

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7666832号
(P7666832)

(45)発行日 令和7年4月22日(2025.4.22)

(24)登録日 令和7年4月14日(2025.4.14)

(51)国際特許分類

FI

G 0 6 N 20/00 (2019.01)

G 0 6 N 20/00

G 0 6 N 3/08 (2023.01)

G 0 6 N 3/08

G 1 0 L 15/16 (2006.01)

G 1 0 L 15/16

請求項の数 40 (全27頁)

(21)出願番号	特願2021-68894(P2021-68894)	(73)特許権者	390019839
(22)出願日	令和3年4月15日(2021.4.15)		三星電子株式会社
(65)公開番号	特開2021-190101(P2021-190101 A)		Samsung Electronics Co., Ltd.
(43)公開日	令和3年12月13日(2021.12.13)		大韓民国京畿道水原市靈通区三星路129
審査請求日	令和6年3月6日(2024.3.6)		129, Samsung-ro, Yeongtong-gu, Suwon-si
(31)優先権主張番号	10-2020-0062450		, Gyeonggi-do, Republic of Korea
(32)優先日	令和2年5月25日(2020.5.25)	(74)代理人	100107766
(33)優先権主張国・地域又は機関	韓国(KR)		弁理士 伊東 忠重
		(74)代理人	100070150
			弁理士 伊東 忠彦
		(74)代理人	100135079

最終頁に続く

(54)【発明の名称】 アテンション基盤シーケンスツースーケンスモデルの性能向上方法及び装置

(57)【特許請求の範囲】

【請求項1】

電子装置で実施されるアテンション基盤シーケンスツースーケンスモデルの性能向上方法であって、

前記電子装置のプロセッサによって、アテンション基盤シーケンスツースーケンスモデルに基づいて、入力シーケンスに対応する出力シーケンスを決定するステップと、

前記プロセッサによって、それぞれのアテンション重み行列を作り出すように、それぞれが構成されている複数のアテンションヘッドのうち、1つ以上のターゲットアテンションヘッドを選択するステップと、

前記プロセッサによって、前記出力シーケンスを構成する出力トークンのうち、前記ターゲットアテンションヘッドに基づいて、1つ以上のエラー出力トークンを検出するステップと、

前記プロセッサによって、前記1つ以上のエラー出力トークンに基づいて、前記出力シーケンスを補正するステップと、

を含む、性能向上方法。

【請求項2】

前記選択するステップは、

前記ターゲットアテンションヘッドに対応するターゲットアテンション重み行列になるように学習された特定のアテンション重み行列を作り出すアテンションヘッドを、前記ターゲットアテンションヘッドとして選択する、ステップを含む、

請求項 1 に記載の性能向上方法。

【請求項 3】

前記特定のアテンション重み行列は、予め決定された形態を有するガイド重み行列に基づいて学習される、

請求項 2 に記載の性能向上方法。

【請求項 4】

前記ガイド重み行列は、出力シーケンス長さ、入力フレーム長さ、開始シフト、終了シフト、及びディフュージョン比率のうち少なくとも 1 つに基づいて、決定される、

請求項 3 に記載の性能向上方法。

【請求項 5】

前記特定のアテンション重み行列は、ステップごとにアテンション重みの分布が異なるように学習される、

請求項 2 に記載の性能向上方法。

【請求項 6】

前記特定のアテンション重み行列は、以前ステップのアテンション重み累積和に基づいて、現在ステップのアテンション重みが決定されるように学習される、

請求項 2 に記載の性能向上方法。

【請求項 7】

前記選択するステップは、

予め決定された目的に最も適するアテンション重み行列を生成するアテンションヘッドを、前記ターゲットアテンションヘッドとして選択する、ステップを含む、

請求項 1 に記載の性能向上方法。

【請求項 8】

前記選択するステップは、

予め決定された目的による予め決定された形態を有するガイド重み行列に基づいて、前記ターゲットアテンションヘッドを選択する、ステップを含む、

請求項 1 に記載の性能向上方法。

【請求項 9】

前記選択するステップは、

前記アテンション基盤シーケンスツリーモデルが単調性質を有する場合に、前記複数のアテンションヘッドが作り出すアテンション重み行列に対して単調回帰分析を行って、前記ターゲットアテンションヘッドを選択する、ステップを含む、

請求項 1 に記載の性能向上方法。

【請求項 10】

前記選択するステップは、

前記複数のアテンションヘッドが作り出すアテンション重み行列のエントロピーに基づいて、前記ターゲットアテンションヘッドを選択する、ステップを含む、

請求項 1 に記載の性能向上方法。

【請求項 11】

前記エントロピーに基づいて、前記ターゲットアテンションヘッドを選択するステップは、

前記アテンション重み行列のうち、前記エントロピーが最大のアテンション重み行列を作り出すアテンションヘッドを、前記ターゲットアテンションヘッドとして選択する、ステップを含む、

請求項 10 に記載の性能向上方法。

【請求項 12】

前記エントロピーに基づいて、前記ターゲットアテンションヘッドを選択するステップは、

カルバック・ライブラー情報量に基づいて、前記ターゲットアテンションヘッドを選択する、ステップを含む、

10

20

30

40

50

請求項 10 に記載の性能向上方法。

【請求項 13】

前記選択するステップは、

アテンション重み行列の各行の分布間の距離が最も遠いアテンション重み行列を作り出すアテンションヘッドを、前記ターゲットアテンションヘッドとして選択する、ステップを含む、

