(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
2021年5月20日(20.05.2021)

WIPO | PCT

(10) 国際公開番号
WO 2021/095212 A1

(51) 国際特許分類:
G06N 20/00 (2019.01)

(21) 国際出願番号： PCT/JP2019/044770

(22) 国際出願日： 2019年11月14日(14.11.2019)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(71) 出願人: 富士通株式会社(**FUJITSU LIMITED**) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 Kanagawa (JP).

(72) 発明者: 山田 萌(**YAMADA, Moyuru**); 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 富士通株式会社内 Kanagawa (JP).

(74) 代理人: 酒井 昭徳(**SAKAI, Akinori**); 〒1020094 東京都千代田区紀尾井町３番１２号 紀尾井町ビル７階 酒井総合特許事務所 Tokyo (JP).

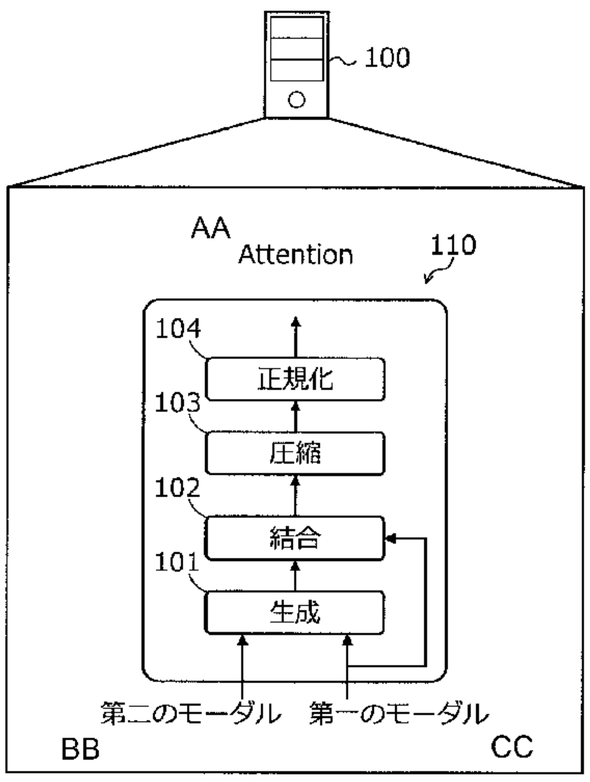
(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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).

(54) Title: OUTPUT METHOD, OUTPUT PROGRAM, AND OUTPUT DEVICE

(54) 発明の名称： 出力方法、出力プログラム、および出力装置

[図1]



101 Generation
102 Coupling
103 Compression
104 Normalization
AA Attention
BB Second modal
CC First modal

(57) Abstract: An output device (100) generates, by using a generation model (101), a correction vector for correcting a vector based on first modal information, the correction vector being generated on the basis of a correlation between a vector based on the first modal information and a vector based on second modal information. The output device (100) couples, by using a coupling model (102), the generated correction vector to the vector based on the first modal information. The output device (100) compresses, by using a compression model (103), the vector based on the first modal information after coupling, in accordance with a prescribed rule. The output device (100) carries out, by using a normalization model (104), a normalization process on the vector based on the first modal information after compression. The output device (100) outputs the vector obtained through the normalization process.

[続葉有]

WO 2021/095212 A1

添付公開書類：

一　国際調査報告（条約第21条(3)）

(57) 要約：出力装置（１００）は、生成モデル（１０１）を用いて、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとの相関に基づいて、第一のモーダルの情報に基づくベクトルを補正する補正ベクトルを生成する。出力装置（１００）は、結合モデル（１０２）を用いて、生成した補正ベクトルを、第一のモーダルの情報に基づくベクトルに結合する。出力装置（１００）は、圧縮モデル（１０３）を用いて、所定のルールに従って、結合後の第一のモーダルの情報に基づくベクトルを圧縮する。出力装置（１００）は、正規化モデル（１０４）を用いて、圧縮後の第一のモーダルの情報に基づくベクトルに対して、正規化処理を実施する。出力装置（１００）は、正規化処理により得たベクトルを出力する。

明　細　書

発明の名称 ： 出力方法、出力プログラム、および出力装置

技術分野

[0001] 本発明は、出力方法、出力プログラム、および出力装置に関する。

背景技術

[0002] 従来、複数のモーダルの情報を用いて問題を解く技術がある。この技術は、例えば、文書翻訳や質疑応答、物体検出、状況判断などの問題を解く際に利用される。ここで、モーダルとは、情報の様式や種類を示す概念であり、具体例としては、画像、文書（テキスト）、音声などを挙げることができる。複数のモーダルを用いた機械学習はマルチモーダル学習と呼ばれる。

[0003] 先行技術としては、例えば、Ａｔｔｅｎｔｉｏｎにより情報を変換するＴｒａｎｓｆｏｒｍｅｒと呼ばれるものがある。Ａｔｔｅｎｔｉｏｎは、具体的には、第一のモーダルの情報に基づくベクトルから得たクエリと、第二のモーダルの情報に基づくベクトルから得たキーとの相関に基づいて、第二のモーダルの情報に基づくベクトルから得たバリューの重み付け和を算出し、第一のモーダルの情報に基づくベクトルに加算する。

先行技術文献

非特許文献

[0004] 非特許文献1：Ｖａｓｗａｎｉ，Ａｓｈｉｓｈ，ｅｔ ａｌ．“Ａｔｔｅｎｔｉｏｎ ｉｓ ａｌｌ ｙｏｕ ｎｅｅｄ．” Ａｄｖａｎｃｅｓ ｉｎ ｎｅｕｒａｌ ｉｎｆｏｒｍａｔｉｏｎ ｐｒｏｃｅｓｓｉｎｇ ｓｙｓｔｅｍｓ．２０１７．

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、複数のモーダルの情報を用いて問題を解いた際の解の精度が悪い場合がある。例えば、画像と文書とを基に状況を判断する問題を解くにあたり、Ａｔｔｅｎｔｉｏｎにより、画像に関するモーダ

ルの情報に基づくベクトルに、文書に関するモーダルの情報に基づくベクトルから得たバリューの重み付け和を、単純に加算すると、問題の解決に有用な情報が失われやすい。このため、問題を解いた際の解の精度が悪くなりやすい。

[0006] １つの側面では、本発明は、複数のモーダルの情報を用いて問題を解いた際の解の精度の向上を図ることを目的とする。

課題を解決するための手段

[0007] １つの実施態様によれば、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとの相関に基づいて、前記第一のモーダルの情報に基づくベクトルを補正する補正ベクトルを生成し、生成した前記補正ベクトルを、前記第一のモーダルの情報に基づくベクトルに結合し、所定のルールに従って、結合後の前記第一のモーダルの情報に基づくベクトルを圧縮し、圧縮後の前記第一のモーダルの情報に基づくベクトルに対して、正規化処理を実施し、前記正規化処理により得たベクトルを出力する出力方法、出力プログラム、および出力装置が提案される。

発明の効果

[0008] 一態様によれば、複数のモーダルの情報を用いて問題を解いた際の解の精度の向上を図ることが可能になる。

図面の簡単な説明

[0009] [図1]図１は、実施の形態にかかる出力方法の一実施例を示す説明図である。

[図2]図２は、情報処理システム２００の一例を示す説明図である。

[図3]図３は、出力装置１００のハードウェア構成例を示すブロック図である。

[図4]図４は、出力装置１００の機能的構成例を示すブロック図である。

[図5]図５は、Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋ５００の具体例を示す説明図である。

[図6]図６は、ＳＡ層６００の具体例と、ＴＡ層６１０の具体例とを示す説明図である。

[図7]図７は、画像ＴＡ層５０１の具体例を示す説明図である。

[図8]図８は、画像ＴＡ層５０１の別の具体例を示す説明図である。

[図9]図９は、画像ＴＡ層５０１と文書ＴＡ層５０３との比較例を示す説明図である。

[図10]図１０は、ＣＡＮ５００を用いた動作の一例を示す説明図である。

[図11]図１１は、出力装置１００の利用例１を示す説明図（その１）である。

[図12]図１２は、出力装置１００の利用例１を示す説明図（その２）である。

[図13]図１３は、出力装置１００の利用例２を示す説明図（その１）である。

[図14]図１４は、出力装置１００の利用例２を示す説明図（その２）である。

[図15]図１５は、学習処理手順の一例を示すフローチャートである。

[図16]図１６は、推定処理手順の一例を示すフローチャートである。

[図17]図１７は、アテンション処理手順の一例を示すフローチャートである。

発明を実施するための形態

[0010] 以下に、図面を参照して、本発明にかかる出力方法、出力プログラム、および出力装置の実施の形態を詳細に説明する。

[0011] （実施の形態にかかる出力方法の一実施例）

図１は、実施の形態にかかる出力方法の一実施例を示す説明図である。出力装置１００は、複数のモーダルの情報を用いて、問題の解決に有用な情報を得やすくすることにより、問題を解いた際の解の精度の向上を図るためのコンピュータである。

[0012] 従来、問題を解くための手法として、例えば、Ａｔｔｅｎｔｉｏｎにより情報を変換するＴｒａｎｓｆｏｒｍｅｒを利用した、ＢＥＲＴ（Ｂｉｄｉｒｅｃｔｉｏｎａｌ Ｅｎｃｏｄｅｒ Ｒｅｐｒｅｓｅｎｔａｔｉｏｎｓ ｆ

ｒｏｍ　Ｔｒａｎｓｆｏｒｍｅｒｓ）と呼ばれるものがある。ＢＥＲＴは、具体的には、ＴｒａｎｓｆｏｒｍｅｒのＥｎｃｏｄｅｒ部を積み重ねて形成される。ＢＥＲＴについては、例えば、下記非特許文献２を参照することができる。

[0013] 非特許文献２：Ｄｅｖｌｉｎ，　Ｊａｃｏｂ　ｅｔ　ａｌ．“ＢＥＲＴ：Ｐｒｅ－ｔｒａｉｎｉｎｇ　ｏｆ　Ｄｅｅｐ　Ｂｉｄｉｒｅｃｔｉｏｎａｌ　Ｔｒａｎｓｆｏｒｍｅｒｓ　ｆｏｒ　Ｌａｎｇｕａｇｅ　Ｕｎｄｅｒｓｔａｎｄｉｎｇ．”　ＮＡＡＣＬ－ＨＬＴ　（２０１９）．

[0014] ここで、ＢＥＲＴは、文書に関するモーダルの情報を用いて問題を解くような状況に適用することが想定されており、複数のモーダルの情報を用いて問題を解くような状況に適用することができない。

[0015] これに対し、例えば、ＶｉｄｅｏＢＥＲＴと呼ばれる手法がある。ＶｉｄｅｏＢＥＲＴは、具体的には、ＢＥＲＴを、文書に関するモーダルの情報と、画像に関するモーダルの情報とを用いて問題を解くような状況に適用可能に拡張したものである。ＶｉｄｅｏＢＥＲＴについては、例えば、下記非特許文献３を参照することができる。

[0016] 非特許文献３：Ｓｕｎ，　Ｃｈｅｎ，　ｅｔ　ａｌ．“Ｖｉｄｅｏｂｅｒｔ：Ａ　ｊｏｉｎｔ　ｍｏｄｅｌ　ｆｏｒ　ｖｉｄｅｏ　ａｎｄ　ｌａｎｇｕａｇｅ　ｒｅｐｒｅｓｅｎｔａｔｉｏｎ　ｌｅａｒｎｉｎｇ．”　ａｒＸｉｖ　ｐｒｅｐｒｉｎｔ　ａｒＸｉｖ：１９０４．０１７６６　（２０１９）．

[0017] また、例えば、ＭＣＡＮ（Ｍｏｄｕｌａｒ　Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋ）と呼ばれる手法がある。ＭＣＡＮは、文書に関するモーダルの情報に基づくベクトルと、文書に関するモーダルの情報に基づくベクトルを基に補正した、画像に関するモーダルの情報に基づくベクトルとを参照し、問題を解くものである。ＭＣＡＮについては、例えば、下記非特許文献４を参照することができる。

[0018] 非特許文献４：Ｙｕ，　Ｚｈｏｕ，　ｅｔ　ａｌ．“Ｄｅｅｐ　Ｍ

odular Co-Attention Networks for Visual Question Answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[0019] また、例えば、ＶｉＬＢＥＲＴ（Ｖｉｓｉｏｎ－ａｎｄ－Ｌａｎｇｕａｇｅ Ｂｉｄｉｒｅｃｔｉｏｎａｌ Ｅｎｃｏｄｅｒ Ｒｅｐｒｅｓｅｎｔａｔｉｏｎｓ ｆｒｏｍ Ｔｒａｎｓｆｏｒｍｅｒｓ）と呼ばれる手法がある。ＶｉＬＢＥＲＴは、画像に関するモーダルの情報に基づくベクトルを基に補正した、文書に関するモーダルの情報に基づくベクトルと、文書に関するモーダルの情報に基づくベクトルを基に補正した、画像に関するモーダルの情報に基づくベクトルとを参照し、問題を解く技術である。

[0020] 非特許文献５：Ｌｕ，Ｊｉａｓｅｎ，ｅｔ ａｌ．“ｖｉｌｂｅｒｔ：Ｐｒｅｔｒａｉｎｉｎｇ ｔａｓｋ－ａｇｎｏｓｔｉｃ ｖｉｓｉｏｌｉｎｇｕｉｓｔｉｃ ｒｅｐｒｅｓｅｎｔａｔｉｏｎｓ ｆｏｒ ｖｉｓｉｏｎ－ａｎｄ－ｌａｎｇｕａｇｅ ｔａｓｋｓ．” ａｒＸｉｖ ｐｒｅｐｒｉｎｔ ａｒＸｉｖ：１９０８．０２２６５ （２０１９）．

[0021] しかしながら、上述したＶｉｄｅｏＢＥＲＴ、ＭＣＡＮ、およびＶｉＬＢＥＲＴなどの手法でも、複数のモーダルの情報を用いて問題を解いた際の解の精度が悪い場合がある。具体的には、いずれの手法でも、Ａｔｔｅｎｔｉｏｎにより、画像に関するモーダルの情報に基づくベクトルに、文書に関するモーダルの情報に基づくベクトルから得たバリューの重み付け和を、単純に加算するため、問題の解決に有用な情報が失われやすいという性質が存在する。このため、いずれの手法でも、問題を解いた際の解の精度が悪くなりやすい。また、ＶｉｄｅｏＢＥＲＴでは、問題を解くにあたり、文書に関するモーダルの情報と、画像に関するモーダルの情報とを明示的に区別せずに扱うため、問題を解いた際の解の精度が悪い。

[0022] そこで、本実施の形態では、問題を解くにあたり有用なベクトルを生成可

能にすることにより、複数のモーダルの情報を用いて問題を解くような状況に適用可能でありつつ、問題を解いた際の解の精度を向上可能にすることができる出力方法について説明する。

[0023] 図１において、出力装置１００は、例えば、Ａｔｔｅｎｔｉｏｎを実現する変換モデル１１０を有する。変換モデルは、生成モデル１０１と、結合モデル１０２と、圧縮モデル１０３と、正規化モデル１０４とを含む。

[0024] 出力装置１００は、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとを取得する。モーダルは、情報の様式を意味する。第一のモーダルと、第二のモーダルとは、それぞれ異なるモーダルである。第一のモーダルは、例えば、画像に関するモーダルである。第二のモーダルは、例えば、文書に関するモーダルである。

[0025] 第一のモーダルの情報に基づくベクトルは、例えば、第一のモーダルに従って表現されたベクトルである。第一のモーダルの情報に基づくベクトルは、例えば、第一のモーダルの情報に基づいて生成される。第一のモーダルの情報は、例えば、画像である。第一のモーダルの情報に基づくベクトルは、例えば、画像に基づいて生成されたベクトルである。

[0026] 第二のモーダルの情報に基づくベクトルは、例えば、第二のモーダルに従って表現されたベクトルである。第二のモーダルの情報に基づくベクトルは、例えば、第二のモーダルの情報に基づいて生成される。第二のモーダルの情報は、例えば、文書である。第二のモーダルの情報に基づくベクトルは、例えば、文書に基づいて生成されたベクトルである。

[0027] （１－１）出力装置１００は、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとの相関に基づいて、第一のモーダルの情報に基づくベクトルを補正する補正ベクトルを生成する。出力装置１００は、例えば、生成モデル１０１を用いて、第一のモーダルの情報に基づくベクトルを補正する補正ベクトルを生成する。

[0028] 相関は、例えば、第一のモーダルの情報に基づくベクトルから得たベクトルと、第二のモーダルの情報に基づくベクトルから得たベクトルとの類似度

によって表現される。第一のモーダルの情報に基づくベクトルから得たベクトルは、例えば、クエリである。第二のモーダルの情報に基づくベクトルから得たベクトルは、例えば、キーである。類似度は、例えば、内積によって表現される。類似度は、例えば、差分の二乗和などによって表現されてもよい。

[0029] （１－２）出力装置１００は、生成した補正ベクトルを、第一のモーダルの情報に基づくベクトルに結合する。出力装置１００は、例えば、結合モデル１０２を用いて、生成した補正ベクトルを、第一のモーダルの情報に基づくベクトルに結合する。

[0030] （１－３）出力装置１００は、所定のルールに従って、結合後の第一のモーダルの情報に基づくベクトルを圧縮する。出力装置１００は、例えば、圧縮モデル１０３を用いて、結合後の第一のモーダルの情報に基づくベクトルを圧縮する。圧縮は、次元数を低減しない変換を含む。

[0031] （１－４）出力装置１００は、圧縮後の第一のモーダルの情報に基づくベクトルに対して、正規化処理を実施する。出力装置１００は、例えば、正規化モデル１０４を用いて、正規化処理を実施する。正規化処理を実施する具体例については、例えば、図７を用いて後述する。

[0032] （１－５）出力装置１００は、正規化処理により得たベクトルを出力する。出力形式は、例えば、ディスプレイへの表示、プリンタへの印刷出力、他のコンピュータへの送信、または、記憶領域への記憶などである。これにより、出力装置１００は、第一のモーダルの情報に基づくベクトルと第二のモーダルの情報に基づくベクトルとのうち問題の解決に有用な情報が反映される傾向があるベクトルを生成し、利用可能にすることができる。結果として、出力装置１００は、以降の、問題を解いた際の解の精度の向上を図ることができる。

[0033] ここで、例えば、第一のモーダルが画像に関し、第二のモーダルが文書に関する場合、第二のモーダルは、第一のモーダルの上位の階層であるという特徴を有していると考えることができる。具体的には、「りんご（単語）」

