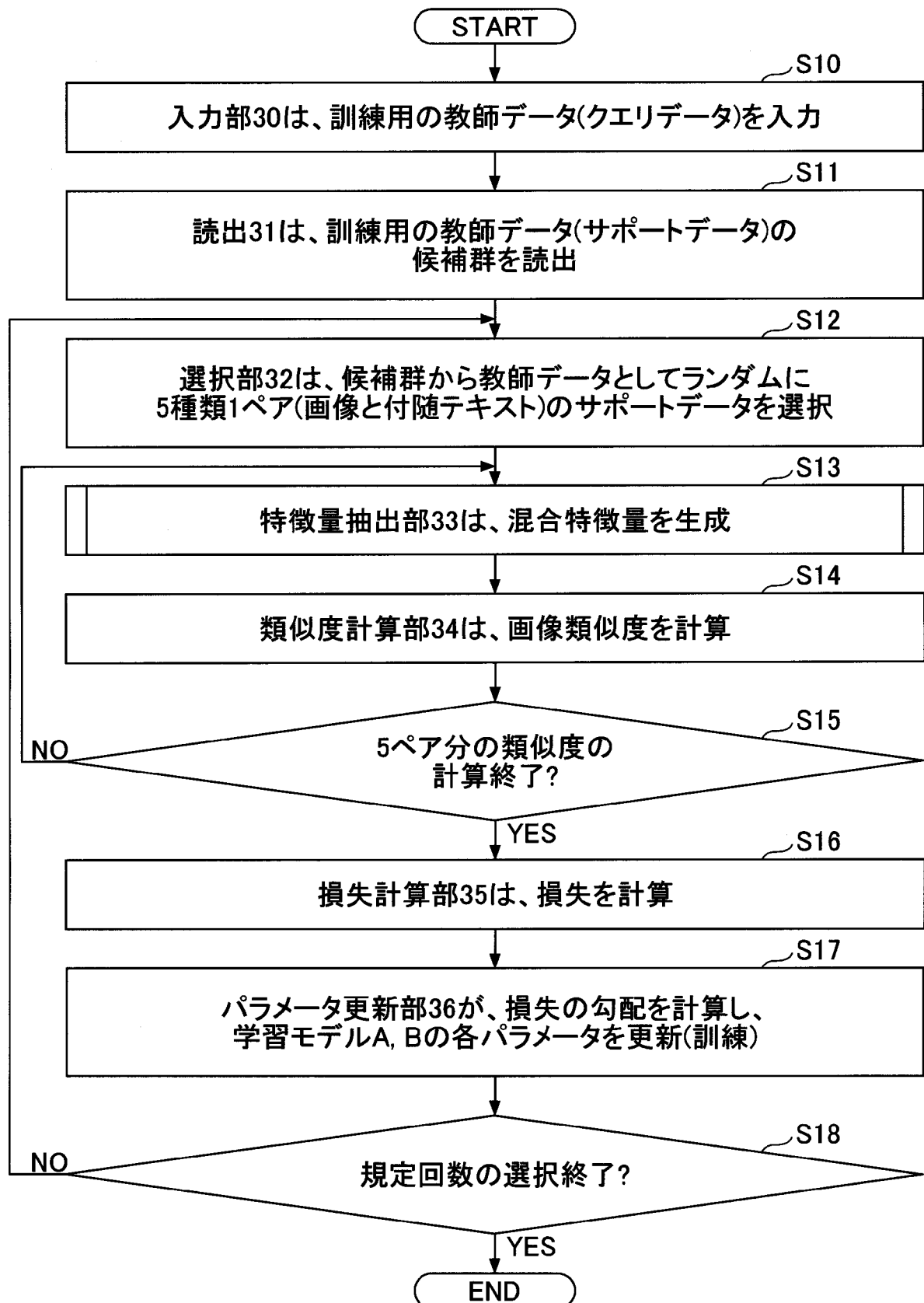
[図4]