請求項 1 に記載の性能向上方法。

【請求項 14】

前記検出するステップは、

前記ターゲットアテンションヘッドの前記入力シーケンスと前記出力シーケンスとの間のアテンション重みのうち、前記ターゲットアテンションヘッドのアテンション重みとガイド重み行列との差が閾値以上である、1つ以上のエラーアテンション重みを検出するステップと、

前記1つ以上のエラーアテンション重みに対応する出力トークンを、前記1つ以上のエラー出力トークンとして決定するステップと、

を含む、請求項 1 に記載の性能向上方法。

【請求項 15】

前記検出するステップは、

前記ターゲットアテンションヘッドの前記入力シーケンスと前記出力シーケンスとの間のアテンション重みのうち、以前ステップのアテンション重みと閾値以上の類似度を有する、1つ以上のエラーアテンション重みを検出するステップと、

前記1つ以上のエラーアテンション重みに対応する出力トークンを、前記1つ以上のエラー出力トークンとして決定するステップと、

を含む、請求項 1 に記載の性能向上方法。

【請求項 16】

前記補正するステップは、

前記1つ以上のエラー出力トークンを前記出力シーケンスから取り除くステップを含む、請求項 1 に記載の性能向上方法。

【請求項 17】

前記補正するステップは、

前記1つ以上のエラー出力トークンを取り除いた他の出力トークン候補のうち、次の入力トークンを決定する、ステップを含む、

請求項 1 に記載の性能向上方法。

【請求項 18】

前記性能向上方法は、さらに、

前記1つ以上のエラー出力トークンが出力されたステップの入力トークンを、前記次の入力トークンとして決定する、ステップを含む、

請求項 17 に記載の性能向上方法。

【請求項 19】

前記複数のアテンションヘッドの数は、前記アテンション基盤シーケンスツリーシーケンスモデルのアテンションレイヤー数とデコーダーレイヤー数の積に該当する、

請求項 1 に記載の性能向上方法。

【請求項 20】

コンピュータで読取り可能な記憶媒体に保管された、インストラクションを含むコンピュータプログラムであり、

前記インストラクションが実行されると、前記コンピュータに、請求項 1 乃至 19 いずれか一項に記載の方法を実施させる、

コンピュータプログラム。

【請求項 21】

プロセッサを含む、アテンション基盤シーケンスツリーシーケンスモデルの電子装置であ

10

20

30

40

50

って、前記プロセッサは、

アテンション基盤シーケンスツースーケンスモデルに基づいて、入力シーケンスに対応する出力シーケンスを決定し、
それぞれのアテンション重み行列を作り出すように、それぞれが構成されている複数のアテンションヘッドのうち1つ以上のターゲットアテンションヘッドを選択し、

前記出力シーケンスを構成する出力トークンのうち、前記ターゲットアテンションヘッドに基づいて、1つ以上のエラー出力トークンを検出し、

前記1つ以上のエラー出力トークンに基づいて、前記出力シーケンスを補正する、
電子装置。

【請求項22】

前記プロセッサは、

前記ターゲットアテンションヘッドに対応するターゲットアテンション重み行列になるよう学習された特定のアテンション重み行列を作り出すアテンションヘッドを、前記ターゲットアテンションヘッドとして選択する、

請求項21に記載の電子装置。

【請求項23】

前記特定のアテンション重み行列は、予め決定された形態を有するガイド重み行列に基づいて学習される、

請求項22に記載の電子装置。

【請求項24】

前記特定のアテンション重み行列は、ステップごとにアテンション重みの分布が異なるように学習される、

請求項22に記載の電子装置。

【請求項25】

前記プロセッサは、

予め決定された目的に最も適するアテンション重み行列を生成するアテンションヘッドを、前記ターゲットアテンションヘッドとして選択する、

請求項21に記載の電子装置。

【請求項26】

前記プロセッサは、

予め決定された目的による予め決定された形態を有するガイド重み行列に基づいて、前記ターゲットアテンションヘッドを選択する、

請求項21に記載の電子装置。

【請求項27】

前記プロセッサは、

前記アテンション基盤シーケンスツースーケンスモデルが単調性質を有する場合に、前記複数のアテンションヘッドが作り出すアテンション重み行列に対して単調回帰分析を行って、前記ターゲットアテンションヘッドを選択する、

請求項21に記載の電子装置。

【請求項28】

前記プロセッサは、

前記複数のアテンションヘッドが作り出すアテンション重み行列のエントロピーに基づいて、前記ターゲットアテンションヘッドを選択する、

請求項21に記載の電子装置。

【請求項29】

前記プロセッサは、

前記アテンション重み行列のうち、前記エントロピーが最大のアテンション重み行列を作り出すアテンションヘッドを、前記ターゲットアテンションヘッドとして選択する、

請求項28に記載の電子装置。

【請求項30】

10

20

30

40

50

前記プロセッサは、

アテンション重み行列の各行の分布間の距離が最も遠いアテンション重み行列を作り出すアテンションヘッドを、前記ターゲットアテンションヘッドとして選択する、

請求項 2 1 に記載の電子装置。

【請求項 3 1】

前記プロセッサは、

前記ターゲットアテンションヘッドの前記入力シーケンスと前記出力シーケンスとの間のアテンション重みのうち、前記ターゲットアテンションヘッドのアテンション重みとガイド重み行列との差が第 1 閾値以上である、1 つ以上のエラーアテンション重みを検出し、