は、複数の「りんご（画像）」を包含する概念である。

[0034] 出力装置１００は、この特徴を利用し、画像に関する第一のモーダルの情報に基づくベクトルに、文書に関する第二のモーダルの情報に基づくベクトルに基づく補正ベクトルを結合した上で、圧縮することができる。このため、出力装置１００は、圧縮後のベクトルにおいて、画像と文書とのうち問題の解決に有用な情報が失われ辛く、反映され易くすることができる。出力装置１００は、例えば、実世界の画像や文書の特徴のうち、問題の解決に有用な特徴を、コンピュータ上で効果的に表現した圧縮後のベクトルを利用可能にすることができる。結果として、出力装置１００は、複数のモーダルの情報を用いて問題を解くにあたり、有用なベクトルを得ることができ、問題を解いた際の解の精度を向上可能にすることができる。

[0035] ここでは、第一のモーダルと、第二のモーダルとが、それぞれ異なるモーダルである場合について説明したが、これに限らない。例えば、第一のモーダルと、第二のモーダルとが同一のモーダルである場合があってもよい。

[0036] （情報処理システム２００の一例）

次に、図２を用いて、図１に示した出力装置１００を適用した、情報処理システム２００の一例について説明する。

[0037] 図２は、情報処理システム２００の一例を示す説明図である。図２において、情報処理システム２００は、出力装置１００と、クライアント装置２０１と、端末装置２０２とを含む。

[0038] 情報処理システム２００において、出力装置１００とクライアント装置２０１とは、有線または無線のネットワーク２１０を介して接続される。ネットワーク２１０は、例えば、ＬＡＮ（Ｌｏｃａｌ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）、ＷＡＮ（Ｗｉｄｅ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）、インターネットなどである。また、情報処理システム２００において、出力装置１００と端末装置２０２とは、有線または無線のネットワーク２１０を介して接続される。

[0039] 出力装置１００は、第一のモーダルの情報に基づくベクトルと、第二のモ

ーダルの情報に基づくベクトルとに基づいて、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとを統合した統合ベクトルを生成するＣｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋを有する。第一のモーダルは、例えば、画像に関するモーダルである。第二のモーダルは、例えば、文書に関するモーダルである。Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋは、例えば、図１に示した変換モデル１１０を用いて形成される。

[0040] 出力装置１００は、教師データに基づいて、Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋを更新する。教師データは、例えば、標本となる第一のモーダルの情報に基づくベクトルを生成する元となる第一のモーダルの情報と、標本となる第二のモーダルの情報に基づくベクトルを生成する元となる第二のモーダルの情報と、正解データとを対応付けた対応情報である。教師データは、例えば、出力装置１００のユーザにより出力装置１００に入力される。正解データは、例えば、問題を解いた場合についての正解を示す。例えば、第一のモーダルが、画像に関するモーダルであれば、第一のモーダルの情報は、画像である。例えば、第二のモーダルが、文書に関するモーダルであれば、第二のモーダルの情報は、文書である。

[0041] 出力装置１００は、例えば、第一のモーダルの情報となる教師データの画像から、第一のモーダルの情報に基づくベクトルを生成することにより取得し、第二のモーダルの情報となる教師データの文書から、第二のモーダルの情報に基づくベクトルを生成することにより取得する。そして、出力装置１００は、取得した第一のモーダルの情報に基づくベクトルと、取得した第二のモーダルの情報に基づくベクトルと、教師データの正解データとに基づいて、誤差逆伝搬などにより、Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋを更新する。出力装置１００は、誤差逆伝搬以外の学習方法により、Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋを更新してもよい。

[0042] 出力装置１００は、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとを取得する。そして、出力装置１００は、

Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋを用いて、取得した第一のモーダルの情報に基づくベクトルと、取得した第二のモーダルの情報に基づくベクトルとに基づいて、統合ベクトルを生成し、生成した統合ベクトルに基づいて、問題を解く。その後、出力装置１００は、問題を解いた結果を、クライアント装置２０１に送信する。

[0043] 出力装置１００は、例えば、出力装置１００のユーザにより出力装置１００に入力された第一のモーダルの情報に基づくベクトルを取得する。また、出力装置１００は、第一のモーダルの情報に基づくベクトルを、クライアント装置２０１または端末装置２０２から受信することにより取得してもよい。また、出力装置１００は、例えば、第一のモーダルの情報を、クライアント装置２０１または端末装置２０２から受信し、受信した第一のモーダルの情報から、第一のモーダルの情報に基づくベクトルを生成することにより取得してもよい。

[0044] 出力装置１００は、例えば、出力装置１００のユーザにより出力装置１００に入力された第二のモーダルの情報に基づくベクトルを取得する。また、出力装置１００は、第二のモーダルの情報に基づくベクトルを、クライアント装置２０１または端末装置２０２から受信することにより取得してもよい。また、出力装置１００は、例えば、第二のモーダルの情報を、クライアント装置２０１または端末装置２０２から受信し、受信した第二のモーダルの情報から、第二のモーダルの情報に基づくベクトルを生成することにより取得してもよい。

[0045] そして、出力装置１００は、Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋを用いて、取得した第一のモーダルの情報に基づくベクトルと、取得した第二のモーダルの情報に基づくベクトルとに基づいて、統合ベクトルを生成し、生成した統合ベクトルに基づいて、問題を解く。その後、出力装置１００は、問題を解いた結果を、クライアント装置２０１に送信する。出力装置１００は、例えば、サーバやＰＣ（Ｐｅｒｓｏｎａｌ　Ｃｏｍｐｕｔｅｒ）などである。

[0046] クライアント装置２０１は、出力装置１００と通信可能なコンピュータである。クライアント装置２０１は、例えば、第一のモーダルの情報に基づくベクトルを、出力装置１００に送信してもよい。また、クライアント装置２０１は、例えば、第一のモーダルの情報を、出力装置１００に送信してもよい。クライアント装置２０１は、例えば、第二のモーダルの情報に基づくベクトルを、出力装置１００に送信してもよい。クライアント装置２０１は、例えば、第二のモーダルの情報を、出力装置１００に送信してもよい。

[0047] クライアント装置２０１は、出力装置１００が問題を解いた結果を受信して出力する。出力形式は、例えば、ディスプレイへの表示、プリンタへの印刷出力、他のコンピュータへの送信、または、記憶領域への記憶などである。クライアント装置２０１は、例えば、ＰＣ、タブレット端末、またはスマートフォンなどである。

[0048] 端末装置２０２は、出力装置１００と通信可能なコンピュータである。端末装置２０２は、例えば、第一のモーダルの情報に基づくベクトルを、出力装置１００に送信してもよい。また、端末装置２０２は、例えば、第一のモーダルの情報を、出力装置１００に送信してもよい。端末装置２０２は、例えば、第二のモーダルの情報に基づくベクトルを、出力装置１００に送信してもよい。また、端末装置２０２は、例えば、第二のモーダルの情報を、出力装置１００に送信してもよい。端末装置２０２は、例えば、ＰＣ、タブレット端末、スマートフォン、電子機器、ＩｏＴ機器、またはセンサ装置などである。端末装置２０２は、具体的には、監視カメラであってもよい。

[0049] ここでは、出力装置１００が、Ｃｏ－Ａｔｔｅｎｔｉｏｎ Ｎｅｔｗｏｒｋを更新し、かつ、Ｃｏ－Ａｔｔｅｎｔｉｏｎ Ｎｅｔｗｏｒｋを用いて、問題を解く場合について説明したが、これに限らない。例えば、他のコンピュータが、Ｃｏ－Ａｔｔｅｎｔｉｏｎ Ｎｅｔｗｏｒｋを更新し、出力装置１００が、他のコンピュータから受信したＣｏ－Ａｔｔｅｎｔｉｏｎ Ｎｅｔｗｏｒｋを用いて、問題を解く場合があってもよい。また、例えば、出力

装置１００が、Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋを更新し、他のコンピュータに提供し、他のコンピュータで、Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋを用いて、問題を解く場合があってもよい。

[0050] ここでは、教師データが、第一のモーダルの情報に基づくベクトルを生成する元となる第一のモーダルの情報と、第二のモーダルの情報に基づくベクトルを生成する元となる第二のモーダルの情報と、正解データとを対応付けた対応情報である場合について説明したが、これに限らない。例えば、教師データが、標本となる第一のモーダルの情報に基づくベクトルと、標本となる第二のモーダルの情報に基づくベクトルと、正解データとを対応付けた対応情報である場合があってもよい。

[0051] ここでは、出力装置１００が、クライアント装置２０１や端末装置２０２とは異なる装置である場合について説明したが、これに限らない。例えば、出力装置１００が、クライアント装置２０１と一体である場合があってもよい。また、例えば、出力装置１００が、端末装置２０２と一体である場合があってもよい。

[0052] ここでは、出力装置１００が、ソフトウェア的に、Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋを実現する場合について説明したが、これに限らない。例えば、出力装置１００が、Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋを、電子回路的に実現する場合があってもよい。

[0053]（情報処理システム２００の適用例１）

適用例１において、出力装置１００は、画像と、画像についての質問文となる文書とを記憶する。質問文は、例えば、「画像内で何を切っているか」である。そして、出力装置１００は、画像と文書とに基づいて、質問文に対する回答文を推定する問題を解く。出力装置１００は、例えば、画像と文書とに基づいて、画像内で何を切っているかの質問文に対する回答文を推定し、クライアント装置２０１に送信する。

[0054]（情報処理システム２００の適用例２）

適用例２において、端末装置２０２は、監視カメラであり、対象を撮像し

た画像を、出力装置１００に送信する。対象は、具体的には、試着室の外観である。また、出力装置１００は、対象についての説明文となる文書を記憶している。説明文は、具体的には、人間が試着室を利用中は、試着室のカーテンが閉まっている傾向があることの説明文である。そして、出力装置１００は、画像と文書とに基づいて、危険度を判断する問題を解く。危険度は、例えば、試着室に避難が未完了の人間が残っている可能性の高さを示す指標値である。出力装置１００は、例えば、災害時に、試着室に避難が未完了の人間が残っている可能性の高さを示す危険度を判断する。

[0055]（情報処理システム２００の適用例３）

適用例３において、出力装置１００は、動画を形成する画像と、画像についての説明文となる文書を記憶している。動画は、例えば、料理の様子を写した動画である。説明文は、具体的には、料理の手順についての説明文である。そして、出力装置１００は、画像と文書とに基づいて、危険度を判断する問題を解く。危険度は、例えば、料理中の危険性の高さを示す指標値である。出力装置１００は、例えば、料理中の危険性の高さを示す危険度を判断する。

[0056]（出力装置１００のハードウェア構成例）

次に、図３を用いて、出力装置１００のハードウェア構成例について説明する。

[0057] 図３は、出力装置１００のハードウェア構成例を示すブロック図である。図３において、出力装置１００は、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）３０１と、メモリ３０２と、ネットワークＩ／Ｆ（Ｉｎｔｅｒｆａｃｅ）３０３と、記録媒体Ｉ／Ｆ３０４と、記録媒体３０５とを有する。また、各構成部は、バス３００によってそれぞれ接続される。

[0058] ここで、ＣＰＵ３０１は、出力装置１００の全体の制御を司る。メモリ３０２は、例えば、ＲＯＭ（Ｒｅａｄ　Ｏｎｌｙ　Ｍｅｍｏｒｙ）、ＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）およびフラッシュＲＯＭなどを有する。具体的には、例えば、フラッシュＲＯＭやＲＯＭが各種プログラ

ムを記憶し、ＲＡＭがＣＰＵ３０１のワークエリアとして使用される。メモリ３０２に記憶されるプログラムは、ＣＰＵ３０１にロードされることで、コーディングされている処理をＣＰＵ３０１に実行させる。

[0059] ネットワークＩ／Ｆ３０３は、通信回線を通じてネットワーク２１０に接続され、ネットワーク２１０を介して他のコンピュータに接続される。そして、ネットワークＩ／Ｆ３０３は、ネットワーク２１０と内部のインターフェースを司り、他のコンピュータからのデータの入出力を制御する。ネットワークＩ／Ｆ３０３は、例えば、モデムやＬＡＮアダプタなどである。

[0060] 記録媒体Ｉ／Ｆ３０４は、ＣＰＵ３０１の制御に従って記録媒体３０５に対するデータのリード／ライトを制御する。記録媒体Ｉ／Ｆ３０４は、例えば、ディスクドライブ、ＳＳＤ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ）、ＵＳＢ（Ｕｎｉｖｅｒｓａｌ　Ｓｅｒｉａｌ　Ｂｕｓ）ポートなどである。記録媒体３０５は、記録媒体Ｉ／Ｆ３０４の制御で書き込まれたデータを記憶する不揮発メモリである。記録媒体３０５は、例えば、ディスク、半導体メモリ、ＵＳＢメモリなどである。記録媒体３０５は、出力装置１００から着脱可能であってもよい。

[0061] 出力装置１００は、上述した構成部のほか、例えば、キーボード、マウス、ディスプレイ、プリンタ、スキャナ、マイク、スピーカーなどを有してもよい。また、出力装置１００は、記録媒体Ｉ／Ｆ３０４や記録媒体３０５を複数有していてもよい。また、出力装置１００は、記録媒体Ｉ／Ｆ３０４や記録媒体３０５を有していなくてもよい。

[0062]（クライアント装置２０１のハードウェア構成例）

クライアント装置２０１のハードウェア構成例は、具体的には、図３に示した出力装置１００のハードウェア構成例と同様であるため、説明を省略する。

[0063]（端末装置２０２のハードウェア構成例）

端末装置２０２のハードウェア構成例は、具体的には、図３に示した出力装置１００のハードウェア構成例と同様であるため、説明を省略する。

[0064] （出力装置１００の機能的構成例）

次に、図４を用いて、出力装置１００の機能的構成例について説明する。

[0065] 図４は、出力装置１００の機能的構成例を示すブロック図である。出力装置１００は、記憶部４００と、取得部４０１と、生成部４０２と、結合部４０３と、変換部４０４と、正規化部４０５と、出力部４０６とを含む。

[0066] 記憶部４００は、例えば、図３に示したメモリ３０２や記録媒体３０５などの記憶領域によって実現される。以下では、記憶部４００が、出力装置１００に含まれる場合について説明するが、これに限らない。例えば、記憶部４００が、出力装置１００とは異なる装置に含まれ、記憶部４００の記憶内容が出力装置１００から参照可能である場合があってもよい。

[0067] 取得部４０１～出力部４０６は、制御部の一例として機能する。取得部４０１～出力部４０６は、具体的には、例えば、図３に示したメモリ３０２や記録媒体３０５などの記憶領域に記憶されたプログラムをＣＰＵ３０１に実行させることにより、または、ネットワークＩ／Ｆ３０３により、その機能を実現する。各機能部の処理結果は、例えば、図３に示したメモリ３０２や記録媒体３０５などの記憶領域に記憶される。

[0068] 記憶部４００は、各機能部の処理において参照され、または更新される各種情報を記憶する。記憶部４００は、Ａｔｔｅｎｔｉｏｎを実現し、第一のモーダルの情報に基づくベクトルを、第二のモーダルの情報に基づくベクトルに基づいて補正し、補正後の第一のモーダルの情報に基づくベクトルを出力する変換モデルを記憶する。

[0069] 例えば、第一のモーダルは、画像に関するモーダルであり、第二のモーダルは、文書に関するモーダルである。例えば、第一のモーダルは、画像に関するモーダルであり、第二のモーダルは、音声に関するモーダルである。例えば、第一のモーダルは、第一の言語の文書に関するモーダルであり、第二のモーダルは、第二の言語の文書に関するモーダルである。例えば、第一のモーダルは、第二のモーダルと同一であってもよい。

[0070] 取得部４０１は、各機能部の処理に用いられる各種情報を取得する。取得

部４０１は、取得した各種情報を、記憶部４００に記憶し、または、各機能部に出力する。また、取得部４０１は、記憶部４００に記憶しておいた各種情報を、各機能部に出力してもよい。取得部４０１は、例えば、ユーザの操作入力に基づき、各種情報を取得する。取得部４０１は、例えば、出力装置１００とは異なる装置から、各種情報を受信してもよい。

[0071] 取得部４０１は、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとを取得する。取得部４０１は、例えば、ユーザによる、第一のモーダルの情報に基づくベクトルを生成する元となる第一のモーダルの情報と、第二のモーダルの情報に基づくベクトルを生成する元となる第二のモーダルの情報との入力を受け付ける。そして、取得部４０１は、入力された各種情報に基づいて、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとを生成する。

[0072] 取得部４０１は、具体的には、第一のモーダルの情報として、画像を取得し、第一のモーダルの情報に基づくベクトルとして、取得した画像に関する特徴量ベクトルを生成する。画像に関する特徴量ベクトルは、例えば、画像に写る物体ごとの特徴量ベクトルを並べたものである。また、取得部４０１は、具体的には、第二のモーダルの情報として、文書を取得し、第二のモーダルの情報に基づくベクトルとして、取得した文書に関する特徴量ベクトルを生成する。文書に関する特徴量ベクトルは、例えば、文書に含まれる単語ごとの特徴量ベクトルを並べたものである。

[0073] 取得部４０１は、例えば、第一のモーダルの情報に基づくベクトルを生成する元となる第一のモーダルの情報と、第二のモーダルの情報に基づくベクトルを生成する元となる第二のモーダルの情報とを、クライアント装置２０１または端末装置２０２から受信してもよい。そして、取得部４０１は、取得した各種情報に基づいて、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとを生成する。

[0074] 取得部４０１は、具体的には、第一のモーダルの情報として、画像を取得し、第一のモーダルの情報に基づくベクトルとして、取得した画像に関する

特徴量ベクトルを生成する。画像に関する特徴量ベクトルは、例えば、画像に写る物体ごとの特徴量ベクトルを並べたものである。また、取得部４０１は、具体的には、第二のモーダルの情報として、文書を取得し、第二のモーダルの情報に基づくベクトルとして、取得した文書に関する特徴量ベクトルを生成する。文書に関する特徴量ベクトルは、例えば、文書に含まれる単語ごとの特徴量ベクトルを並べたものである。

[0075] 取得部４０１は、例えば、ユーザによる、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとの入力を受け付けることにより、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとを取得してもよい。取得部４０１は、例えば、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとを、クライアント装置２０１または端末装置２０２から受信することにより取得してもよい。

[0076] 取得部４０１は、いずれかの機能部の処理を開始する開始トリガーを受け付けてもよい。開始トリガーは、例えば、ユーザによる所定の操作入力があったことである。開始トリガーは、例えば、他のコンピュータから、所定の情報を受信したことであってもよい。開始トリガーは、例えば、いずれかの機能部が所定の情報を出力したことであってもよい。取得部４０１は、例えば、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとを取得したことを、各機能部の処理を開始する開始トリガーとして受け付ける。