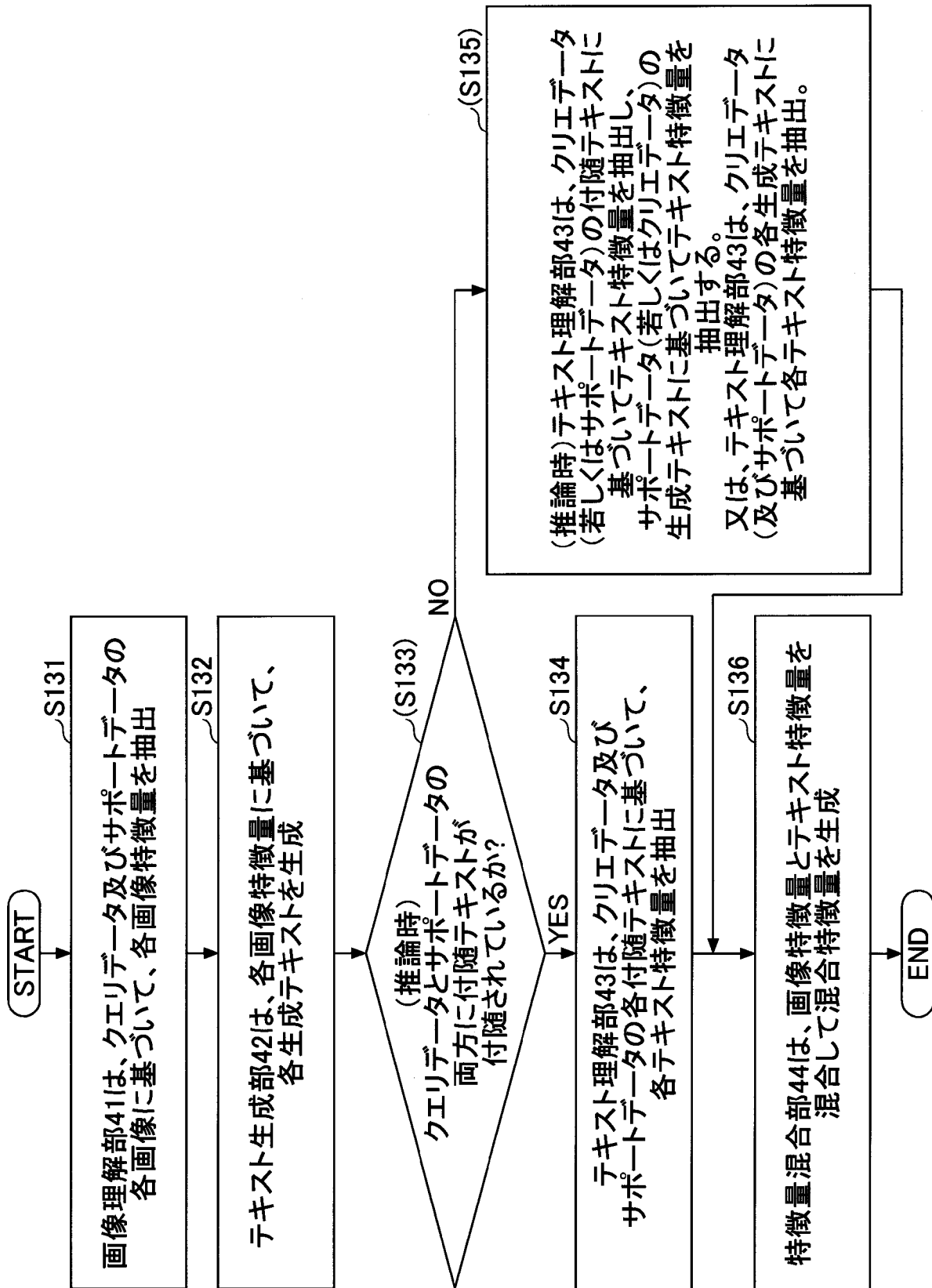
[図5]