前記 1 つ以上のエラーアテンション重みに対応する出力トークンを、前記 1 つ以上のエラー出力トークンとして決定する、

請求項 2 1 に記載の電子装置。

【請求項 3 2】

前記プロセッサは、

前記ターゲットアテンションヘッドの前記入力シーケンスと前記出力シーケンスとの間のアテンション重みのうち、以前ステップのアテンション重みと第 2 閾値以上の類似度を有する、1 つ以上のエラーアテンション重みを検出し、

前記エラーアテンション重みに対応する出力トークンを、前記 1 つ以上のエラー出力トークンとして決定する、

請求項 2 1 に記載の電子装置。

【請求項 3 3】

前記プロセッサは、

前記 1 つ以上のエラー出力トークンを前記出力シーケンスから取り除く、

請求項 2 1 に記載の電子装置。

【請求項 3 4】

前記プロセッサは、

前記 1 つ以上のエラー出力トークンを取り除いた他の出力トークン候補のうち次の入力トークンを決定する、

請求項 2 1 に記載の電子装置。

【請求項 3 5】

前記プロセッサは、

前記 1 つ以上のエラー出力トークンを出力したステップの入力トークンを、前記次の入力トークンとして決定する、

請求項 3 4 に記載の電子装置。

【請求項 3 6】

電子装置であって、

入力シーケンスを受信し、前記入力シーケンスに基づいて、出力シーケンスを出力する、エンコーダー及びデコーダーと、

少なくとも 1 つのプロセッサと、

を含み、

前記少なくとも 1 つのプロセッサは、

それぞれのアテンション重み行列を作り出すように、それぞれが構成されており、かつ、前記デコーダーのアテンション基盤シーケンスツーシーケンスモデルに含まれた、複数のアテンションヘッドのうちターゲットアテンションヘッドを選択し、

前記ターゲットアテンションヘッドに基づいて、前記出力シーケンスに含まれているエラー出力トークンを検出し、

前記エラー出力トークンに基づいて、前記出力シーケンスを補正し、

前記補正された出力シーケンスを出力する、

電子装置。

【請求項 3 7】

前記エンコーダー及び前記デコーダーは、人工神経網に含まれている、
請求項 36 に記載の電子装置。

【請求項 38】

前記少なくとも 1 つのプロセッサは、
前記アテンション基盤シーケンスツリーモデルに格納されている複数のアテンション重み行列のうち、前記ターゲットアテンションヘッドを選択する、
請求項 36 に記載の電子装置。

【請求項 39】

前記アテンション基盤シーケンスツリーモデルは、
正しい結果が出る例題入力を前記エンコーダー及び前記デコーダーに入力することで、
複数のアテンションヘッドに対応する複数のアテンション重み行列を格納し、
前記少なくとも 1 つのプロセッサは、
前記シーケンスツリーモデルの予め決定された目的に最も適するアテンション重み行列を生成するアテンションヘッドを、前記ターゲットアテンションヘッドとして選択する、
請求項 38 に記載の電子装置。

【請求項 40】

前記少なくとも 1 つのプロセッサは、
前記複数のアテンションヘッドのうち特定アテンションヘッドを希望する形態のアテンション重み行列を生成するよう学習し、
前記特定アテンションヘッドを、前記特定アテンションヘッドとして決定する、
請求項 36 に記載の電子装置。

【発明の詳細な説明】

【技術分野】

【0001】

下記の実施形態は、アテンション基盤シーケンスツリーモデルの性能向上方法及び装置に関する。

【背景技術】

【0002】

人工神経網を用いて解決する問題のうち長さが決まっていないシーケンスを入力として受け、決まっていない長さの出力を生成しなければならない問題を解決するための方法として、シーケンスツリーモデル(sequence-to-sequence)モデルの 1 つであるエンコーダー-デコーダーの人工神経網モデルが導入された。

【0003】

可変の長さを入力として可変の長さを生成するシーケンスツリーモデルは、翻訳、音声認識、カンバセーション、文書要約などの様々なタスクで使用される人工神経網モデルであって、可変入力を上位表現であるコンテキスト・ベクトルに作り出すエンコーダーと、そのコンテキスト・ベクトルに基づいて可変出力を生成するデコーダーからなる。しかし、シーケンスツリーモデルは、コンテキスト・ベクトルが長くなれば性能が低下するという短所がある。

【発明の概要】

【発明が解決しようとする課題】

【0004】

本発明の目的は、アテンション基盤シーケンスツリーモデルの性能向上方法及び装置を提供する。

【課題を解決するための手段】

【0005】

一実施形態に係るアテンション基盤シーケンスツリーモデルの性能向上方法は、アテンション基盤シーケンスツリーモデルに基づいて入力シーケンスに対応する出力シーケンスを決定するステップと、複数のアテンションヘッドのうち、1 つ以上の

10

20

30

40

50

ターゲットアテンションヘッドを選択するステップと、前記出力シーケンスを構成する出力トークンのうち、前記ターゲットアテンションヘッドに基づいて、1つ以上のエラー出力トークンを検出するステップと、前記1つ以上のエラー出力トークンに基づいて前記出力シーケンスを補正するステップと、を含む。

【0006】

前記選択するステップは、前記ターゲットアテンションヘッドに対応するターゲットアテンション重み行列になるように学習された特定のアテンション重み行列を作り出すアテンションヘッドを前記ターゲットアテンションヘッドとして選択するステップを含み得る。