[0077] 生成部４０２は、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとの相関に基づいて、第一のモーダルの情報に基づくベクトルを補正する補正ベクトルを生成する。相関は、例えば、第一のモーダルの情報に基づくベクトルから得たベクトルと、第二のモーダルの情報に基づくベクトルから得たベクトルとの類似度によって表現される。第一のモーダルの情報に基づくベクトルから得たベクトルは、例えば、クエリである。第二のモーダルの情報に基づくベクトルから得たベクトルは、例え

ば、キーである。類似度は、例えば、内積によって表現される。類似度は、例えば、差分の二乗和などによって表現されてもよい。

[0078] 生成部４０２は、例えば、第一のモーダルの情報に基づくベクトルから得たベクトルと、第二のモーダルの情報に基づくベクトルから得たベクトルとの内積に基づいて、補正ベクトルを生成する。生成部４０２は、具体的には、第一のモーダルの情報に基づくベクトルから得たクエリと、第二のモーダルの情報に基づくベクトルから得たキーとの内積に基づいて、第一のモーダルの情報に基づくベクトルを補正する補正ベクトルを生成する。

[0079] 生成部４０２は、より具体的には、画像に関するモーダルの情報に基づくベクトルから得たクエリと、文書に関するモーダルの情報に基づくベクトルから得たキーとの内積に基づいて、画像に関するモーダルの情報に基づくベクトルを補正する補正ベクトルを生成する。ここで、補正ベクトルを生成する一例は、例えば、図７を用いて後述する動作例に示す。これにより、生成部４０２は、第二のモーダルの情報に基づくベクトルのうち、第一のモーダルの情報に基づくベクトルと相対的に関連深い成分ほど、第一のモーダルの情報に基づくベクトルに強く反映されるように、第一のモーダルの情報に基づくベクトルを補正可能な補正ベクトルを生成することができる。

[0080] 結合部４０３は、生成した補正ベクトルを、第一のモーダルの情報に基づくベクトルに結合する。結合部４０３は、例えば、補正ベクトルを、第一のモーダルの情報に基づくベクトルに加算せず、第一のモーダルの前後いずれかに結合する。これにより、結合部４０３は、第一のモーダルの情報に基づくベクトルと第二のモーダルの情報に基づくベクトルとのうち問題の解決に有用な情報が失われ辛く、反映され易いように、第一のモーダルの情報に基づくベクトルを加工することができる。

[0081] 変換部４０４は、所定のルールに従って、結合後の第一のモーダルの情報に基づくベクトルを圧縮する。所定のルールは、例えば、学習により自動で設定される。変換部４０４は、例えば、多層ニューラルネットワークを用いて、結合後の第一のモーダルの情報に基づくベクトルを圧縮する。これによ

り、変換部４０４は、結合後の第一のモーダルの情報に基づくベクトルの次元数を、扱いやすい次元数に変換することができる。

[0082] 正規化部４０５は、圧縮後の第一のモーダルの情報に基づくベクトルに対して、正規化処理を実施する。正規化部４０５は、例えば、第一のモーダルの情報に基づくベクトルと、補正ベクトルとの和を正規化し、当該正規化により得たベクトルと、圧縮後の第一のモーダルの情報に基づくベクトルとの和を正規化する。これにより、正規化部４０５は、第一のモーダルの情報に基づくベクトルと第二のモーダルの情報に基づくベクトルとのうち問題の解決に有用な情報が効率よく反映された、問題の解決に有用なベクトルを得ることができる。

[0083] 正規化部４０５は、例えば、結合後の第一のモーダルの情報に基づくベクトルと、圧縮後の第一のモーダルの情報に基づくベクトルとの和を正規化する。これにより、正規化部４０５は、第一のモーダルの情報に基づくベクトルと第二のモーダルの情報に基づくベクトルとのうち問題の解決に有用な情報が効率よく反映された、問題の解決に有用なベクトルを得ることができる。

[0084] 出力部４０６は、いずれかの機能部の処理結果を出力する。出力形式は、例えば、ディスプレイへの表示、プリンタへの印刷出力、ネットワークＩ／Ｆ３０３による外部装置への送信、または、メモリ３０２や記録媒体３０５などの記憶領域への記憶である。これにより、出力部４０６は、各機能部の処理結果をユーザに通知可能にし、出力装置１００の利便性の向上を図ることができる。

[0085] 出力部４０６は、正規化処理により得たベクトルを出力する。これにより、出力部４０６は、正規化処理により得たベクトルを利用し、Ａｔｔｅｎｔｉｏｎを実現することができる。そして、出力部４０６は、Ａｔｔｅｎｔｉｏｎにより、Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋを実現可能にすることができる。

[0086] 出力部４０６は、例えば、Ａｔｔｅｎｔｉｏｎにより、問題の解決に有用

に、正規化処理により得られたベクトルを出力することができる。このため、出力部４０６は、Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋを、問題の解決に有用になるように学習可能にすることができる。また、出力部４０６は、問題を解いた際の解の精度を向上可能にすることができる。

[0087] （出力装置１００の動作例）

次に、図５～図７を用いて、出力装置１００の動作例について説明する。まず、図５を用いて、出力装置１００によって用いられるＣｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋ５００の具体例について説明する。

[0088] 図５は、Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋ５００の具体例を示す説明図である。以下の説明では、Ｃｏ－Ａｔｔｅｎｔｉｏｎ　Ｎｅｔｗｏｒｋ５００を「ＣＡＮ５００」と表記する場合がある。また、ターゲットアテンションを「ＴＡ」と表記する場合がある。また、セルフアテンションを「ＳＡ」と表記する場合がある。

[0089] 図５に示すように、ＣＡＮ５００は、画像ＴＡ層５０１と、画像ＳＡ層５０２と、文書ＴＡ層５０３と、文書ＳＡ層５０４と、結合層５０５と、統合ＳＡ層５０６とを有する。

[0090] 図５において、ＣＡＮ５００は、文書に関する特徴量ベクトルＬと画像に関する特徴量ベクトルＩとが入力されたことに応じて、ベクトルZ_Tを出力する。文書に関する特徴量ベクトルＬは、例えば、文書に関するＭ個の特徴量ベクトルを並べたものである。Ｍ個の特徴量ベクトルは、例えば、文書に含まれるＭ個の単語を示す特徴量ベクトルである。画像に関する特徴量ベクトルＩは、例えば、画像に関するＮ個の特徴量ベクトルを並べたものである。Ｎ個の特徴量ベクトルは、例えば、画像に写ったＮ個の物体を示す特徴量ベクトルである。

[0091] 具体的には、画像ＴＡ層５０１は、画像に関する特徴量ベクトルＩと、文書に関する特徴量ベクトルＬとの入力を受け付ける。画像ＴＡ層５０１は、画像に関する特徴量ベクトルＩから得たクエリと、文書に関する特徴量ベクトルＬから得たキーおよびバリューとに基づいて、画像に関する特徴量ベク

トルＩを補正する。画像ＴＡ層５０１は、補正後の画像に関する特徴量ベクトルＩを、画像ＳＡ層５０２に出力する。画像ＴＡ層５０１の具体例については、例えば、図７および図８を用いて後述する。

[0092] また、画像ＳＡ層５０２は、補正後の画像に関する特徴量ベクトルＩの入力を受け付ける。画像ＳＡ層５０２は、補正後の画像に関する特徴量ベクトルＩから得たクエリ、キーおよびバリューに基づいて、補正後の画像に関する特徴量ベクトルＩをさらに補正し、新たな特徴量ベクトルZ_Iを生成し、結合層５０５に出力する。画像ＳＡ層５０２を実現するＳＡ層の具体例については、例えば、図６を用いて後述する。

[0093] また、文書ＴＡ層５０３は、文書に関する特徴量ベクトルＬと、画像に関する特徴量ベクトルＩとの入力を受け付ける。文書ＴＡ層５０３は、文書に関する特徴量ベクトルＬから得たクエリと、画像に関する特徴量ベクトルＩから得たキーおよびバリューとに基づいて、文書に関する特徴量ベクトルＬを補正する。文書ＴＡ層５０３は、補正後の文書に関する特徴量ベクトルＬを、文書ＳＡ層５０４に出力する。文書ＴＡ層５０３を実現するＴＡ層の具体例については、例えば、図６を用いて後述する。

[0094] また、文書ＳＡ層５０４は、補正後の文書に関する特徴量ベクトルＬの入力を受け付ける。文書ＳＡ層５０４は、補正後の文書に関する特徴量ベクトルＬから得たクエリ、キーおよびバリューに基づいて、補正後の文書に関する特徴量ベクトルＬをさらに補正し、新たな特徴量ベクトルZ_Lを生成して出力する。文書ＳＡ層５０４を実現するＳＡ層の具体例については、例えば、図６を用いて後述する。

[0095] また、結合層５０５は、集約用ベクトルＨと、特徴量ベクトルZ_Iと、特徴量ベクトルZ_Lとの入力を受け付ける。結合層５０５は、集約用ベクトルＨと、特徴量ベクトルZ_Iと、特徴量ベクトルZ_Lとを結合し、結合ベクトルＣを生成し、統合ＳＡ層５０６に出力する。

[0096] また、統合ＳＡ層５０６は、結合ベクトルＣの入力を受け付ける。統合ＳＡ層５０６は、結合ベクトルＣから得たクエリ、キーおよびバリューに基づ

いて、結合ベクトルＣを補正し、特徴量ベクトルZ_Tを生成して出力する。特徴量ベクトルZ_Tは、集約ベクトルZ_Hと、文書に関する統合特徴量ベクトルZ_1～Z_Mと、画像に関する統合特徴量ベクトルZ_{M+1}～Z_{M+N}とを含む。これにより、出力装置１００は、問題を解いた際の解の精度を向上させる観点で有用な集約ベクトルZ_Hを含む特徴量ベクトルZ_Tを生成し、参照可能にすることができる。このため、出力装置１００は、問題を解いた際の解の精度を向上可能にすることができる。

[0097] ここでは、説明の簡略化のため、画像ＴＡ層５０１と、画像ＳＡ層５０２と、文書ＴＡ層５０３と、文書ＳＡ層５０４とのグループ５１０が、１段である場合について説明したが、これに限らない。例えば、画像ＴＡ層５０１と、画像ＳＡ層５０２と、文書ＴＡ層５０３と、文書ＳＡ層５０４とのグループ５１０が、複数段存在する場合があってもよい。これによれば、出力装置１００は、問題を解いた際の解の精度のさらなる向上を図ることができる。

[0098] ここでは、ＣＡＮ５００が、画像ＴＡ層５０１と、画像ＳＡ層５０２と、文書ＴＡ層５０３と、文書ＳＡ層５０４と、結合層５０５と、統合ＳＡ層５０６とを有する場合について説明したが、これに限らない。例えば、ＣＡＮ５００が、結合層５０５と、統合ＳＡ層５０６とを有していない場合があってもよい。この場合、出力装置１００は、例えば、問題を解くにあたり、画像ＳＡ層５０２の出力と、文書ＳＡ層５０４の出力とを利用する。

[0099] 次に、図６の説明に移行し、ＣＡＮ５００を形成する画像ＳＡ層５０２と文書ＳＡ層５０４と統合ＳＡ層５０６となどを実現するＳＡ層６００の具体例と、ＣＡＮ５００を形成する文書ＴＡ層５０３などを実現するＴＡ層６１０の具体例とについて説明する。ＣＡＮ５００を形成する画像ＴＡ層５０１の具体例については、図７を用いて後述する。

[0100] 図６は、ＳＡ層６００の具体例と、ＴＡ層６１０の具体例とを示す説明図である。以下の説明では、Ｍｕｌｔｉ－Ｈｅａｄ　Ａｔｔｅｎｔｉｏｎを「ＭＨＡ」と表記する場合がある。また、Ａｄｄ＆Ｎｏｒｍを「Ａ＆Ｎ」と表

記する場合がある。また、Ｆｅｅｄ　Ｆｏｒｗａｒｄを「ＦＦ」と表記する場合がある。

[0101] 図６に示すように、ＳＡ層６００は、ＭＨＡ層６０１と、Ａ＆Ｎ層６０２と、ＦＦ層６０３と、Ａ＆Ｎ層６０４とを有する。ＭＨＡ層６０１は、入力ベクトルＸから得たクエリＱとキーＫとバリューＶとに基づいて、入力ベクトルＸを補正する補正ベクトルＲを生成し、Ａ＆Ｎ層６０２に出力する。ＭＨＡ層６０１は、具体的には、入力ベクトルＸを、Ｈｅａｄ個のベクトルに分割して処理する。Ｈｅａｄは、１以上の自然数である。

[0102] Ａ＆Ｎ層６０２は、入力ベクトルＸと補正ベクトルＲとを加算した上で正規化し、正規化後のベクトルを、ＦＦ層６０３とＡ＆Ｎ層６０４とに出力する。ＦＦ層６０３は、正規化後のベクトルを圧縮し、圧縮後のベクトルを、Ａ＆Ｎ層６０４に出力する。Ａ＆Ｎ層６０４は、正規化後のベクトルと、圧縮後のベクトルとを加算した上で正規化し、出力ベクトルＺを生成して出力する。

[0103] また、ＴＡ層６１０は、ＭＨＡ層６１１と、Ａ＆Ｎ層６１２と、ＦＦ層６１３と、Ａ＆Ｎ層６１４とを有する。ＭＨＡ層６１１は、入力ベクトルＸから得たクエリＱと、入力ベクトルＹから得たキーＫとバリューＶとに基づいて、入力ベクトルＸを補正する補正ベクトルＲを生成し、Ａ＆Ｎ層６１２に出力する。Ａ＆Ｎ層６１２は、入力ベクトルＸと補正ベクトルＲとを加算した上で正規化し、正規化後のベクトルを、ＦＦ層６１３とＡ＆Ｎ層６１４とに出力する。ＦＦ層６１３は、正規化後のベクトルを圧縮し、圧縮後のベクトルを、Ａ＆Ｎ層６１４に出力する。Ａ＆Ｎ層６１４は、正規化後のベクトルと、圧縮後のベクトルとを加算した上で正規化し、出力ベクトルＺを生成して出力する。

[0104] 上述したＭＨＡ層６０１やＭＨＡ層６１１は、より具体的には、Ｈｅａｄの個数分のＡｔｔｅｎｔｉｏｎ層６２０により形成される。Ａｔｔｅｎｔｉｏｎ層６２０は、ＭａｔＭｕｌ層６２１と、Ｓｃａｌｅ層６２２と、Ｍａｓｋ層６２３と、ＳｏｆｔＭａｘ層６２４と、ＭａｔＭｕｌ層６２５とを有す

る。

[0105] ＭａｔＭｕｌ層６２１は、クエリＱとキーＫとの内積を算出し、Ｓｃｏｒｅに設定する。Ｓｃａｌｅ層６２２は、Ｓｃｏｒｅ全体を定数ａで除算し、更新する。Ｍａｓｋ層６２３は、更新後のＳｃｏｒｅをマスク処理してもよい。ＳｏｆｔＭａｘ層６２４は、更新後のＳｃｏｒｅを、正規化し、Ａｔｔに設定する。ＭａｔＭｕｌ層６２５は、ＡｔｔとバリューＶとの内積を算出し、補正ベクトルＲに設定する。次に、図７および図８を用いて、ＣＡＮ５００を形成する画像ＴＡ層５０１の具体例について説明する。

[0106] 図７は、画像ＴＡ層５０１の具体例を示す説明図である。図７において、画像ＴＡ層５０１は、ＭＨＡ層７０１と、Ａ＆Ｎ層７０２と、Ｃｏｎ層７０３と、ＦＦ層７０４と、Ａ＆Ｎ層７０５とを含む。ＭＨＡ層７０１は、入力ベクトルＸから得たクエリＱと、入力ベクトルＹから得たキーＫとバリューＶとに基づいて、入力ベクトルＸを補正する補正ベクトルＲを生成し、Ａ＆Ｎ層７０２およびＣｏｎ層７０３に出力する。Ａ＆Ｎ層７０２は、入力ベクトルＸと補正ベクトルＲとを加算した上で正規化し、正規化後のベクトルを、Ａ＆Ｎ層７０５に出力する。

[0107] Ｃｏｎ層７０３は、入力ベクトルＸと補正ベクトルＲとを結合し、結合ベクトルをＦＦ層７０４に出力する。ＦＦ層７０４は、結合ベクトルを圧縮し、圧縮後のベクトルを、Ａ＆Ｎ層７０５に出力する。Ａ＆Ｎ層７０５は、正規化後のベクトルと、圧縮後のベクトルとを加算した上で正規化し、正規化で得た出力ベクトルを出力する。次に、図８を用いて、画像ＴＡ層５０１の別の具体例について説明する。

[0108] 図８は、画像ＴＡ層５０１の別の具体例を示す説明図である。図８において、画像ＴＡ層５０１は、ＭＨＡ層８０１と、Ｃｏｎ層８０２と、ＦＦ層８０３と、Ａ＆Ｎ層８０４とを含む。ＭＨＡ層８０１は、入力ベクトルＸから得たクエリＱと、入力ベクトルＹから得たキーＫとバリューＶとに基づいて、入力ベクトルＸを補正する補正ベクトルＲを生成し、Ｃｏｎ層８０２に出力する。

[0109] Ｃｏｎ層８０２は、入力ベクトルＸと補正ベクトルＲとを結合し、結合ベクトルをＦＦ層８０３およびＡ＆Ｎ層８０４に出力する。ＦＦ層８０３は、結合ベクトルを圧縮し、圧縮後のベクトルを、Ａ＆Ｎ層８０４に出力する。Ａ＆Ｎ層８０４は、結合ベクトルと、圧縮後のベクトルとを加算した上で正規化し、正規化で得た出力ベクトルを出力する。次に、図９を用いて、画像ＴＡ層５０１と文書ＴＡ層５０３との比較例について説明する。

[0110] 図９は、画像ＴＡ層５０１と文書ＴＡ層５０３との比較例を示す説明図である。図９に示すように、画像ＴＡ層５０１と、文書ＴＡ層５０３とは、文書に関する特徴量ベクトルＬと、画像に関する特徴量ベクトルＩとの入力を受け付ける。しかしながら、画像ＴＡ層５０１と、文書ＴＡ層５０３とは、それぞれ、異なる手法で、文書に関する特徴量ベクトルＬと、画像に関する特徴量ベクトルＩとを扱うことになる。

[0111] 例えば、画像ＴＡ層５０１は、画像に関する特徴量ベクトルＩに、ベクトルZ_{I1}を結合することにより、新たな特徴量ベクトルZ_{I2}を生成する。一方で、文書ＴＡ層５０３は、文書に関する特徴量ベクトルＬに、ベクトルZ_{L1}を加算することにより、新たな特徴量ベクトルZ_{L2}を生成する。これにより、出力装置１００は、それぞれ性質が異なる、文書に関する特徴量ベクトルＬと、画像に関する特徴量ベクトルＩとに対し、異なる扱い方をすることができる。