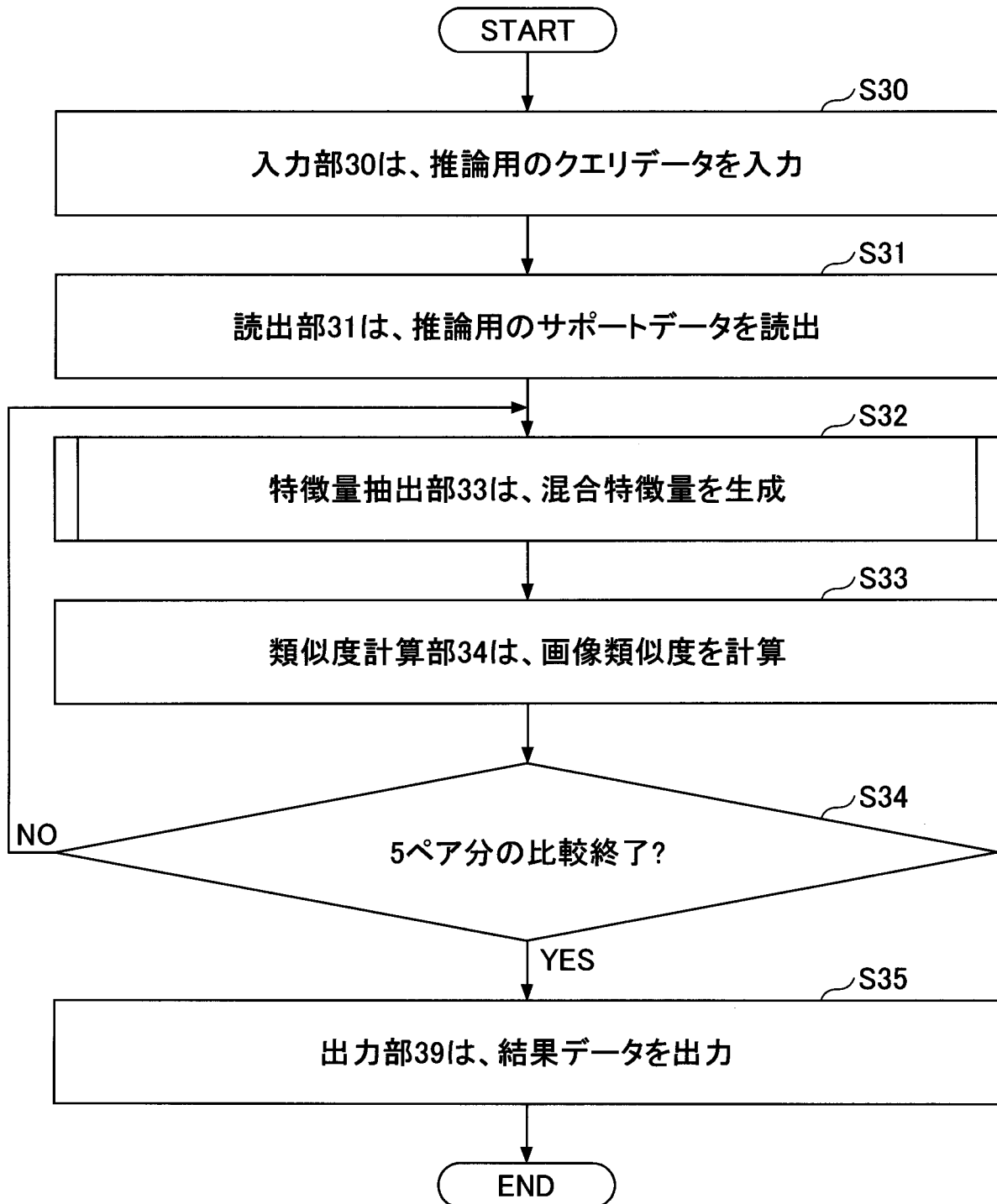
[図6]



[図7]



[図8]



[図9]

LIDE	67.53 ± 0.91
No $L_{class,gen}$	61.96 ± 0.90
No Contrastive Loss	64.60 ± 0.88
No Weighted Average Pooling	66.16 ± 0.93
No Random Sampling	66.40 ± 0.93
None of Above	59.68 ± 0.98

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2022/025412

A. CLASSIFICATION OF SUBJECT MATTER**G06T 7/00**(2017.01)i; **G06F 16/56**(2019.01)i; **G06F 16/583**(2019.01)i

FI: G06T7/00 300F; G06F16/56; G06F16/583

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06T7/00; G06F16/56; G06F16/583

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996
 Published unexamined utility model applications of Japan 1971-2022
 Registered utility model specifications of Japan 1996-2022
 Published registered utility model applications of Japan 1994-2022

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP 2020-52463 A (MACROMILL INC.) 02 April 2020 (2020-04-02) paragraphs [0017], [0037]-[0044], [0051]	1-2, 10-12
Y		3-9
Y	渡辺靖彦, 外 1 名. 画像の内容を説明するテキストを利用した画像解析. 人工知能学会誌. 01 January 1998, vol. 13, no. 1, pp. 66-74, (WATANABE, Yasuhiko, NAGAO, Makoto. Image Analysis Using Natural Language Information Extracted from Explanation Text. Journal of the Japanese Society for Artificial Intelligence.) p. 66, left column to p. 69, left column	3-9
A	US 2020/0311467 A1 (MICROSOFT TECHNOLOGY LICENSING, LLC) 01 October 2020 (2020-10-01) entire text, all drawings	1-11