【0007】

前記特定のアテンション重み行列は、予め決定された形態を有するガイド重み行列に基づいて、学習され得る。

10

【0008】

前記ガイド重み行列は、出力シーケンス長さ、入力フレーム長さ、開始シフト、終了シフト、及びディフュージョン比率のうち少なくとも1つに基づいて決定され得る。

【0009】

前記特定のアテンション重み行列は、ステップごとにアテンション重みの分布が異なるように学習され得る。

【0010】

前記特定のアテンション重み行列は、以前ステップのアテンション重み累積和に基づいて、現在ステップのアテンション重みが決定されるように学習され得る。

20

【0011】

前記選択するステップは、予め決定された目的に最も適するアテンション重み行列を生成するアテンションヘッドを、前記ターゲットアテンションヘッドとして選択するステップを含み得る。

【0012】

前記選択するステップは、予め決定された目的による予め決定された形態を有するガイド重み行列に基づいて、前記ターゲットアテンションヘッドを選択するステップを含み得る。

【0013】

前記選択するステップは、前記アテンション基盤シーケンスツリーモデルが単調性質を有する場合、前記複数のアテンションヘッドが作り出すアテンション重み行列に対して単調回帰分析 (monotonic regression analysis) を行って前記ターゲットアテンションヘッドを選択するステップを含み得る。

30

【0014】

前記選択するステップは、前記複数のアテンションヘッドが作り出すアテンション重み行列のエントロピーに基づいて前記ターゲットアテンションヘッドを選択するステップを含み得る。

【0015】

前記エントロピーに基づいて前記ターゲットアテンションヘッドを選択するステップは、前記アテンション重み行列のうち、前記エントロピーが最大のアテンション重み行列を作り出すアテンションヘッドを前記ターゲットアテンションヘッドとして選択するステップを含み得る。

40

【0016】

前記エントロピーに基づいて前記ターゲットアテンションヘッドを選択するステップは、カルバック・ライブラー情報量 (Kullback-Leibler divergence) に基づいて前記ターゲットアテンションヘッドを選択するステップを含み得る。

【0017】

前記選択するステップは、アテンション重み行列の各行の分布間の距離が最も遠いアテンション重み行列を作り出すアテンションヘッドを前記ターゲットアテンションヘッドとして選択するステップを含み得る。

50

【0018】

前記検出するステップは、前記ターゲットアテンションヘッドの前記入力シーケンスと前記出力シーケンスとの間のアテンション重みのうち、前記ターゲットアテンションヘッドのアテンション重みとガイド重み行列との差が閾値以上の1つ以上のエラーアテンション重みを検出するステップと、前記1つ以上のエラーアテンション重みに対応する出力トークンを前記1つ以上のエラー出力トークンとして決定するステップとを含み得る。

【0019】

前記検出するステップは、前記ターゲットアテンションヘッドの前記入力シーケンスと前記出力シーケンスとの間のアテンション重みのうち、以前ステップのアテンション重みと閾値以上の類似度を有する1つ以上のエラーアテンション重みを検出するステップと、前記1つ以上のエラーアテンション重みに対応する出力トークンを前記1つ以上のエラー出力トークンとして決定するステップとを含み得る。

10

【0020】

前記補正するステップは、前記1つ以上のエラー出力トークンを前記出力シーケンスから取り除くステップを含み得る。

【0021】

前記補正するステップは、前記1つ以上のエラー出力トークンを取り除いた他の出力トークン候補のうち、次の入力トークンを決定するステップを含み得る。

【0022】

一実施形態に係るアテンション基盤シーケンスツーシーケンスモデルの性能向上方法は、前記1つ以上のエラー出力トークンが出力されたステップの入力トークンを前記次の入力トークンとして決定するステップをさらに含み得る。

20

【0023】

前記複数のアテンションヘッドの数は、前記アテンション基盤シーケンスツーシーケンスモデルのアテンションレイヤー数とデコーダーレイヤー数の積に該当し得る。

【0024】

一実施形態に係るアテンション基盤シーケンスツーシーケンスモデルの電子装置は、アテンション基盤シーケンスツーシーケンスモデルに基づいて、入力シーケンスに対応する出力シーケンスを決定し、複数のアテンションヘッドのうち1つ以上のターゲットアテンションヘッドを選択し、前記出力シーケンスを構成する出力トークンのうち、前記ターゲットアテンションヘッドに基づいて、1つ以上のエラー出力トークンを検出し、前記1つ以上のエラー出力トークンに基づいて、前記出力シーケンスを補正する、プロセッサを含む。

30

【0025】

前記プロセッサは、前記ターゲットアテンションヘッドに対応するターゲットアテンション重み行列になるよう学習された特定のアテンション重み行列を作り出すアテンションヘッドを、前記ターゲットアテンションヘッドとして選択し得る。

【0026】

前記特定のアテンション重み行列は、予め決定された形態を有するガイド重み行列に基づいて、学習され得る。

40

【0027】

前記特定のアテンション重み行列は、ステップごとにアテンション重みの分布が異なるように学習され得る。

【0028】

前記プロセッサは、予め決定された目的に最も適するアテンション重み行列を生成するアテンションヘッドを、前記ターゲットアテンションヘッドとして選択し得る。

【0029】

前記プロセッサは、予め決定された目的による予め決定された形態を有するガイド重み行列に基づいて、前記ターゲットアテンションヘッドを選択し得る。

【0030】

50

前記プロセッサは、前記アテンション基盤シーケンスツースーケンスモデルが単調性質を有する場合に、前記複数のアテンションヘッドが作り出すアテンション重み行列に対して単調回帰分析を行って、前記ターゲットアテンションヘッドを選択し得る。