[0112] そして、出力装置１００は、画像ＴＡ層５０１において、文書に関する特徴量ベクトルＬと、画像に関する特徴量ベクトルＩとのうち、問題の解決に有用な情報が失われ辛くすることができる。結果として、出力装置１００は、複数のモーダルの情報を用いて問題を解くにあたり有用なベクトルを得ることができ、問題を解いた際の解の精度を向上可能にすることができる。

[0113] ここでは、画像ＴＡ層５０１を、図７および図８に示す具体例のように形成する場合について説明したが、これに限らない。例えば、画像ＳＡ層５０２と、文書ＴＡ層５０３と、文書ＳＡ層５０４と、統合ＳＡ層５０６との少なくともいずれかを、図７および図８に示す具体例と同様に形成する場合が

あってもよい。次に、図１０を用いて、出力装置１００による、ＣＡＮ５００を用いた動作の一例について説明する。

[0114] 図１０は、ＣＡＮ５００を用いた動作の一例を示す説明図である。図１０において、出力装置１００は、文書１０００を取得し、画像１０１０を取得する。出力装置１００は、文書１０００をトークン化し、トークン集合１００１をベクトル化し、文書１０００に関する特徴量ベクトル１００２を生成し、ＣＡＮ５００に入力する。また、出力装置１００は、画像１０１０から物体を検出し、物体ごとの部分画像の集合１０１１をベクトル化し、画像１０１０に関する特徴量ベクトル１０１２を生成し、ＣＡＮ５００に入力する。

[0115] 出力装置１００は、ＣＡＮ５００から、特徴量ベクトルZ_Tを取得し、特徴量ベクトルZ_Tに含まれる集約ベクトルZ_Hを、危険度推定器１０３０に入力する。出力装置１００は、危険度推定器１０３０から推定結果Ｎｏを取得する。これにより、出力装置１００は、画像と文書との特徴が反映された集約ベクトルZ_Hを用いて、危険度推定器１０３０に危険であるか否かを推定させることができ、危険であるか否かを精度よく推定可能にすることができる。危険度推定器１０３０は、例えば、銃を持った人物が写っている画像１０１０があるが、ミュージアムの展示物であることを示す文書もあるため、推定結果Ｎｏ＝危険ではないと推定することができる。

[0116] （出力装置１００の利用例）

次に、図１１～図１４を用いて、出力装置１００の利用例について説明する。

[0117] 図１１および図１２は、出力装置１００の利用例１を示す説明図である。図１１において、出力装置１００は、学習フェーズを実施し、ＣＡＮ５００を学習する。出力装置１００は、例えば、何らかのシーンを写した画像１１００と、画像１１００に対応する字幕となる文書１１１０とを取得する。画像１１００は、例えば、りんごを切るシーンを写す。

[0118] 出力装置１００は、画像１１００を変換器１１２０により特徴量ベクトル

に変換し、ＣＡＮ５００に入力する。また、出力装置１００は、文書１１１０の単語ａｐｐｌｅをマスクした上で、変換器１１３０により特徴量ベクトルに変換し、ＣＡＮ５００に入力する。

[0119] 出力装置１００は、ＣＡＮ５００により生成された特徴量ベクトルを、識別器１１４０に入力し、マスクされた単語を予測した結果を取得し、マスクされた単語の正解「ａｐｐｌｅ」との誤差を算出する。出力装置１００は、算出した誤差に基づいて、誤差逆伝搬によりＣＡＮ５００を学習する。さらに、出力装置１００は、誤差逆伝搬により、変換器１１２０，１１３０や識別器１１４０を学習してもよい。

[0120] これにより、出力装置１００は、画像１１００と字幕となる文書１１１０の文脈とを考慮して単語を推定する観点で有用なように、ＣＡＮ５００、および変換器１１２０，１１３０や識別器１１４０を更新することができる。次に、図１２の説明に移行する。

[0121] 図１２において、出力装置１００は、試験フェーズを実施し、学習した変換器１１２０，１１３０と、学習したＣＡＮ５００とを用いて、回答を生成して出力する。出力装置１００は、例えば、何らかのシーンを写した画像１２００と、画像１２００に対応する質問文となる文書１２１０とを取得する。画像１２００は、例えば、りんごを切るシーンを写す。

[0122] 出力装置１００は、画像１２００を変換器１１２０により特徴量ベクトルに変換し、ＣＡＮ５００に入力する。また、出力装置１００は、文書１２１０を変換器１１３０により特徴量ベクトルに変換し、ＣＡＮ５００に入力する。出力装置１００は、ＣＡＮ５００により生成された特徴量ベクトルを、回答生成器１２２０に入力し、回答となる単語を取得して出力する。これにより、出力装置１００は、画像１２００と質問文となる文書１２１０の文脈とを考慮して、精度よく回答となる単語を推定することができる。

[0123] 図１３および図１４は、出力装置１００の利用例２を示す説明図である。図１３において、出力装置１００は、学習フェーズを実施し、ＣＡＮ５００を学習する。出力装置１００は、例えば、何らかのシーンを写した画像１３

００と、画像１３００に対応する字幕となる文書１３１０とを取得する。画像１３００は、例えば、りんごを切るシーンを写す。

[0124] 出力装置１００は、画像１３００を変換器１３２０により特徴量ベクトルに変換し、ＣＡＮ５００に入力する。また、出力装置１００は、文書１３１０の単語ａｐｐｌｅをマスクした上で、変換器１３３０により特徴量ベクトルに変換し、ＣＡＮ５００に入力する。

[0125] 出力装置１００は、ＣＡＮ５００により生成された特徴量ベクトルを、識別器１３４０に入力し、画像に写ったシーンの危険度を予測した結果を取得し、危険度の正解との誤差を算出する。出力装置１００は、算出した誤差に基づいて、誤差逆伝搬によりＣＡＮ５００を学習する。また、出力装置１００は、誤差逆伝搬により、変換器１３２０，１３３０や識別器１３４０を学習する。

[0126] これにより、出力装置１００は、画像１３００と字幕となる文書１３１０の文脈とを考慮して危険度を予測する観点で有用なように、ＣＡＮ５００、および変換器１１２０，１１３０や識別器１１４０を更新することができる。次に、図１４の説明に移行する。

[0127] 図１４において、出力装置１００は、試験フェーズを実施し、学習した変換器１３２０，１３３０や識別器１３４０と、学習したＣＡＮ５００とを用いて、危険度を予測して出力する。出力装置１００は、例えば、何らかのシーンを写した画像１４００と、画像に対応する説明文となる文書１４１０とを取得する。画像１４００は、例えば、ももを切るシーンを写す。

[0128] 出力装置１００は、画像１４００を変換器１３２０により特徴量ベクトルに変換し、ＣＡＮ５００に入力する。また、出力装置１００は、文書１４１０を変換器１３３０により特徴量ベクトルに変換し、ＣＡＮ５００に入力する。出力装置１００は、ＣＡＮ５００により生成された特徴量ベクトルを、識別器１３４０に入力し、危険度を取得して出力する。これにより、出力装置１００は、画像１４００と説明文となる文書１４１０の文脈とを考慮して、精度よく危険度を予測することができる。

[0129] （学習処理手順）

次に、図１５を用いて、出力装置１００が実行する、学習処理手順の一例について説明する。学習処理は、例えば、図３に示したＣＰＵ３０１と、メモリ３０２や記録媒体３０５などの記憶領域と、ネットワークＩ／Ｆ３０３とによって実現される。

[0130] 図１５は、学習処理手順の一例を示すフローチャートである。図１５において、出力装置１００は、画像の特徴量ベクトルと、文書の特徴量ベクトルとを取得する（ステップＳ１５０１）。

[0131] 次に、出力装置１００は、取得した画像の特徴量ベクトルから生成したクエリと、取得した文書の特徴量ベクトルから生成したキーおよびバリューとに基づいて、画像ＴＡ層５０１を用いて、画像の特徴量ベクトルを補正する（ステップＳ１５０２）。ここで、出力装置１００は、具体的には、図１４に後述するアテンション処理を実行することにより、画像の特徴量ベクトルを補正する。

[0132] そして、出力装置１００は、補正後の画像の特徴量ベクトルに基づいて、画像ＳＡ層５０２を用いて、補正後の画像の特徴量ベクトルをさらに補正し、新たに画像の特徴量ベクトルを生成する（ステップＳ１５０３）。

[0133] 次に、出力装置１００は、取得した文書の特徴量ベクトルから生成したクエリと、取得した画像の特徴量ベクトルから生成したキーおよびバリューとに基づいて、文書ＴＡ層５０３を用いて、文書の特徴量ベクトルを補正する（ステップＳ１５０４）。

[0134] そして、出力装置１００は、補正後の文書の特徴量ベクトルに基づいて、文書ＳＡ層５０４を用いて、補正後の文書の特徴量ベクトルをさらに補正し、新たに文書の特徴量ベクトルを生成する（ステップＳ１５０５）。

[0135] 次に、出力装置１００は、集約用ベクトルを初期化する（ステップＳ１５０６）。そして、出力装置１００は、集約用ベクトルと、生成した画像の特徴量ベクトルと、生成した文書の特徴量ベクトルとを結合し、結合ベクトルを生成する（ステップＳ１５０７）。

[0136] 次に、出力装置１００は、結合ベクトルに基づいて、統合ＳＡ層５０６を用いて、結合ベクトルを補正し、集約ベクトルを生成する（ステップＳ１５０８）。そして、出力装置１００は、集約ベクトルに基づいて、ＣＡＮ５００を学習する（ステップＳ１５０９）。

[0137] その後、出力装置１００は、学習処理を終了する。これにより、出力装置１００は、ＣＡＮ５００を用いて問題を解くにあたり、問題を解いた際の解の精度が向上するように、ＣＡＮ５００のパラメータを更新することができる。

[0138] ここで、出力装置１００は、図１５の一部ステップの処理の順序を入れ替えて実行してもよい。例えば、ステップＳ１５０２，Ｓ１５０３の処理と、ステップＳ１５０４，Ｓ１５０５の処理との順序は入れ替え可能である。また、出力装置１００は、ステップＳ１５０２～Ｓ１５０５の処理を繰り返し実行してもよい。

[0139] （推定処理手順）

次に、図１６を用いて、出力装置１００が実行する、推定処理手順の一例について説明する。推定処理は、例えば、図３に示したＣＰＵ３０１と、メモリ３０２や記録媒体３０５などの記憶領域と、ネットワークＩ／Ｆ３０３とによって実現される。

[0140] 図１６は、推定処理手順の一例を示すフローチャートである。図１６において、出力装置１００は、画像の特徴量ベクトルと、文書の特徴量ベクトルとを取得する（ステップＳ１６０１）。

[0141] 次に、出力装置１００は、取得した画像の特徴量ベクトルから生成したクエリと、取得した文書の特徴量ベクトルから生成したキーおよびバリューとに基づいて、画像ＴＡ層５０１を用いて、画像の特徴量ベクトルを補正する（ステップＳ１６０２）。ここで、出力装置１００は、具体的には、図１４に後述するアテンション処理を実行することにより、画像の特徴量ベクトルを補正する。

[0142] そして、出力装置１００は、補正後の画像の特徴量ベクトルに基づいて、

画像ＳＡ層５０２を用いて、補正後の画像の特徴量ベクトルをさらに補正し、新たに画像の特徴量ベクトルを生成する（ステップＳ１６０３）。

[0143] 次に、出力装置１００は、取得した文書の特徴量ベクトルから生成したクエリと、取得した画像の特徴量ベクトルから生成したキーおよびバリューとに基づいて、文書ＴＡ層５０３を用いて、文書の特徴量ベクトルを補正する（ステップＳ１６０４）。

[0144] そして、出力装置１００は、補正後の文書の特徴量ベクトルに基づいて、文書ＳＡ層５０４を用いて、補正後の文書の特徴量ベクトルをさらに補正し、新たに文書の特徴量ベクトルを生成する（ステップＳ１６０５）。

[0145] 次に、出力装置１００は、集約用ベクトルを初期化する（ステップＳ１６０６）。そして、出力装置１００は、集約用ベクトルと、生成した画像の特徴量ベクトルと、生成した文書の特徴量ベクトルとを結合し、結合ベクトルを生成する（ステップＳ１６０７）。

[0146] 次に、出力装置１００は、結合ベクトルに基づいて、統合ＳＡ層５０６を用いて、結合ベクトルを補正し、集約ベクトルを生成する（ステップＳ１６０８）。そして、出力装置１００は、集約ベクトルに基づいて、識別モデルを用いて、状況を推定する（ステップＳ１６０９）。

[0147] 次に、出力装置１００は、推定した状況を出力する（ステップＳ１６１０）。そして、出力装置１００は、推定処理を終了する。これにより、出力装置１００は、ＣＡＮ５００を用いて、問題を解いた際の解の精度を向上させることができる。

[0148] ここで、出力装置１００は、図１６の一部ステップの処理の順序を入れ替えて実行してもよい。例えば、ステップＳ１６０２，Ｓ１６０３の処理と、ステップＳ１６０４，Ｓ１６０５の処理との順序は入れ替え可能である。また、出力装置１００は、ステップＳ１６０２～Ｓ１６０５の処理を繰り返し実行してもよい。

[0149] （アテンション処理手順）

次に、図１７を用いて、画像ＴＡ層により、出力装置１００が実行する、

アテンション処理手順の一例について説明する。アテンション処理は、例えば、図３に示したＣＰＵ３０１と、メモリ３０２や記録媒体３０５などの記憶領域と、ネットワークＩ／Ｆ３０３とによって実現される。

[0150] 図１７は、アテンション処理手順の一例を示すフローチャートである。図１７において、出力装置１００は、ベクトルＸとなる画像の特徴量ベクトルと、ベクトルＹとなる文書の特徴量ベクトルとを取得する（ステップＳ１７０１）。

[0151] 次に、出力装置１００は、取得した画像の特徴量ベクトルからベクトルＱｕｅｒｙを生成する（ステップＳ１７０２）。そして、出力装置１００は、取得した文書の特徴量ベクトルからベクトルｋｅｙとベクトルＶａｌｕｅを生成する（ステップＳ１７０３）。

[0152] 次に、出力装置１００は、生成したベクトルＱｕｅｒｙと、生成したベクトルｋｅｙとの内積を算出する（ステップＳ１７０４）。そして、出力装置１００は、内積のｓｏｆｔｍａｘによりベクトルＡｔｔを生成する（ステップＳ１７０５）。

[0153] 次に、出力装置１００は、ベクトルＡｔｔとベクトルＶａｌｕｅとの内積によりベクトルＲを生成する（ステップＳ１７０６）。そして、出力装置１００は、ベクトルＲとベクトルＸとを結合したベクトルＸ’を生成する（ステップＳ１７０７）。

[0154] 次に、出力装置１００は、多層ニューラルネットワークにより、ベクトルＸ’を、ベクトルＸと同じ次元に圧縮し、ベクトルＸ”を生成する（ステップＳ１７０８）。そして、出力装置１００は、ベクトルＲとベクトルＸとを用いて、ベクトルＸ”を正規化し、正規化後のベクトルを取得する（ステップＳ１７０９）。

[0155] 次に、出力装置１００は、取得した正規化後のベクトルを出力する（ステップＳ１７１０）。そして、出力装置１００は、アテンション処理を終了する。これにより、出力装置１００は、画像と文書とのうち問題の解決に有用な情報が失われ辛いように、正規化後のベクトルを生成して取得することが

できる。

[0156] ここで、出力装置１００は、図１７の一部ステップの処理の順序を入れ替えて実行してもよい。例えば、ステップＳ１７０２の処理と、ステップＳ１７０３の処理との順序は入れ替え可能である。

[0157] 以上説明したように、出力装置１００によれば、第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとの相関に基づいて、第一のモーダルの情報に基づくベクトルを補正する補正ベクトルを生成することができる。出力装置１００によれば、生成した補正ベクトルを、第一のモーダルの情報に基づくベクトルに結合することができる。出力装置１００によれば、所定のルールに従って、結合後の第一のモーダルの情報に基づくベクトルを圧縮することができる。出力装置１００によれば、圧縮後の第一のモーダルの情報に基づくベクトルに対して、正規化処理を実施することができる。出力装置１００によれば、正規化処理により得たベクトルを出力することができる。これにより、出力装置１００は、第一のモーダルの情報に基づくベクトルと第二のモーダルの情報に基づくベクトルとのうち問題の解決に有用な情報を残して、問題を解くのに有用なベクトルを得ることができ、問題を解いた際の解の精度を向上可能にすることができる。

[0158] 出力装置１００によれば、第一のモーダルの情報に基づくベクトルから得たベクトルと、第二のモーダルの情報に基づくベクトルから得たベクトルとの内積に基づいて、補正ベクトルを生成することができる。これにより、出力装置１００は、アテンションを実現することができる。また、出力装置１００は、問題を解くのに有用な補正ベクトルを得ることができる。

[0159] 出力装置１００によれば、第一のモーダルの情報に基づくベクトルと、補正ベクトルとの和を正規化し、当該正規化により得たベクトルと、圧縮後の第一のモーダルの情報に基づくベクトルとの和を正規化することができる。これにより、出力装置１００は、正規化処理を実現することができる。

[0160] 出力装置１００によれば、結合後の第一のモーダルの情報に基づくベクトルと、圧縮後の第一のモーダルの情報に基づくベクトルとの和を正規化する

ことができる。これにより、出力装置１００は、正規化処理を実現することができる。

[0161] 出力装置１００によれば、第一のモーダルとして、画像に関するモーダルを採用することができる。出力装置１００によれば、第二のモーダルとして、文書に関するモーダルを採用することができる。これにより、出力装置１００は、ターゲットアテンション層を実現することができる。また、出力装置１００は、画像と文書とに基づいて問題を解く場合に適用可能にすることができる。

[0162] 出力装置１００によれば、第一のモーダルとして、画像に関するモーダルを採用することができる。出力装置１００によれば、第二のモーダルとして、音声に関するモーダルを採用することができる。これにより、出力装置１００は、ターゲットアテンション層を実現することができる。また、出力装置１００は、画像と音声とに基づいて問題を解く場合に適用可能にすることができる。

[0163] 出力装置１００によれば、第一のモーダルとして、第一の言語の文書に関するモーダルを採用することができる。出力装置１００によれば、第二のモーダルとして、第二の言語の文書に関するモーダルを採用することができる。これにより、出力装置１００は、ターゲットアテンション層を実現することができる。また、出力装置１００は、異なる言語の２つの文書に基づいて問題を解く場合に適用可能にすることができる。