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance
 “E” earlier application or patent but published on or after the international filing date
 “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
 “O” document referring to an oral disclosure, use, exhibition or other means
 “P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
 “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
 “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
 “&” document member of the same patent family

Date of the actual completion of the international search

16 August 2022

Date of mailing of the international search report

30 August 2022

Name and mailing address of the ISA/JP

Japan Patent Office (ISA/JP)
3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915
Japan

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/JP2022/025412

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
JP	2020-52463	A	02 April 2020	(Family: none)	
US	2020/0311467	A1	01 October 2020	(Family: none)	

A. 発明の属する分野の分類（国際特許分類（IPC）） G06T 7/00(2017.01)i; G06F 16/56(2019.01)i; G06F 16/583(2019.01)i FI: G06T7/00 300F; G06F16/56; G06F16/583										
B. 調査を行った分野										
調査を行った最小限資料（国際特許分類（IPC）） G06T7/00; G06F16/56; G06F16/583										
最小限資料以外の資料で調査を行った分野に含まれるもの										
<table border="0"> <tr> <td>日本国実用新案公報</td> <td>1922-1996年</td> </tr> <tr> <td>日本国公開実用新案公報</td> <td>1971-2022年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996-2022年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994-2022年</td> </tr> </table>			日本国実用新案公報	1922-1996年	日本国公開実用新案公報	1971-2022年	日本国実用新案登録公報	1996-2022年	日本国登録実用新案公報	1994-2022年
日本国実用新案公報	1922-1996年									
日本国公開実用新案公報	1971-2022年									
日本国実用新案登録公報	1996-2022年									
日本国登録実用新案公報	1994-2022年									
国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）										
C. 関連すると認められる文献										
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号								
X	JP 2020-52463 A (株式会社マクロミル) 02.04.2020 (2020-04-02) [0017], [0037]-[0044], [0051]	1-2, 10-12								
Y		3-9								
Y	渡辺 靖彦, 外1名, 画像の内容を説明するテキストを利用した画像解析, 人工知能 学会誌, 1998.01.01, 第13巻, 第1号, pp. 66-74 p. 66 左欄 ~ p. 69 左欄	3-9								
A	US 2020/0311467 A1 (MICROSOFT TECHNOLOGY LICENSING, LLC) 01.10.2020 (2020-10-01) 全文全文	1-11								
☐ C欄の続きにも文献が列挙されている。 ☒ パテントファミリーに関する別紙を参照。										
<table border="0"> <tr> <td> * 引用文献のカテゴリー “A” 特に関連のある文献ではなく、一般的技术水準を示すもの “E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの “L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す） “O” 口頭による開示、使用、展示等に言及する文献 “P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献 </td> <td> “T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの “X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの “Y” 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの “&” 同一パテントファミリー文献 </td> </tr> </table>			* 引用文献のカテゴリー “A” 特に関連のある文献ではなく、一般的技术水準を示すもの “E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの “L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す） “O” 口頭による開示、使用、展示等に言及する文献 “P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献	“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの “X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの “Y” 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの “&” 同一パテントファミリー文献						
* 引用文献のカテゴリー “A” 特に関連のある文献ではなく、一般的技术水準を示すもの “E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの “L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す） “O” 口頭による開示、使用、展示等に言及する文献 “P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献	“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの “X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの “Y” 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの “&” 同一パテントファミリー文献									
国際調査を完了した日 16.08.2022	国際調査報告の発送日 30.08.2022									
名称及びあて先 日本国特許庁(ISA/JP) 〒100-8915 日本国 東京都千代田区霞が関三丁目4番3号	権限のある職員（特許庁審査官） 笠田 和宏 5H 5285 電話番号 03-3581-1101 内線 3531									

国際調査報告
パテントファミリーに関する情報

国際出願番号

PCT/JP2022/025412

引用文献			公表日	パテントファミリー文献	公表日
JP	2020-52463	A	02.04.2020	(ファミリーなし)	
US	2020/0311467	A1	01.10.2020	(ファミリーなし)	