【0031】

前記プロセッサは、前記複数のアテンションヘッドが作り出すアテンション重み行列のエントロピーに基づいて、前記ターゲットアテンションヘッドを選択し得る。

【0032】

前記プロセッサは、前記アテンション重み行列のうち、前記エントロピーが最大のアテンション重み行列を作り出すアテンションヘッドを、前記ターゲットアテンションヘッドとして選択し得る。

【0033】

前記プロセッサは、アテンション重み行列の各行の分布間の距離が最も遠いアテンション重み行列を作り出すアテンションヘッドを、前記ターゲットアテンションヘッドとして選択し得る。

【0034】

前記プロセッサは、前記ターゲットアテンションヘッドの前記入力シーケンスと前記出力シーケンスとの間のアテンション重みのうち、前記ターゲットアテンションヘッドのアテンション重みとガイド重み行列との差が第1閾値以上である、1つ以上のエラーアテンション重みを検出し、前記エラー1つ以上のアテンション重みに対応する出力トークンを、前記1つ以上のエラー出力トークンとして決定し得る。

【0035】

前記プロセッサは、前記ターゲットアテンションヘッドの前記入力シーケンスと前記出力シーケンスとの間のアテンション重みのうち、以前ステップのアテンション重みと第2閾値以上の類似度を有する、1つ以上のエラーアテンション重みを検出し、前記エラーアテンション重みに対応する出力トークンを、前記1つ以上のエラー出力トークンとして決定し得る。

【0036】

前記プロセッサは、前記1つ以上のエラー出力トークンを前記出力シーケンスから取り除くことができる。

【0037】

前記プロセッサは、前記1つ以上のエラー出力トークンを取り除いた他の出力トークン候補のうち次の入力トークンを決定し得る。

【0038】

前記プロセッサは、前記1つ以上のエラー出力トークンを出力したステップの入力トークンを前記次の入力トークンとして決定し得る。

【0039】

一実施形態に係る電子装置は、入力シーケンスを受信し、前記入力シーケンスに基づいて、出力シーケンスを出力するエンコーダー及びデコーダーと、少なくとも1つのプロセッサとを含み、前記少なくとも1つのプロセッサは、前記デコーダーのアテンション基盤シーケンスツースーケンスモデルに含まれた複数のアテンションヘッドのうちターゲットアテンションヘッドを選択し、前記ターゲットアテンションヘッドに基づいて前記出力シーケンスに含まれているエラー出力トークンを検出し、前記エラー出力トークンに基づいて前記出力シーケンスを補正し、前記補正された出力シーケンスを出力する。

【0040】

前記エンコーダー及び前記デコーダーは、人工神経網に含まれてもよい。

【0041】

前記少なくとも1つのプロセッサは、前記アテンション基盤シーケンスツースーケンスモデルに格納されている複数のアテンション重み行列のうち前記ターゲットアテンションヘッドを選択し得る。

【0042】

10

20

30

40

50

前記アテンション基盤シーケンスツースーケンスモデルは、正しい結果が出る例題入力を前記エンコーダー及び前記デコーダーに入力することで、複数のアテンションヘッドに対応する複数のアテンション重み行列を格納し、前記少なくとも1つのプロセッサは、前記シーケンスツースーケンスモデルの予め決定された目的に最も適するアテンション重み行列を生成するアテンションヘッドを、前記ターゲットアテンションヘッドとして選択し得る。

【0043】

前記少なくとも1つのプロセッサは、前記複数のアテンションヘッドのうち特定アテンションヘッドを希望する形態のアテンション重み行列を生成するよう学習し、前記特定アテンションヘッドを前記特定アテンションヘッドとして決定し得る。

【発明の効果】

【0044】

本発明によると、アテンション基盤シーケンスツースーケンスモデルの性能向上方法及び装置を提供することができる。

【図面の簡単な説明】

【0045】

【図1】一実施形態に係るシーケンスツースーケンスモデルが使用されるシステムの構成図である。

【図2A】一実施形態に係るエンコーダー-デコーダー構造のエンドツーエンド人工神経網を有する電子装置の動作を説明するための図である。

【図2B】一実施形態に係るアテンション重みを決定する方法を説明するための図である。

【図2C】一実施形態に係るアテンション重みを決定する方法を説明するための図である。

【図3A】一実施形態に係るアテンション重み行列を説明するための図である。

【図3B】一実施形態に係るアテンション重み行列を説明するための図である。

【図4A】一実施形態に係る電子装置のブロック図である。

【図4B】他の実施形態に係る電子装置のブロック図である。

【図5】一実施形態に係るアテンション基盤シーケンスツースーケンスモデルの性能向上方法を説明するためのフローチャートである。

【図6】一実施形態に係る予め決定された形態を有するガイド重み行列を説明するための図である。

【図7】一実施形態に係る毎出力トークンを生成するとき以前とは異なるアテンション重みを有するように構成されたガイド重み行列を説明するための図である。

【発明を実施するための形態】

【0046】

実施形態に対する特定の構造的又は機能的な説明は単なる例示を目的として開示されたものとして、様々な形態に変更される。したがって、実施形態は特定の開示形態に限定されるものではなく、本明細書の範囲は技術的な思想に含まれる変更、均等物ないし代替物を含む。