[0164] 出力装置１００によれば、第一のモーダルと、第二のモーダルとに、同一のモーダルを採用することができる。これにより、出力装置１００は、セルフアテンション層を実現することができる。また、出力装置１００は、同一のモーダルの異なる情報に基づいて問題を解く場合に適用可能にすることができる。

[0165] なお、本実施の形態で説明した出力方法は、予め用意されたプログラムをＰＣやワークステーションなどのコンピュータで実行することにより実現することができる。本実施の形態で説明した出力プログラムは、コンピュータ

で読み取り可能な記録媒体に記録され、コンピュータによって記録媒体から読み出されることによって実行される。記録媒体は、ハードディスク、フレキシブルディスク、ＣＤ（Ｃｏｍｐａｃｔ　Ｄｉｓｃ）－ＲＯＭ、ＭＯ、ＤＶＤ（Ｄｉｇｉｔａｌ　Ｖｅｒｓａｔｉｌｅ　Ｄｉｓｃ）などである。また、本実施の形態で説明した出力プログラムは、インターネットなどのネットワークを介して配布してもよい。

符号の説明

[0166] １００　出力装置

１０１　生成モデル

１０２　結合モデル

１０３　圧縮モデル

１０４　正規化モデル

１１０　変換モデル

２００　情報処理システム

２０１　クライアント装置

２０２　端末装置

２１０　ネットワーク

３００　バス

３０１　ＣＰＵ

３０２　メモリ

３０３　ネットワークＩ／Ｆ

３０４　記録媒体Ｉ／Ｆ

３０５　記録媒体

４００　記憶部

４０１　取得部

４０２　生成部

４０３　結合部

４０４　変換部

４０５　正規化部

４０６　出力部

５００　ＣＡＮ

５０１　画像ＴＡ層

５０２　画像ＳＡ層

５０３　文書ＴＡ層

５０４　文書ＳＡ層

５０５　結合層

５０６　統合ＳＡ層

５１０　グループ

６００　ＳＡ層

６０１，６１１，７０１，８０１　ＭＨＡ層

６０２，６０４，６１２，６１４，７０２，７０５，８０４　Ａ＆Ｎ層

６０３，６１３，７０４，８０３　ＦＦ層

６１０　ＴＡ層

６２０　Ａｔｔｅｎｔｉｏｎ層

６２１，６２５　ＭａｔＭｕｌ層

６２２　Ｓｃａｌｅ層

６２３　Ｍａｓｋ層

６２４　ＳｏｆｔＭａｘ層

７０３，８０２　Ｃｏｎ層

１０００，１１１０，１２１０，１３１０，１４１０　文書

１００１　トークン集合

１００２，１０１２　特徴量ベクトル

１０１０，１１００，１２００，１３００，１４００　画像

１０１１　集合

１０３０　危険度推定器

１１２０，１１３０，１３２０，１３３０　変換器

１１４０，１３４０　識別器

１２２０　回答生成器

請求の範囲

[請求項1] 第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとの相関に基づいて、前記第一のモーダルの情報に基づくベクトルを補正する補正ベクトルを生成し、

生成した前記補正ベクトルを、前記第一のモーダルの情報に基づくベクトルに結合し、

所定のルールに従って、結合後の前記第一のモーダルの情報に基づくベクトルを圧縮し、

圧縮後の前記第一のモーダルの情報に基づくベクトルに対して、正規化処理を実施し、

前記正規化処理により得たベクトルを出力する、

処理をコンピュータが実行することを特徴とする出力方法。

[請求項2] 前記生成する処理は、

前記第一のモーダルの情報に基づくベクトルから得たベクトルと、前記第二のモーダルの情報に基づくベクトルから得たベクトルとの内積に基づいて、前記補正ベクトルを生成する、ことを特徴とする請求項１に記載の出力方法。

[請求項3] 前記正規化処理を実施する処理は、

前記第一のモーダルの情報に基づくベクトルと、前記補正ベクトルとの和を正規化し、

当該正規化により得たベクトルと、圧縮後の前記第一のモーダルの情報に基づくベクトルとの和を正規化する、ことを特徴とする請求項１または２に記載の出力方法。

[請求項4] 前記正規化処理を実施する処理は、

結合後の前記第一のモーダルの情報に基づくベクトルと、圧縮後の前記第一のモーダルの情報に基づくベクトルとの和を正規化する、ことを特徴とする請求項１または２に記載の出力方法。

[請求項5] 前記第一のモーダルと前記第二のモーダルとの組は、画像に関する

モーダルと文書に関するモーダルとの組、画像に関するモーダルと音声に関するモーダルとの組、第一の言語の文書に関するモーダルと第二の言語の文書に関するモーダルとの組のうちいずれかの組である、ことを特徴とする請求項１～４のいずれか一つに記載の出力方法。

[請求項6] 前記第一のモーダルは、前記第二のモーダルと同一である、ことを特徴とする請求項１～４のいずれか一つに記載の出力方法。

[請求項7] 第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとの相関に基づいて、前記第一のモーダルの情報に基づくベクトルを補正する補正ベクトルを生成し、

生成した前記補正ベクトルを、前記第一のモーダルの情報に基づくベクトルに結合し、

所定のルールに従って、結合後の前記第一のモーダルの情報に基づくベクトルを圧縮し、

圧縮後の前記第一のモーダルの情報に基づくベクトルに対して、正規化処理を実施し、

前記正規化処理により得たベクトルを出力する、

処理をコンピュータに実行させることを特徴とする出力プログラム。

[請求項8] 第一のモーダルの情報に基づくベクトルと、第二のモーダルの情報に基づくベクトルとの相関に基づいて、前記第一のモーダルの情報に基づくベクトルを補正する補正ベクトルを生成し、

生成した前記補正ベクトルを、前記第一のモーダルの情報に基づくベクトルに結合し、

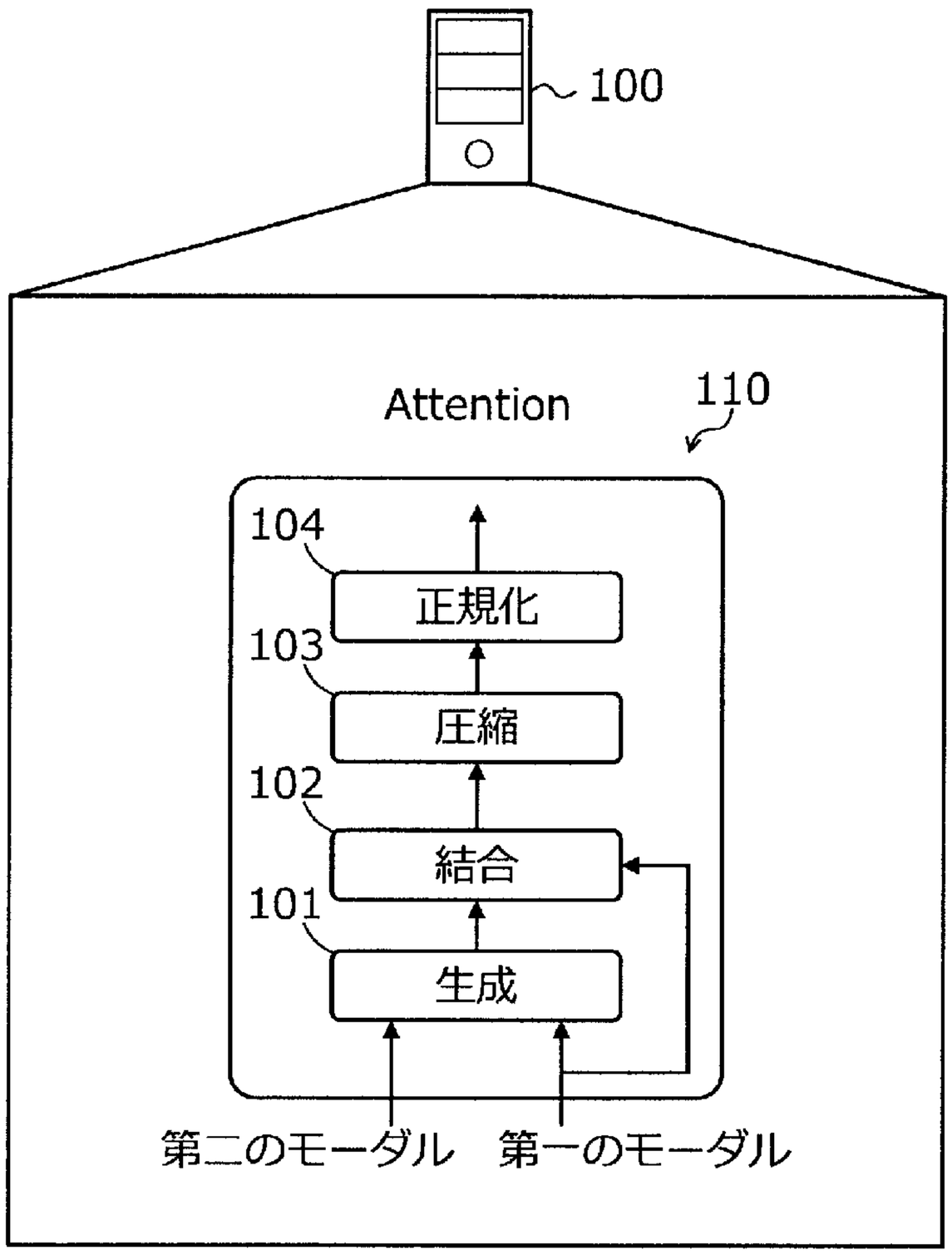
所定のルールに従って、結合後の前記第一のモーダルの情報に基づくベクトルを圧縮し、

圧縮後の前記第一のモーダルの情報に基づくベクトルに対して、正規化処理を実施し、

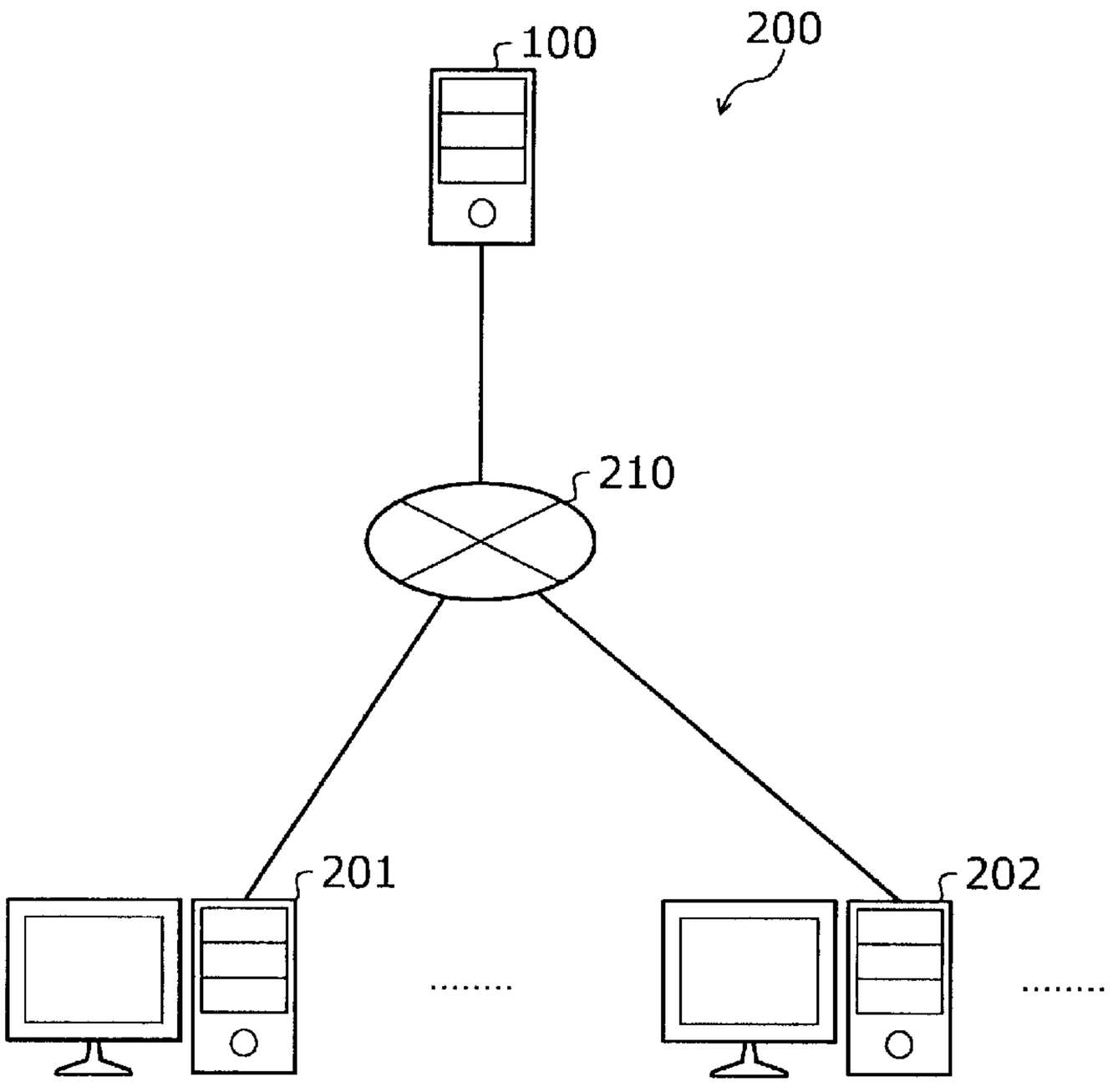
前記正規化処理により得たベクトルを出力する、

制御部を有することを特徴とする出力装置。

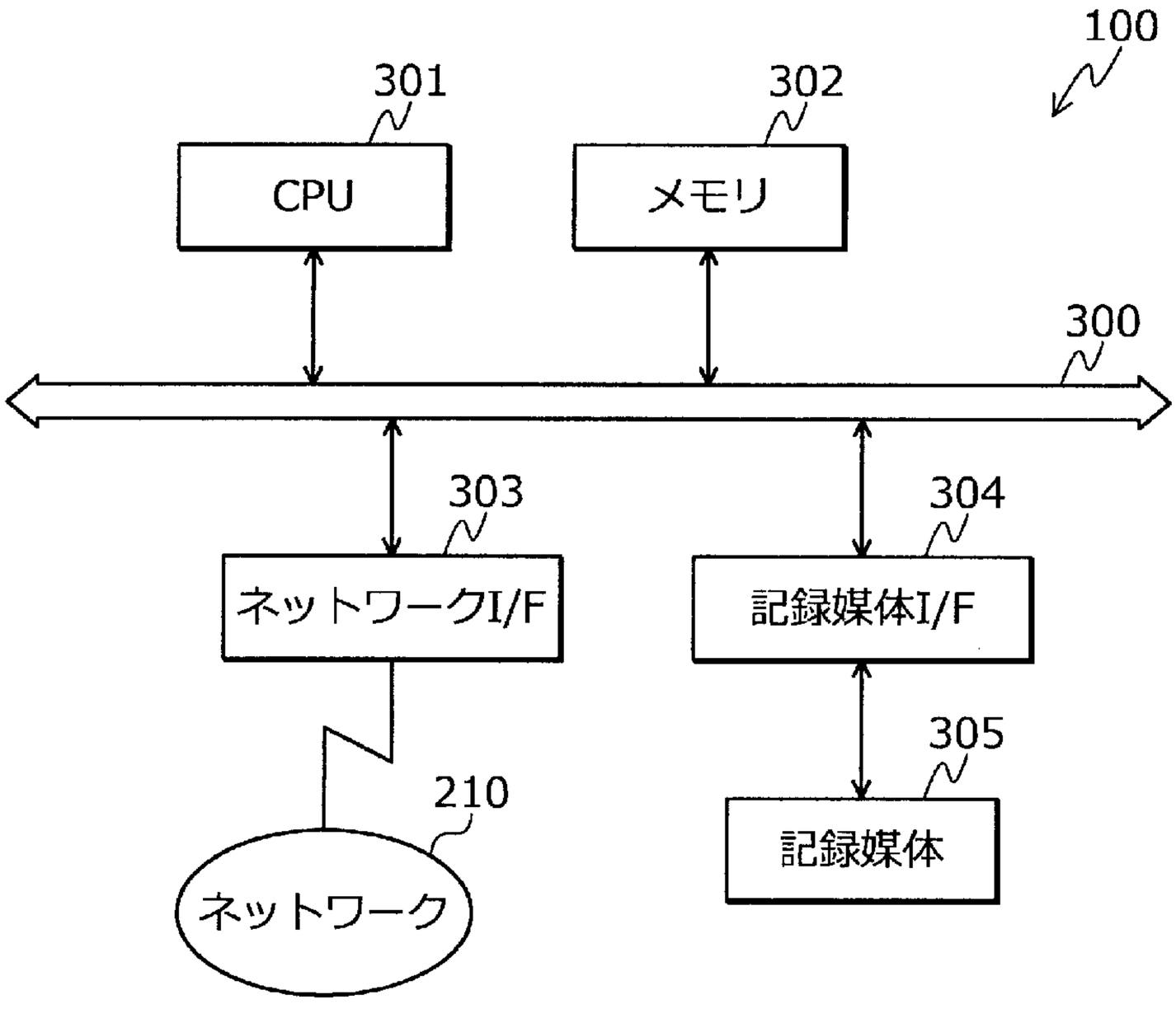
[図1]