【0047】

第1又は第2などの用語を複数の構成要素を説明するために用いることがあるが、このような用語は1つの構成要素を他の構成要素から区別する目的としてのみ解釈されなければならない。例えば、第1構成要素は第2構成要素と命名することができ、同様に第2構成要素は第1構成要素としても命名することができる。

【0048】

また、実施形態の構成要素を説明することにおいて、第1、第2、A、B、(a)、(b)などの用語を使用することができる。これらの用語は、その構成要素を他の構成要素と区別するためのものに過ぎず、その用語によって当該構成要素の本質や順序などが限定されることはない。いずれかの構成要素が他の構成要素に「連結」、「結合」、又は「接続」されていると言及されたときには、その他の構成要素に直接的に連結されているか又は接続されているが、各構成要素との間にさらなる構成要素が「連結」、「結合」、又は

10

20

30

40

50

「接続」され得るものと理解されなければならない。

【0049】

単数の表現は、文脈上、明白に異なる意味をもたない限り複数の表現を含む。本明細書において、「含む」又は「有する」等の用語は明細書上に記載した特徴、数字、ステップ、動作、構成要素、部品、又はこれらを組み合わせたものが存在することを示すものであって、1つ又はそれ以上の他の特徴や数字、ステップ、動作、構成要素、部品、又はこれを組み合わせたものなどの存在、又は、付加の可能性を予め排除しないものとして理解しなければならない。

【0050】

異なるように定義されない限り、技術的であるか又は科学的な用語を含む、ここで用いる全ての用語は、本実施形態が属する技術分野で通常の知識を有する者によって一般的に理解されるものと同じ意味を有する。一般的に用いられる、予め定義された用語は、関連技術の文脈上で有する意味と一致する意味を有するものと解釈すべきであって、本明細書で明白に定義しない限り、理想的又は過度に形式的な意味として解釈されることはない。

【0051】

以下、実施形態を添付の図面を参照しながら詳説する。各図面に提示された同じ参照符号は同じ部材を示す。

【0052】

図1は、一実施形態に係るシーケンスツリーモデルが使用されるシステムの構成図である。

【0053】

図1を参照すると、シーケンスツリーモデル (Sequence-to-sequence) モデルが使用されるシステムは、ユーザ端末110及び電子装置120を含む。図1に示すシーケンスツリーモデルが使用されるシステムの各構成要素は、機能的に区分される機能要素を示すものであって、少なくとも1つの構成要素が実際の物理的環境では互いに統合される形態において実現されてもよい。例えば、図1には、電子装置120が一種のサーバー装置として示されているが、実施形態によって電子装置120の各機能は、ユーザ端末110に内蔵されている形態に実現されてもよい。

【0054】

シーケンスツリーモデルは、入力シーケンスから他のドメインの出力シーケンスを出力する様々な分野で使用され得る。例えば、シーケンスツリーモデルは、音声認識システム、機械翻訳システム、チャットボットシステムなどで使用され得る。音声認識システムは、音声を受信して音声認識文章を出力することができ、機械翻訳システムは、第1言語文章を受信し、これに対応する第2言語文章を出力することができ、チャットボットシステムは、質問を受信して、質問に対応する返答を出力することができる。以下では、説明の便宜のために音声認識システムにおけるシーケンスツリーモデルを基準にして説明するが、実施形態は、シーケンスツリーモデルが使用され得る様々な分野に適用されてもよい。

【0055】

音声認識システムにおいて、ユーザ端末110は、入力シーケンス130を受信して電子装置120に伝達し、電子装置120から受信した出力シーケンス140をユーザに提供する端末である。図1において、ユーザ端末110は、スマートフォンとして示されているが、ユーザ端末110は、コンピュータ、UMPC (Ultra Mobile PC)、ワークステーション、ネットブック (net-book)、PDA (Personal Digital Assistants)、ポータブル (portable) コンピュータ、ウェブタブレット (web tablet)、無線電話機 (wireless phone)、モバイルフォン (mobile phone)、e-ブック (e-book)、PMP (portable multimedia player)、ポータブルゲーム機、ナビゲーション (navigation) 装置、ブラックボックス (black box) 又はデジタルカメラ (digital camera) のような電子装置の1つであ

10

20

30

40

50

り得る。

【0056】

電子装置120は、話者の入力シーケンス130の入力を受けて認識結果である出力シーケンス140を提供するコンピュータ装置であって、シーケンスツーシーケンスモデル基盤の電子装置である。入力シーケンス130は、入力データと入力データから抽出された入力特徴ベクトルを含む。例えば、音声認識システムにおいて、入力シーケンス130は、入力音声と入力音声から抽出された入力音声特徴ベクトルを含む。また、コンピュータ装置は、ノート型パソコン、デスクトップ（desktop）、ラップトップ（laptop）、スマートフォン（smart phone）などのモバイル装置であってもよいが、これに限定されることなく、演算装置が具備された全ての種類の装置を含んでもよい。

10

【0057】

電子装置120は、人工神経網で構成されたシーケンスツーシーケンスモデルを構築し、構築されたシーケンスツーシーケンスモデルを用いて入力シーケンス130に対する認識結果である出力シーケンス140を提供する。人工神経網は、例えば、リカレントニューラルネットワーク（RNN； Recurrent Neural Network）、畳み込みニューラルネットワーク（CNN； Convolutional Neural Network）、セルフアテンション神経網（SANN； Self-Attention Neural Network）、BRNN（Bi-directional RNN）、LSTM（Long Short Term Memory）、BLSTM（Bi-directional LSTM）、GRU（Gated Recurrent Unit）、BGRU（Bi-directional GRU）などであり得るが、これに限定されることはない。

20

【0058】

また、電子装置120は、出力シーケンス140の正確度を向上するための補正装置を含む。補正装置は、電子装置120と機能的に区分される機能要素を示すものであり、実際の物理的環境では、補正装置の少なくとも1つの構成要素が電子装置120と統合される形態において実現されてもよい。電子装置120は、補正装置を用いて出力シーケンス140のエラーを検出し、補正された出力シーケンスを出力することができる。

【0059】

以下では、図2A～図2Cを参照して、多重ヘッドアテンション基盤シーケンスツーシーケンスモデルの動作方法を説明し、図3A～図3Bを参照して、アテンション重み行列を説明し、図4A～図7を参照して、アテンション基盤シーケンスツーシーケンスモデルの性能向上方法を説明する。

30

【0060】

図2Aは、一実施形態に係るエンコーダー-デコーダー構造のエンドツーエンド人工神経網を有する電子装置の動作を説明するための図である。

【0061】

図2Aを参照すると、一実施形態に係る人工神経網は、エンコーダー220及びデコーダー230を含む。

40

【0062】

人工神経網は、エンコーダー220-デコーダー230構造のエンドツーエンド人工神経網であり、モデルの全ての媒介変数が1つの損失関数に対して同時に訓練される経路が可能であるため、エンコーダー220とデコーダー230が同時に学習され得る。人工神経網は、一方のエンドから入力され、他方のエンドに出力を生成するが、入力及び出力を直接考慮してネットワーク重みを最適化することができる。人工神経網内のレイヤーのノードは非線型的で互いに影響を与える関係であり、各ノードから出力される値、ノード間の関係など、人工神経網のパラメータは学習によって最適化され得る。

【0063】

エンコーダー220-デコーダー230構造のエンドツーエンド人工神経網は、エンコ

50

ーダー２２０とデコーダー２３０が統合されたネットワーク構造であって、エンコーダー２２０とデコーダー２３０は、入力シーケンスから入力シーケンスに対応する出力シーケンスを直接生成する。

【００６４】

例えば、電子装置は、入力データ２１０から入力特徴ベクトル２１５を抽出し、エンコーダー２２０は、入力特徴ベクトル２１５を符号化し、符号化された特徴２２５を生成する。エンコーダー２２０は、入力特徴ベクトル２１５の次元を変換させ、符号化された特徴２２５を生成し得る。即ち、エンコーダー２２０は、入力シーケンスの意味を円滑に要約する役割を果たす。

【００６５】

符号化された特徴２２５は、人工神経網のデコーダー２３０に印加されてもよい。デコーダー２３０は、トークン単位で符号化された特徴２２５と以前に決定された認識結果に基づいて出力トークンを決定することができる。

【００６６】

具体的に、符号化された特徴２２５を受信したデコーダー２３０は、トークン単位で認識結果を出力し、トークンは、単語、部分語（*subword*）、単一字（*character*）、あるいは単一字をなす単位（ハングルにおいては初声、中性、終声など）になる。以下では、デコーダー２３０が出力する全体認識結果を出力シーケンスと称する。出力シーケンスは、複数の出力トークンを含む。

【００６７】

デコーダー２３０は、決定されていない長さの出力を人工神経網で算出するために、終わりを意味する特殊トークンである<EOS>が出力される以前まで、今まで推正したトークンを入力として次のトークンを１つずつ、次に、予測することができる。デコーダー２３０は、ステップごとにエンコーダー２２０から算出された情報に基づいて出力トークンを求めるが、ここでは、以前ステップの出力トークンに属して求めることができ、このような方法を自己回帰（*autoregressive*）デコーディングと言う。

【００６８】

即ち、自己回帰デコーディングを行うエンドツーエンド人工神経網モデルから出力トークンに選択されたトークンが、次の入力トークンとして決定され得る。例えば、「*tell me a joke*」という文章において「*me*」を認識するとき、以前ステップの出力トークンである「*tell*」を考慮する。

【００６９】

但し、エンコーダー２２０－デコーダー２３０構造のエンドツーエンド人工神経網は、全体入力データを単一のベクトルに処理するため、入力データが長くなればデータの前の部分の情報が希薄にされ、出力シーケンスの品質が落ちる問題がある。これを解決するために、デコーダー２３０で出力トークンを推正するステップごとに、エンコーダー２２０における全体入力データを再度参考することができる。但し、全体入力データを全て同じ割合で参考にすることなく、該当ステップで予測しなければならない出力トークンと関連のある入力音声部分にもっと集中することができる。

【００７０】

より具体的に、入力シーケンスを構成している各入力フレームは、該当ステップで予測しなければならない出力トークンと関連する程度に応じて、アテンション重み（*attention weight*）が設定され、デコーダー２３０は、設定されたアテンション重みに基づいて、ステップごとに出力トークンを推正することができる。