100
Attention
110
104
正規化
103
圧縮
102
結合
101
生成
第二のモーダル
第一のモーダル

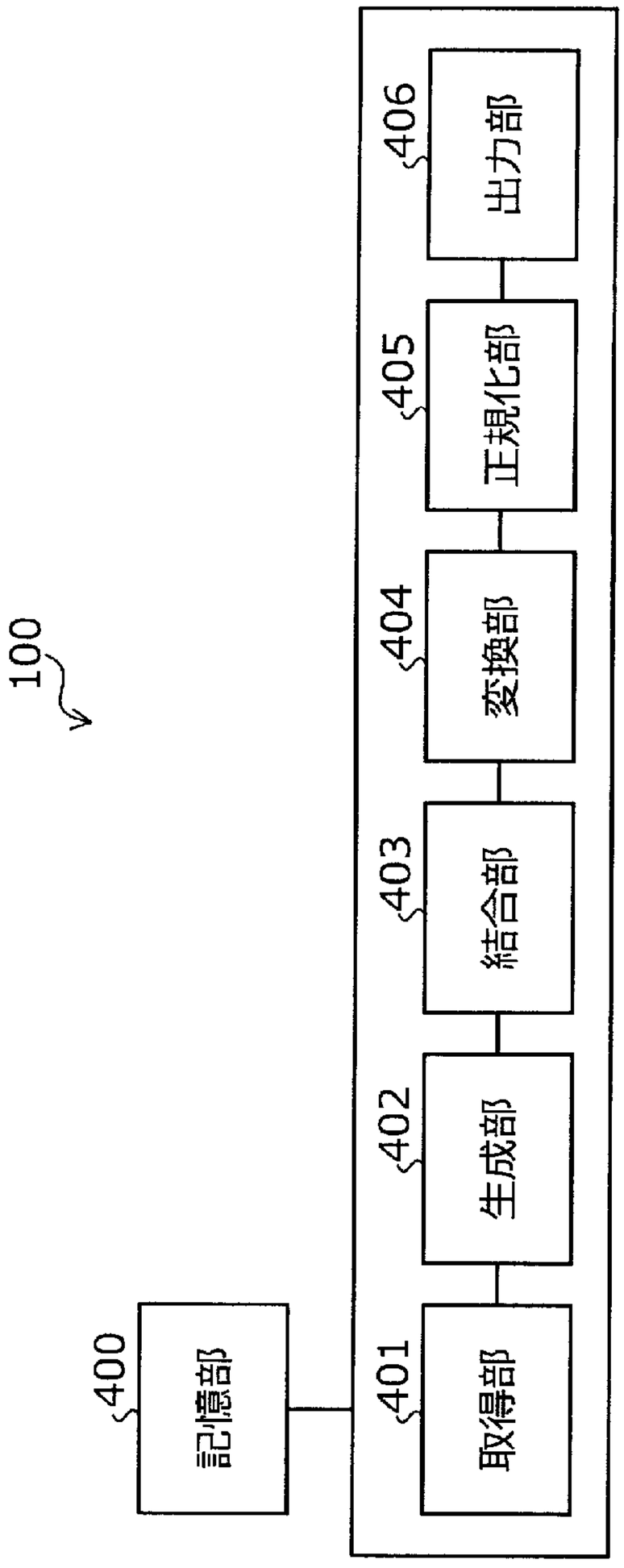
[図2]



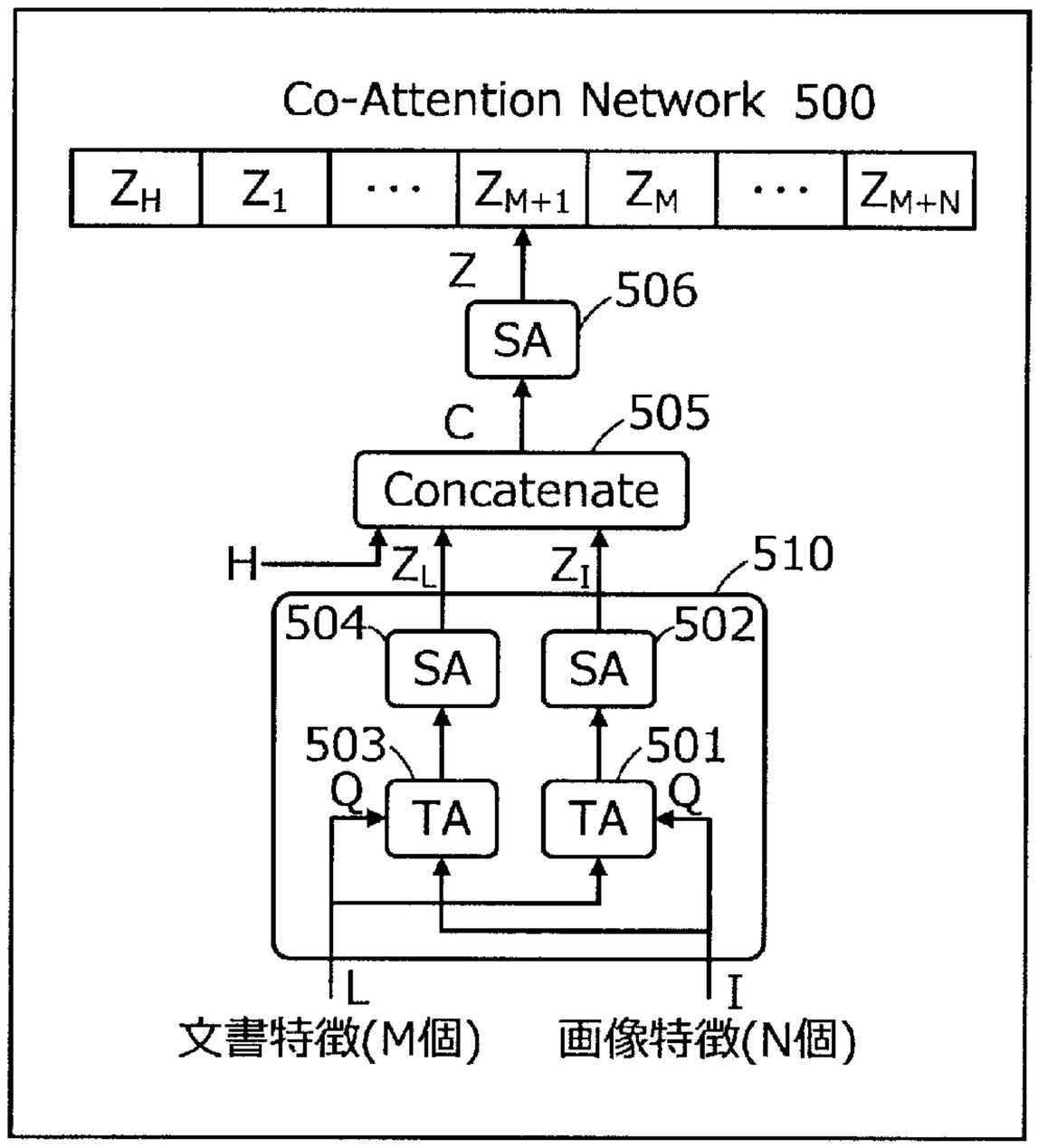
[図3]


100
301
302
CPU
メモリ
300
303
304
ネットワークI/F
記録媒体I/F
305
210
記録媒体
ネットワーク

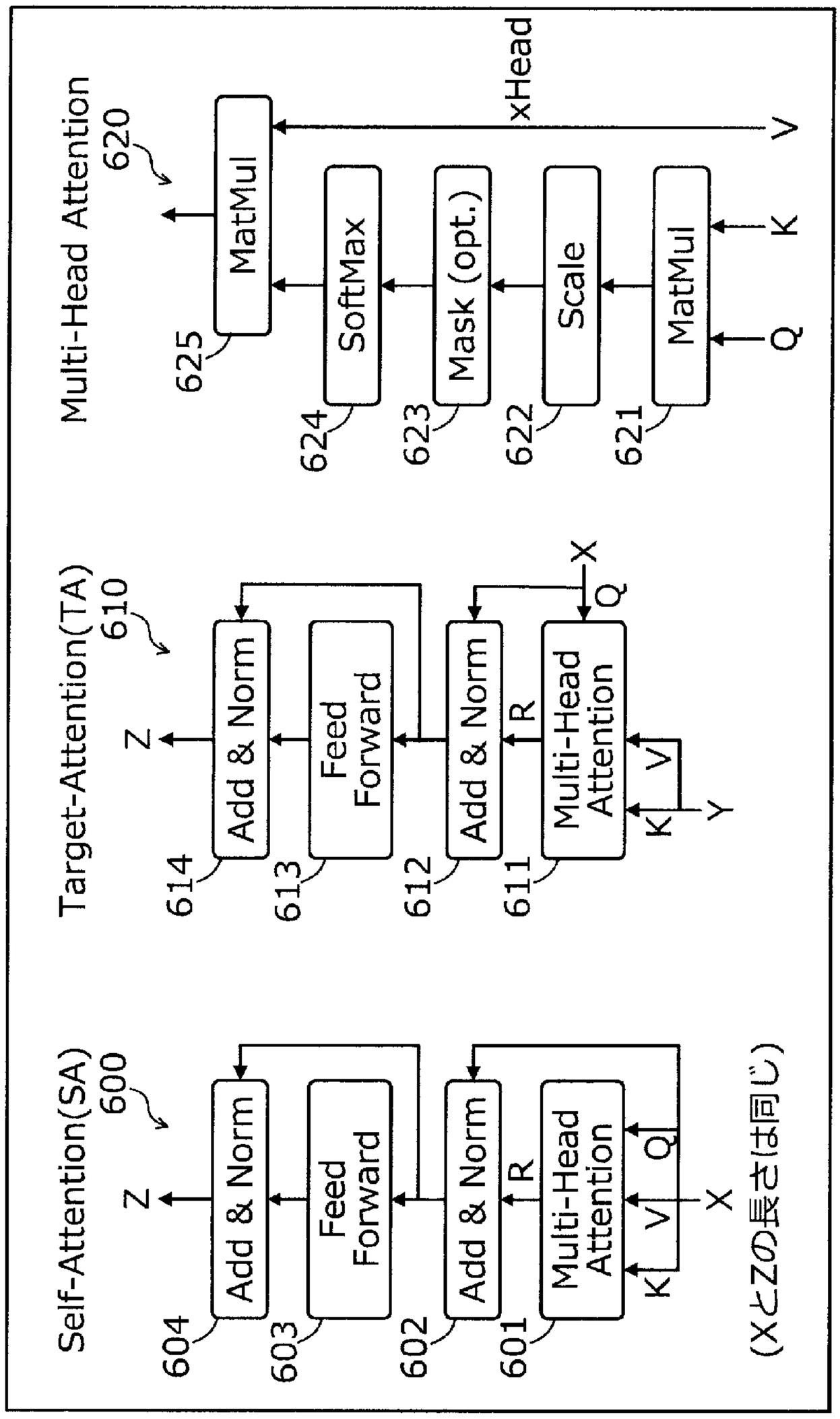
[図4]



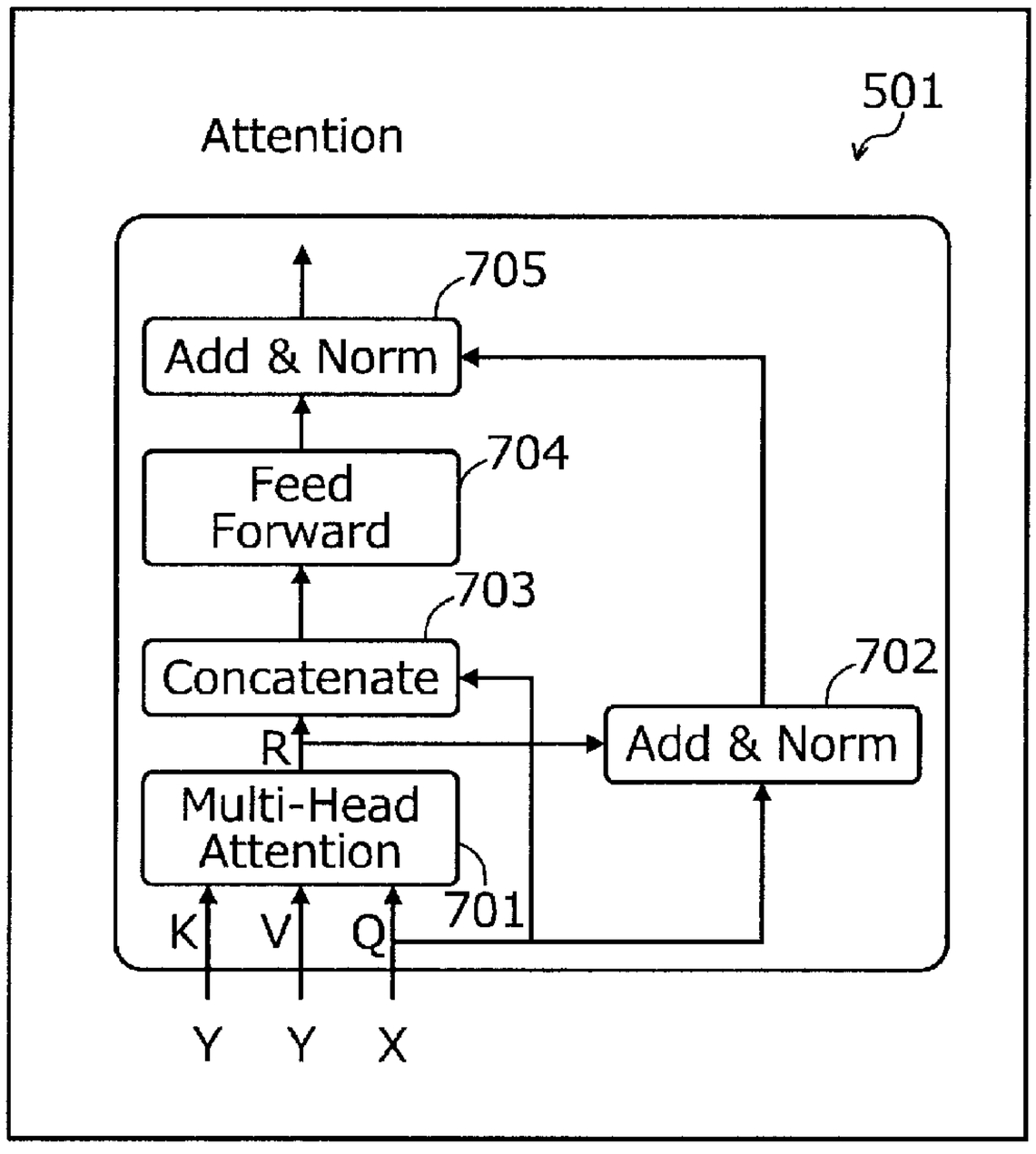
[図5]


Co-Attention Network 500
Z_H
Z_1
・・・
Z_{M+1}
Z_M
・・・
Z_{M+N}
Z
506
SA
C
505
Concatenate
H
Z_L
Z_I
510
504
SA
SA
502
503
Q
TA
TA
Q
501
L
I
文書特徴(M個)
画像特徴(N個)

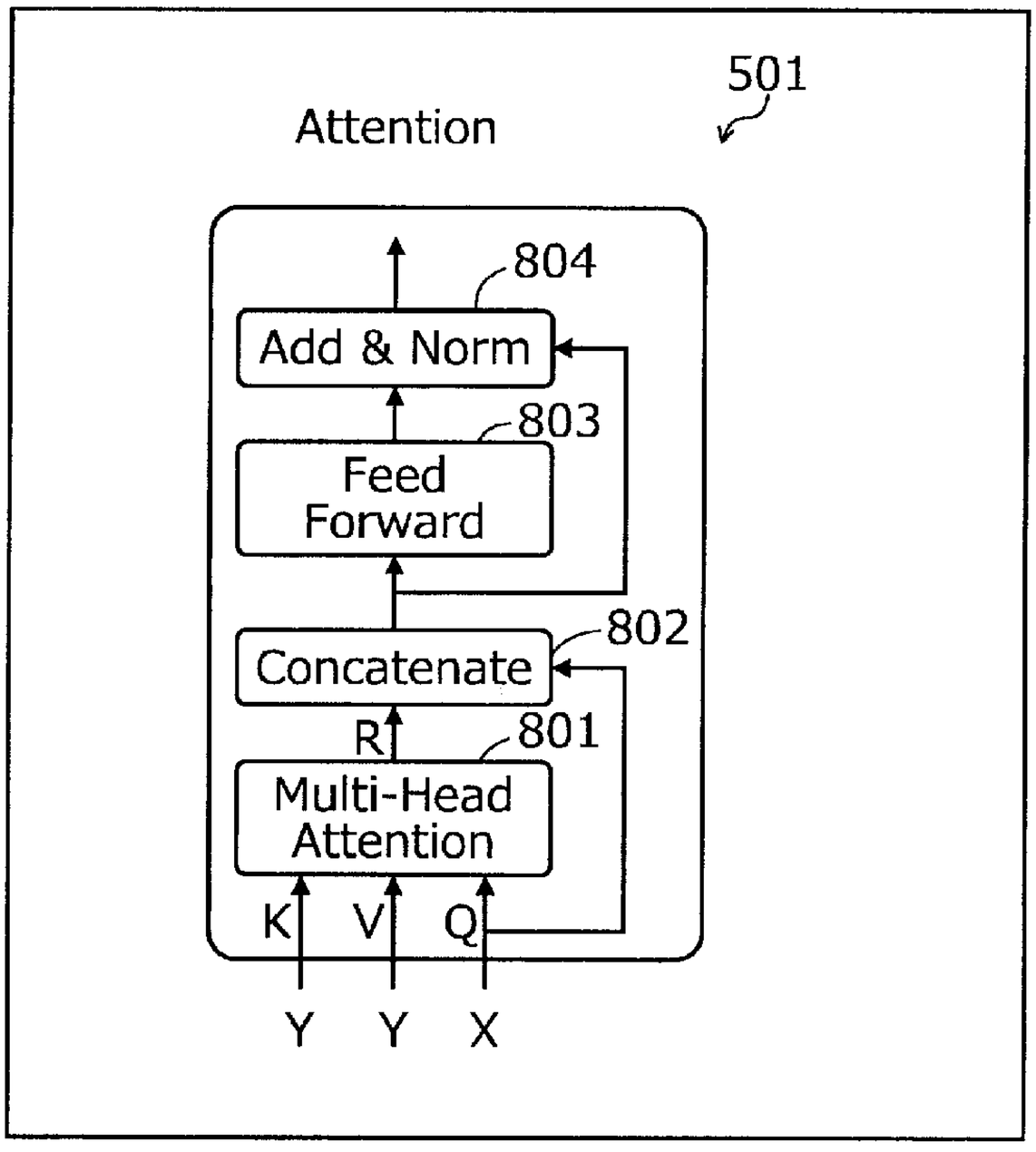
[図6]


Self-Attention(SA)
600
Z
604
Add & Norm
603
Feed Forward
602
Add & Norm
R
601
Multi-Head Attention
K
V
Q
X
(XとZの長さは同じ)
Target-Attention(TA)
610
Z
614
Add & Norm
613
Feed Forward
612
Add & Norm
R
611
Multi-Head Attention
K
V
Q
X
Y
Multi-Head Attention
620
625
MatMul
624
SoftMax
623
Mask (opt.)
622
Scale
621
MatMul
K
Q
V
xHead

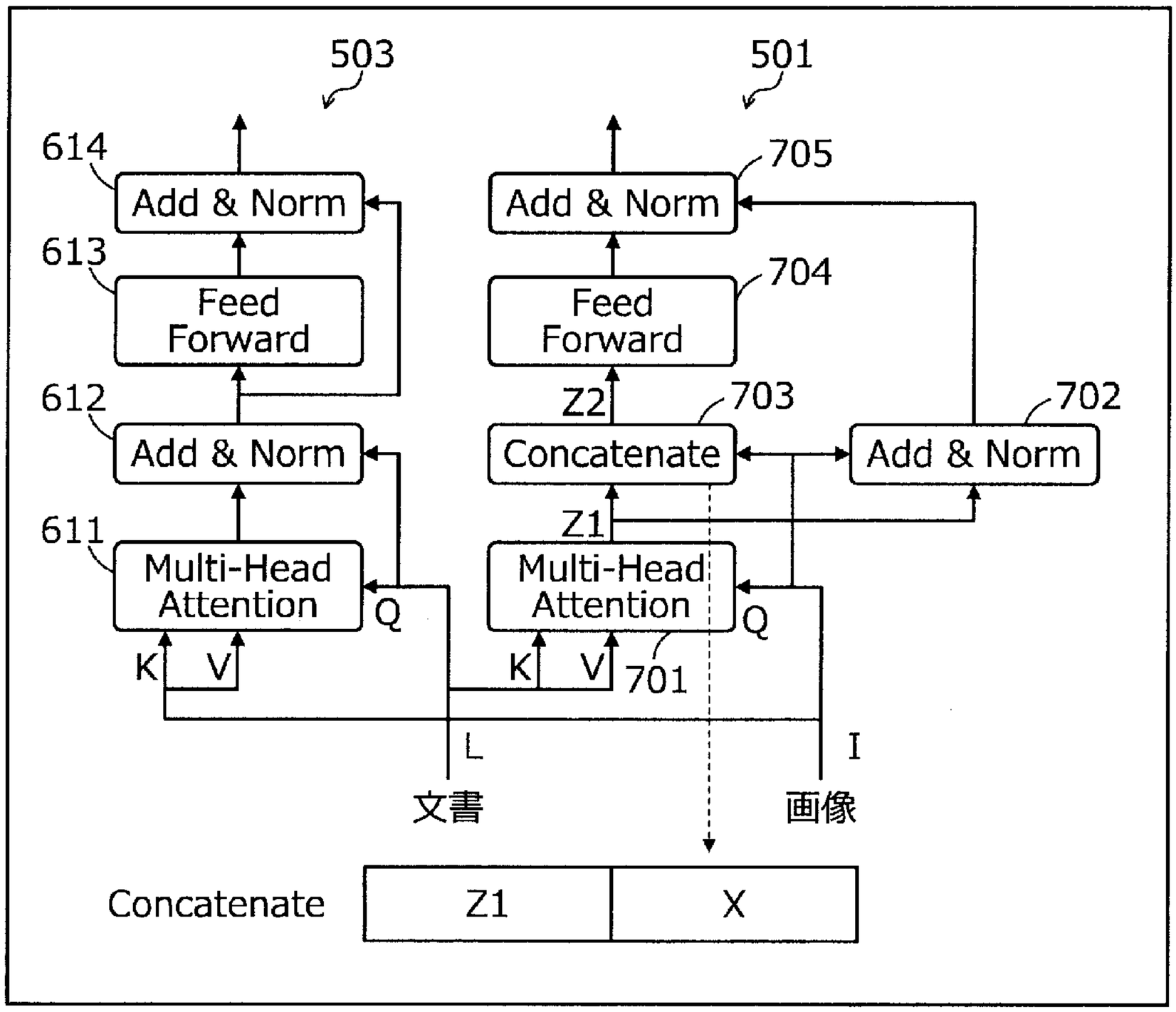
[図7]


Attention
501
705
Add & Norm
704
Feed Forward
703
Concatenate
702
R
Add & Norm
Multi-Head Attention
701
K
V
Q
Y
Y
X

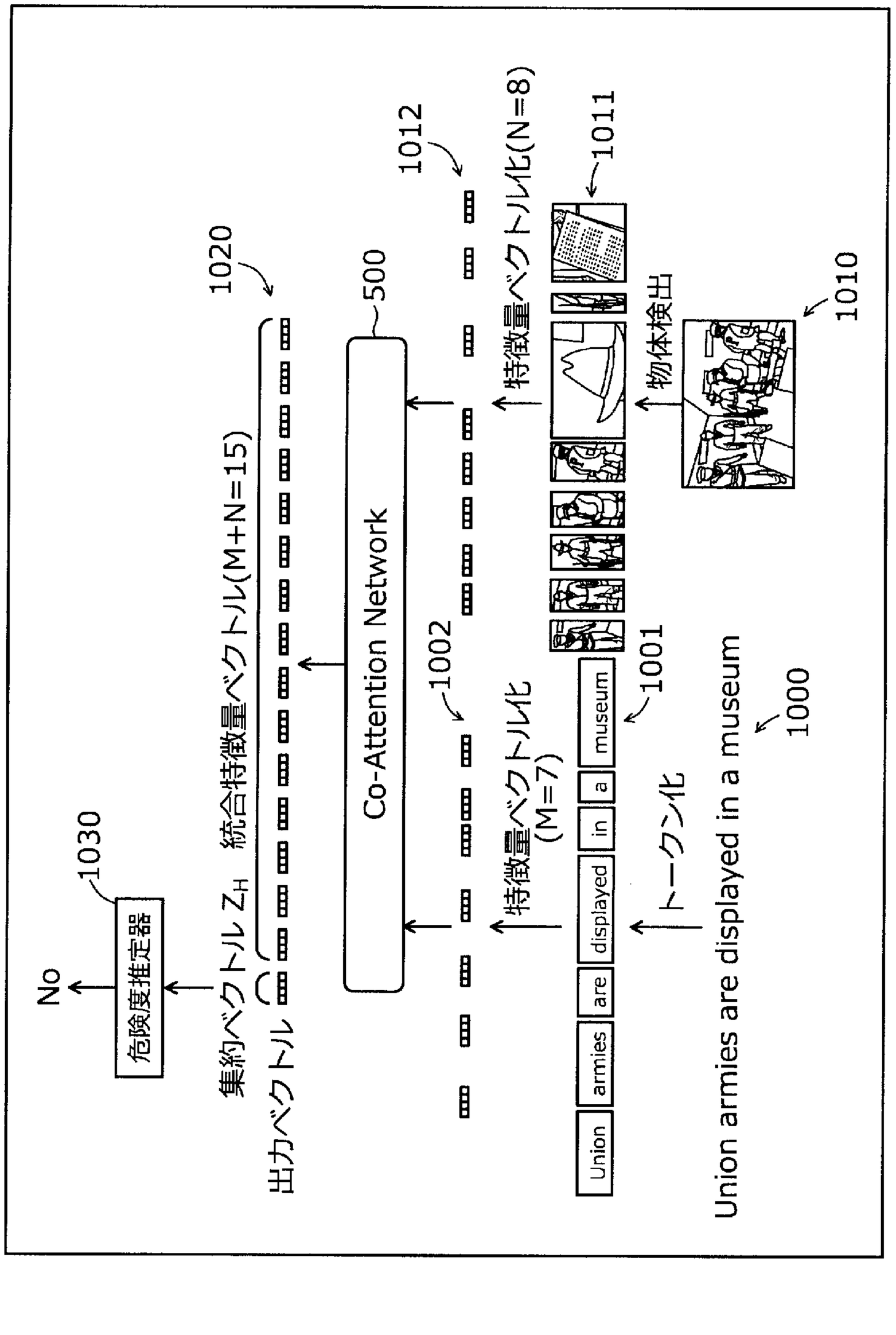
[図8]


Attention
501
804
Add & Norm
803
Feed Forward
802
Concatenate
R
801
Multi-Head Attention
K
V
Q
Y
Y
X

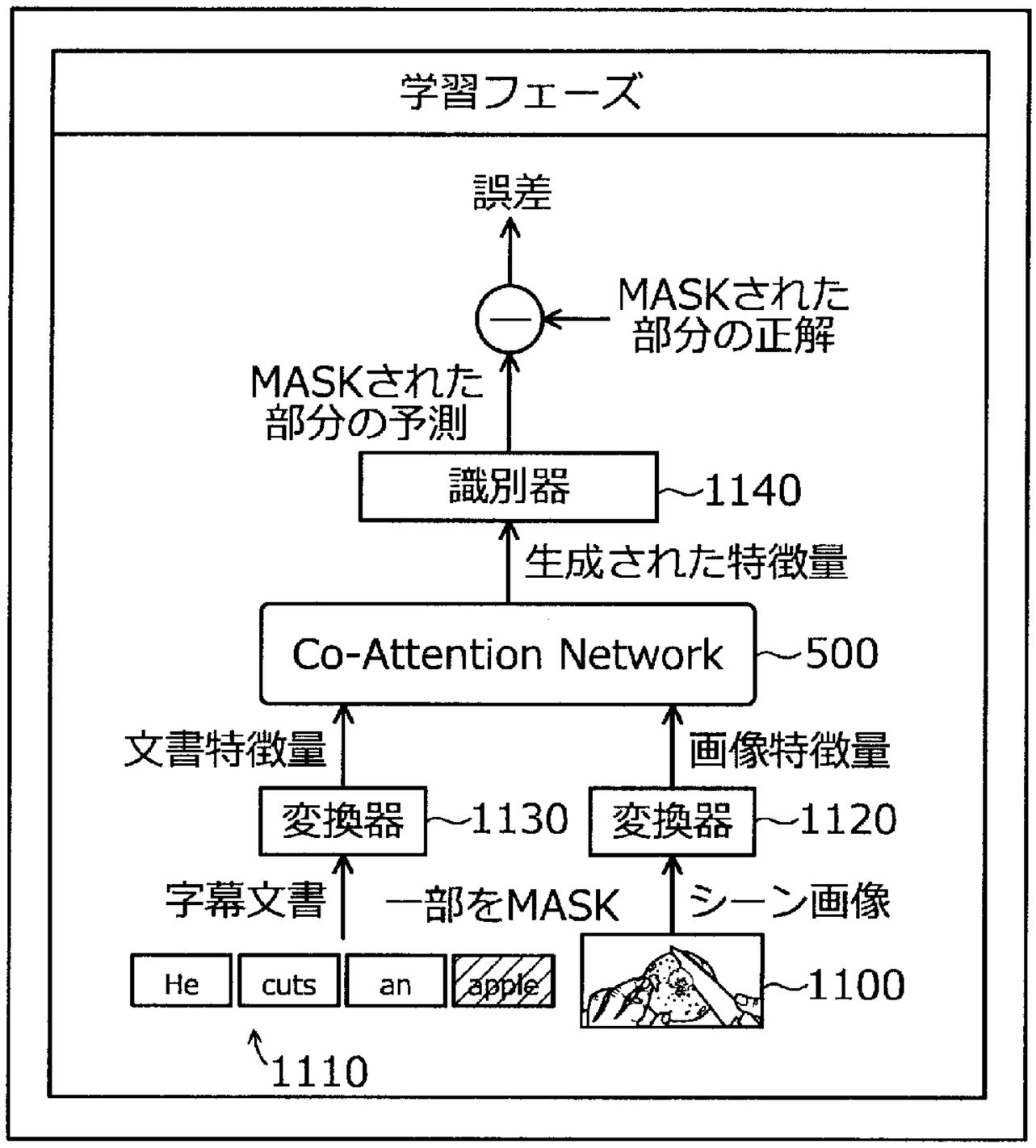
[図9]


503
501
614
Add & Norm
705
Add & Norm
613
Feed Forward
704
Feed Forward
612
Add & Norm
Z2
703
Concatenate
702
Add & Norm
611
Multi-Head Attention
Q
K
V
Z1
Multi-Head Attention
Q
K
V
701
L
I
文書
画像
Concatenate
Z1
X

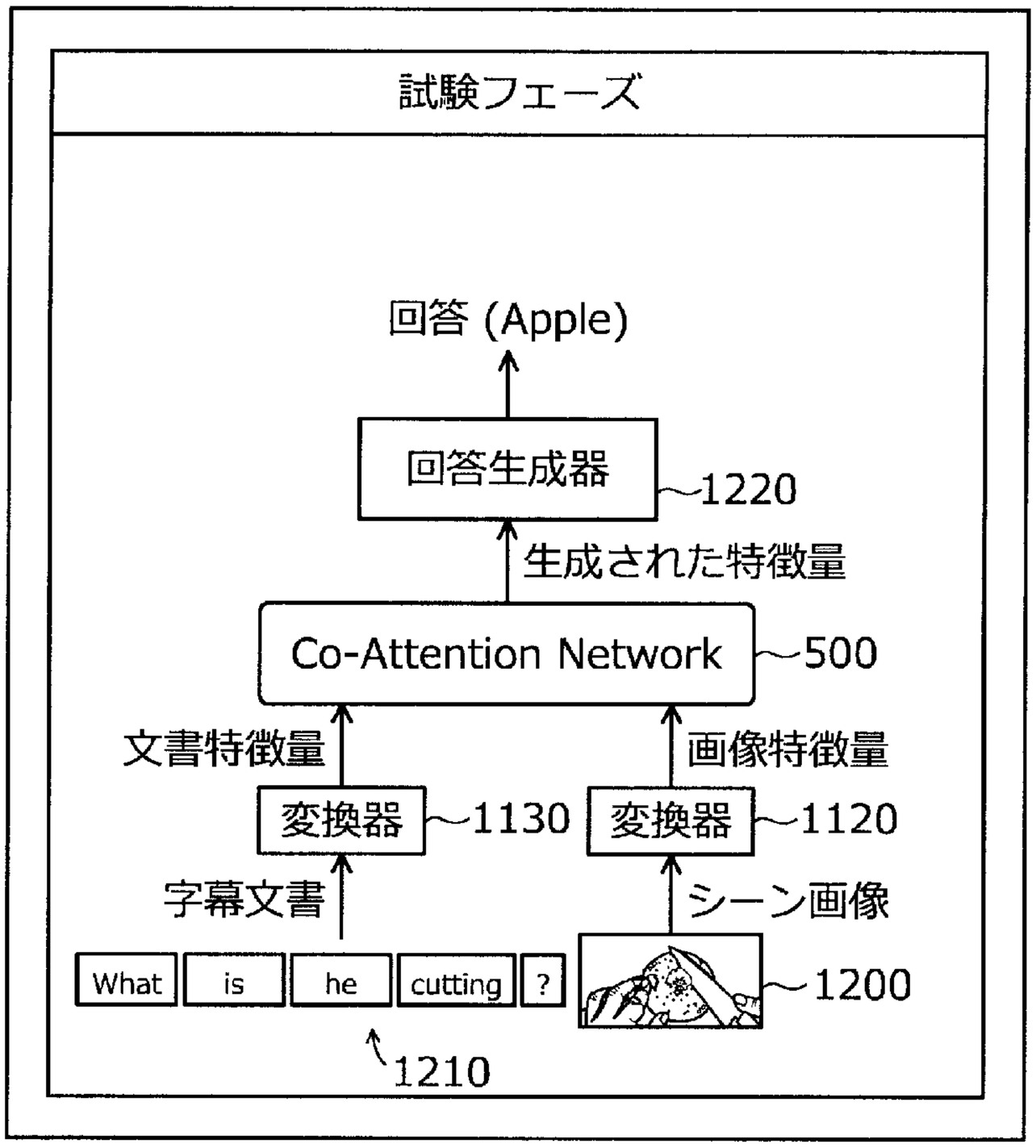
[図10]


1012
1011
特徴量ベクトル化(N=8)
1020
特徴量ベクトル(M+N=15)
500
Co-Attention Network
1002
特徴量ベクトル化
(M=7)
物体検出
museum
1001
1010
a museum
1000
No
1030
危険度推定器
集約ベクトル Z_H 統合特徴
出力ベクトル
特徴量ベクトル
(M=7)
Union armies are displayed in a
トークン化

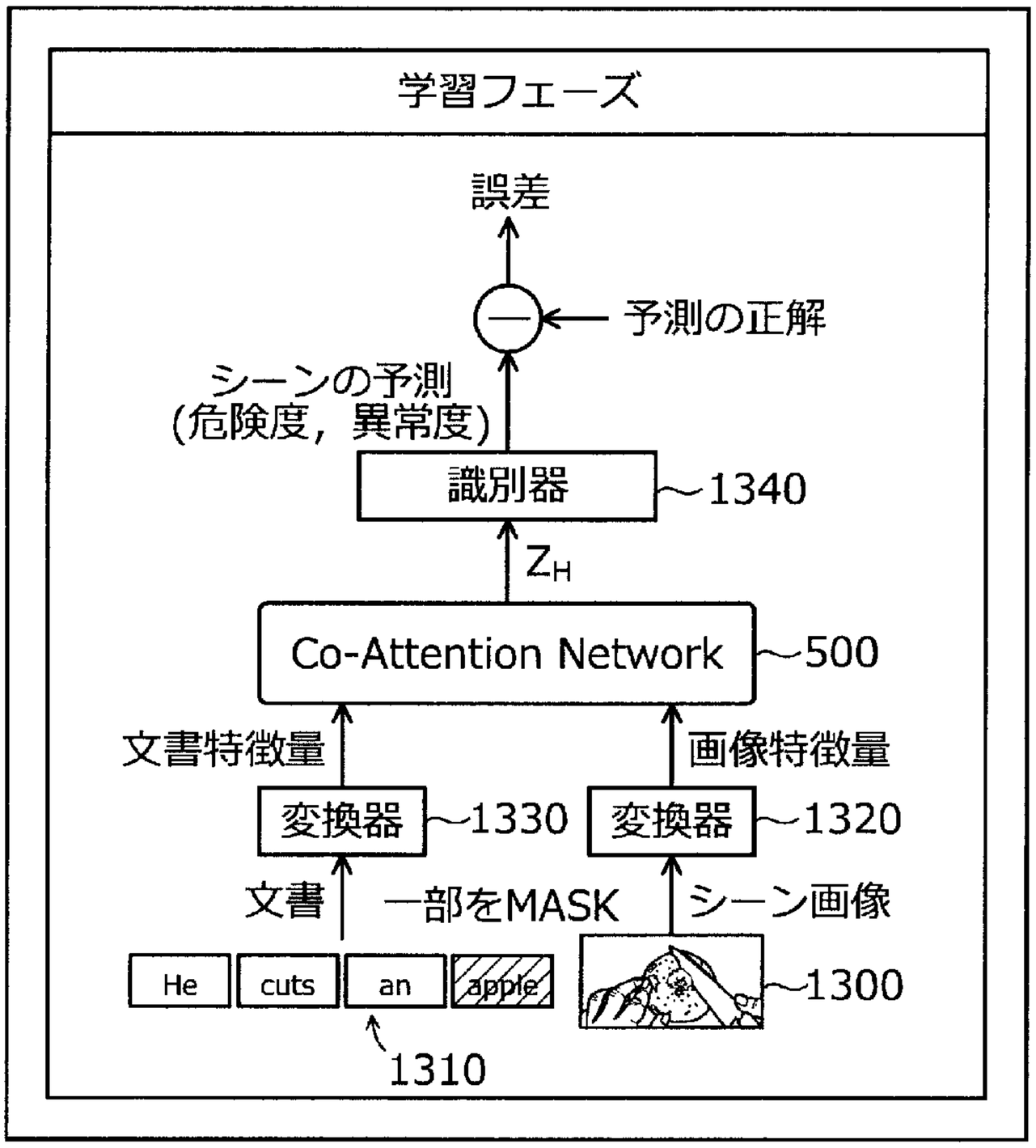
[図11]


学習フェーズ
誤差
MASKされた
部分の正解
MASKされた
部分の予測
識別器
1140
生成された特徴量
Co-Attention Network
500
文書特徴量
画像特徴量
変換器
1130
変換器
1120
字幕文書
一部をMASK
シーン画像
He
cuts
an
apple
1100
1110

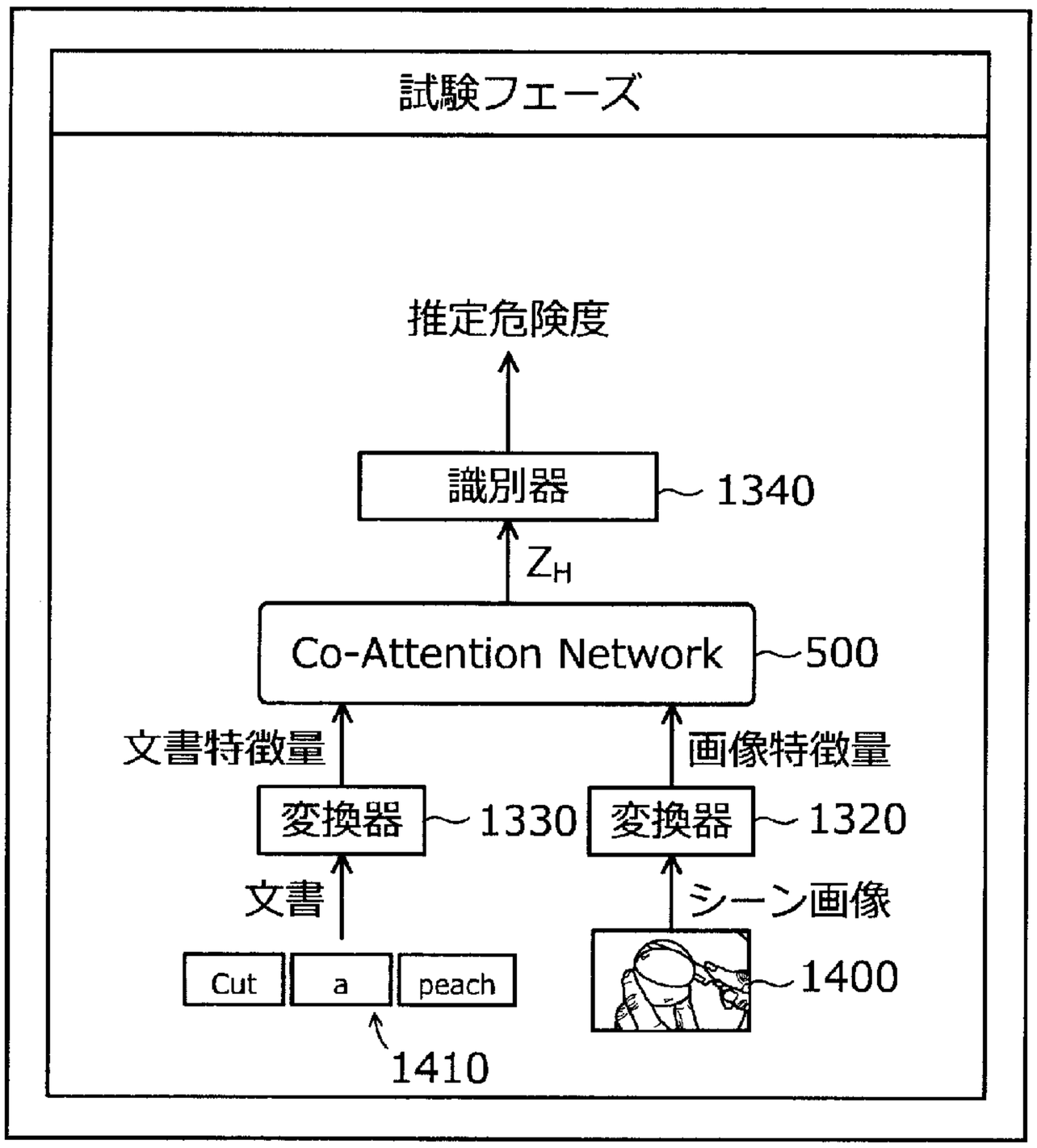
[図12]


試験フェーズ
回答 (Apple)
回答生成器
1220
生成された特徴量
Co-Attention Network
500
文書特徴量
画像特徴量
変換器
1130
変換器
1120
字幕文書
シーン画像
What
is
he
cutting
?
1200
1210

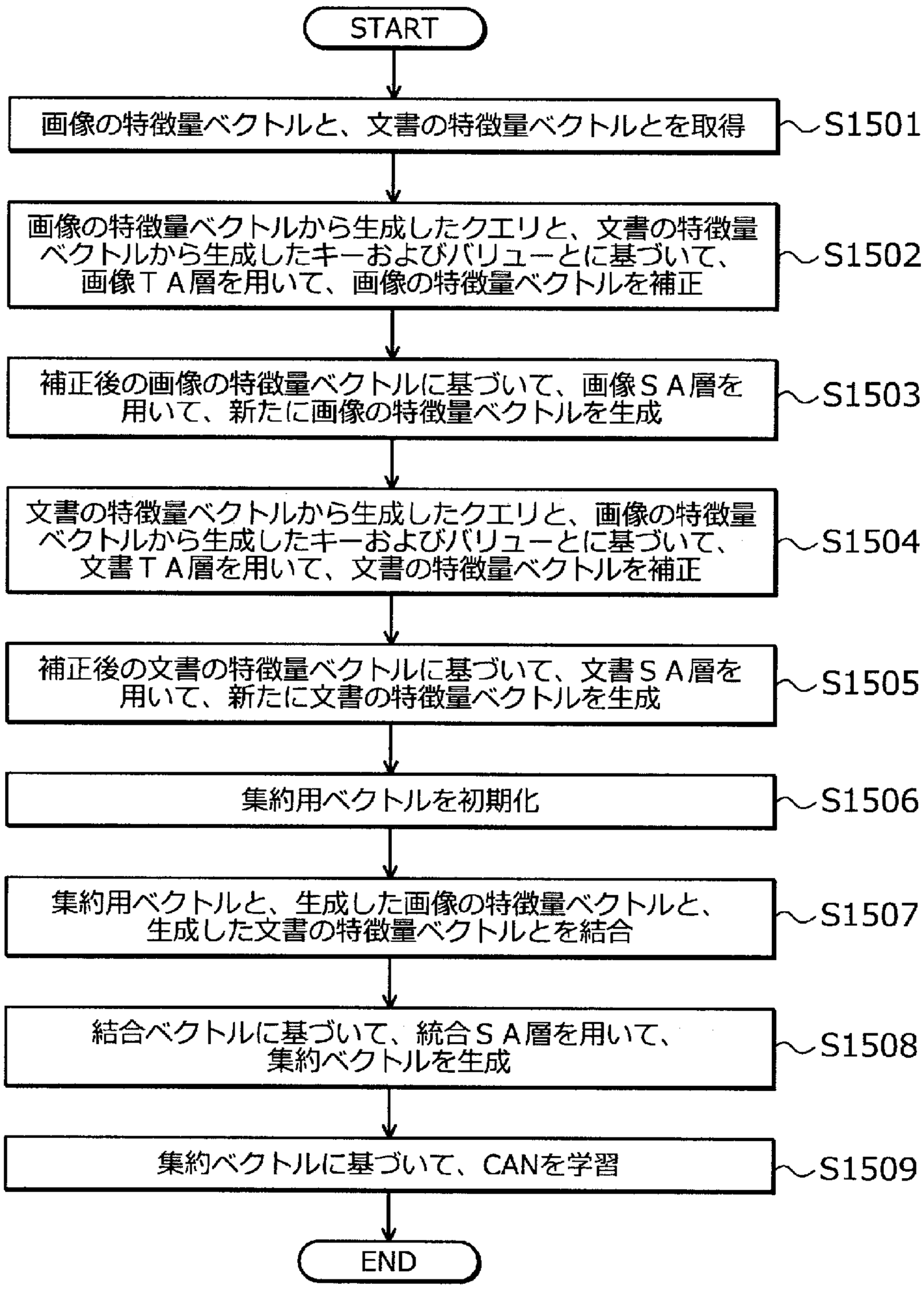
[図13]


学習フェーズ
誤差
予測の正解
シーンの予測
(危険度，異常度)
識別器
1340
Z_H
Co-Attention Network
500
文書特徴量
画像特徴量
変換器
1330
変換器
1320
文書
一部をMASK
シーン画像
He
cuts
an
apple
1300
1310

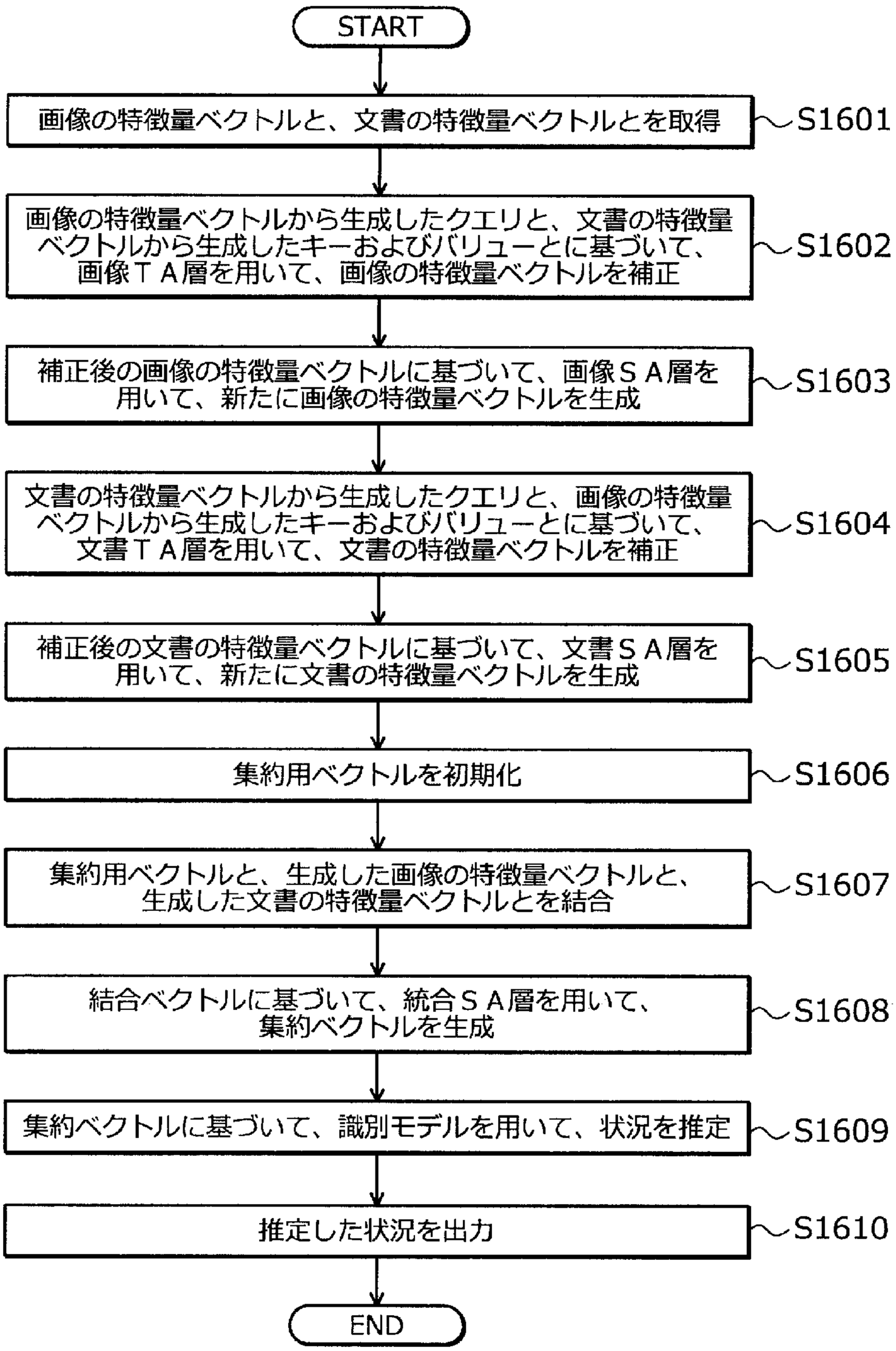
[図14]


試験フェーズ
推定危険度
識別器
1340
Z_H
Co-Attention Network
500
文書特徴量
画像特徴量
変換器
1330
変換器
1320
文書
シーン画像
Cut
a
peach
1400
1410

[図15]


START
画像の特徴量ベクトルと、文書の特徴量ベクトルとを取得
S1501
画像の特徴量ベクトルから生成したクエリと、文書の特徴量ベクトルから生成したキーおよびバリューとに基づいて、画像ＴＡ層を用いて、画像の特徴量ベクトルを補正
S1502
補正後の画像の特徴量ベクトルに基づいて、画像ＳＡ層を用いて、新たに画像の特徴量ベクトルを生成
S1503
文書の特徴量ベクトルから生成したクエリと、画像の特徴量ベクトルから生成したキーおよびバリューとに基づいて、文書ＴＡ層を用いて、文書の特徴量ベクトルを補正
S1504
補正後の文書の特徴量ベクトルに基づいて、文書ＳＡ層を用いて、新たに文書の特徴量ベクトルを生成
S1505
集約用ベクトルを初期化
S1506
集約用ベクトルと、生成した画像の特徴量ベクトルと、生成した文書の特徴量ベクトルとを結合
S1507
結合ベクトルに基づいて、統合ＳＡ層を用いて、集約ベクトルを生成
S1508
集約ベクトルに基づいて、CANを学習
S1509
END

[図16]


START
画像の特徴量ベクトルと、文書の特徴量ベクトルとを取得
S1601
画像の特徴量ベクトルから生成したクエリと、文書の特徴量ベクトルから生成したキーおよびバリューとに基づいて、画像ＴＡ層を用いて、画像の特徴量ベクトルを補正
S1602
補正後の画像の特徴量ベクトルに基づいて、画像ＳＡ層を用いて、新たに画像の特徴量ベクトルを生成
S1603
文書の特徴量ベクトルから生成したクエリと、画像の特徴量ベクトルから生成したキーおよびバリューとに基づいて、文書ＴＡ層を用いて、文書の特徴量ベクトルを補正
S1604
補正後の文書の特徴量ベクトルに基づいて、文書ＳＡ層を用いて、新たに文書の特徴量ベクトルを生成
S1605
集約用ベクトルを初期化
S1606
集約用ベクトルと、生成した画像の特徴量ベクトルと、生成した文書の特徴量ベクトルとを結合
S1607
結合ベクトルに基づいて、統合ＳＡ層を用いて、集約ベクトルを生成
S1608
集約ベクトルに基づいて、識別モデルを用いて、状況を推定
S1609
推定した状況を出力
S1610
END

[図17]


START
画像の特徴量ベクトルと、文書の特徴量ベクトルとを取得
S1701
画像の特徴量ベクトルからベクトルＱｕｅｒｙを生成
S1702
文書の特徴量ベクトルからベクトルｋｅｙとベクトルＶａｌｕｅを生成
S1703
ベクトルＱｕｅｒｙと、ベクトルｋｅｙとの内積を算出
S1704
内積のｓｏｆｔｍａｘによりベクトルＡｔｔを生成
S1705
ベクトルＡｔｔとベクトルＶａｌｕｅとの内積によりベクトルＲを生成
S1706
ベクトルＲとベクトルＸとを結合したベクトルＸ'を生成
S1707
多層ニューラルネットワークにより、ベクトルＸ'を、ベクトルＸと同じ次元に圧縮し、ベクトルＸ"を生成
S1708
ベクトルＲとベクトルＸとを用いて、ベクトルＸ"を正規化し、正規化後のベクトルを取得
S1709
正規化後のベクトルを出力
S1710
END

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2019/044770

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl. G06N20/00(2019.01)i

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)
Int.Cl. G06N3/00-99/00

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 1971-2019
Registered utility model specifications of Japan 1996-2019
Published registered utility model applications of Japan 1994-2019

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.
A	LU, J. S. et al., ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, arXiv.org [online], 06 August 2019, pp. 1-11, [retrieved on 13 December 2019], Internet: <URL:https://arxiv.org/pdf/1908.02265v1.pdf>, particularly, chapter 2	1-8

☒ 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	Date of mailing of the international search report
18.12.2019	07.01.2020
Name and mailing address of the ISA/ Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2019/044770

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	NGUYEN, D. K. et al., Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering, [online], 2018, pp. 6087-6096, [retrieved on 13 December 2019], Internet: <URL:http://openaccess.thecvf.com/content_cvpr_2018/html/Nguyen_Improved_Fusion_of_CVPR_2018_paper.html>, particularly, chapters 1-3	1-8

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１９／０４４７７０

A.　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06N20/00(2019.01)i

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））

Int.Cl.　G06N3/00-99/00

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

日本国実用新案公報　１９２２－１９９６年
日本国公開実用新案公報　１９７１－２０１９年
日本国実用新案登録公報　１９９６－２０１９年
日本国登録実用新案公報　１９９４－２０１９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
A	LU, Jiasen, et al., ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, arXiv.org [online], 2019.08.06, pp.1-11, [検索日 2019.12.13], インターネット:<URL:https://arxiv.org/pdf/1908.02265v1.pdf>, 特に第２章	1-8

☑　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。

＊　引用文献のカテゴリー
「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「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矛盾するものではなく、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	国際調査報告の発送日
１８．１２．２０１９	０７．０１．２０２０

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｂ　１５９９
多賀　実
電話番号　０３－３５８１－１１０１　内線　３５４５

Ｃ（続き）．　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
A	NGUYEN, Duy-Kien, et al., Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering, [online], 2018, pp.6087-6096, [検索日 2019.12.13], インターネット：<URL：http://openaccess.thecvf.com/content_cvpr_2018/html/Nguyen_Improved_Fusion_of_CVPR_2018_paper.html>, 特に第１章-第３章	1-8