【００７１】

エンコーダー２２０とデコーダー２３０は、入力された入力特徴ベクトルから認識結果のシーケンスを生成するように予め学習されてもよい。例えば、エンコーダー２２０とデコーダー２３０のモデルパラメータ（例えば、シナプス接続の重みとニューロンの偏向）は、入力データに対応する正解データ対のシーケンスから認識結果のシーケンスを生成するように予め学習され得る。さらに、エンコーダー２２０とデコーダー２３０のアテンシ

10

20

30

40

50

ヨン重みについても、入力データに対応する正解テキスト対のシーケンスから認識結果のシーケンスを生成するように予め学習されてもよい。

【0072】

図2B～図2Cは、一実施形態に係るアテンション重みを決定する方法を説明するための図である。

【0073】

図2Bを参照すると、エンコーダー240は、入力フレーム（例えば、 $i_1 \sim i_N$ 、 N は入力フレームの数）を含む入力シーケンスを受信し、エンコーダーの隠れ状態ベクトル（例えば、 $h_1 \sim h_N$ ）を生成することができる。

【0074】

アテンションモジュール251は、デコーダー250がステップごとに出力トークンを推正するときに使用される、該当ステップで入力シーケンスを構成する各入力フレーム（例えば、 $i_1 \sim i_N$ ）のアテンション重みを決定し、決定されたアテンション重みに基づいて、コンテキスト・ベクトル（*context vector*）（例えば、 $c_1 \sim c_M$ 、 M は出力トークンの数）を生成する。

【0075】

多重ヘッドアテンション基盤シーケンスツーシーケンスモデルである場合、アテンションモジュール251は多重ヘッドアテンションモジュールであってもよく、アテンションレイヤーは複数のアテンションレイヤー（例えば、第1アテンションレイヤーないし第 k アテンションレイヤー）を含み、各アテンションレイヤーは、並列的にアテンション重みを決定し、決定されたアテンション重みに基づいてコンテキスト・ベクトルを生成することができる。アテンションモジュール251は、暗黙的に様々な役割をするように学習された複数のアテンションレイヤーを用いてもよい。

【0076】

デコーダー250は、以前ステップのデコーダーの隠れ状態ベクトル s_{t-1} 、以前ステップの出力トークン y_{t-1} 、及び現在ステップのコンテキスト・ベクトル c_t が入力され、現在ステップのデコーダーの隠れ状態ベクトル s_t 及び現在ステップの出力トークン y_t を抽出する。デコーダーが現在ステップで出力トークンを抽出するときに使用される現在ステップのコンテキスト・ベクトル c_t は次の数式（1）のように求めることができる。

【数1】

$$c_t = \sum_{i=1}^N \alpha_i^t h_i \quad (1)$$

数式（1）の α_i^t は、ステップ t に対して i 番目の入力フレームに対するアテンション重みを意味する。

【0077】

図2Cを参照すると、 α_i^t は、ステップ t に対して i 番目の入力フレームに対するアテンション重みで全ての入力フレーム（例えば、1番目から N 番目のまで）に対するアテンション重みを和すると1である。もし、 α_1^t ないし α_N^t のうち α_2^t の値が最大である場合、デコーダー250は、現在ステップ t で2番目の入力フレームに最も集中して出力トークンを抽出することができる。

【0078】

多重ヘッドアテンションモデルは、既存の単一アテンションモデルよりも性能が改善されたが、可変の長さを出力するシーケンスツーシーケンスモデルの慢性的な短所である一部の出力をジャンピング（*jumping*）したり、以前出力の内容を繰り返す（*rep*

e a t i n g) という問題が依然として残っている。

【0079】

例えば、多重ヘッド基盤シーケンスツースーケンスモデルは「t e l l m e a j o k e」という文章に対応する音声が入力される場合、場合に応じて、「t e l l」の後に「m e a」をジャンピングして「j o k e」を出力する現象が発生し得る。

【0080】

また、多重ヘッド基盤シーケンスツースーケンスモデルは、下記のように珍しい音声あるいは前に見られなかった単語（e s s e n d i r e r u m）に対応する音声が入力されると、下表1のように節、あるいは一つの単語を繰り返す現象が発生し得る。

【表1】

10

表1

正解	corollary hence it follows that god is not only the cause of things coming into existence but also of their continuing in existence that is in scholastic phraseology god is cause of the being of things <i>essendi rerum</i>
認識結果	corollary hence it follows that god is not only the cause of things coming into existence but also of their continuing in existence that is in scholastic phraseology god is cause of the being of things as sindi rerum god is cause of the being of things as sindi rerum god is cause of the being of things as sindi rerum god is cause of the being of things acindi rerum god is cause of the being of the being of things as sindi

20

30

これに、一実施形態に係る補正装置は、複数のアテンションヘッドのうちターゲットアテンションヘッドを選択し、ターゲットアテンションヘッドに基づいて1つ以上のエラー出力トークンを検出し、エラー出力トークンに基づいて出力シーケンスを補正して電子装置の最終出力シーケンスの正確度を向上することができる。

【0081】

図3A～図3Bは、一実施形態に係るアテンション重み行列を説明するための図である。

40

【0082】

一実施形態に係るアテンション重み行列は、出力シーケンスを構成するそれぞれの出力トークンに対して、入力シーケンスを構成するそれぞれの入力フレームのアテンション重みをエレメントとして有してもよい。例えば、電子装置が「t e l l m e a j o k e」のような出力シーケンスを出力するケースであり、アテンション重み行列は下記の表2の通りである。

50

【表 2】

表 2

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
tell	0.9	0.07	0.02	0.004	0.001	0.001	0.001	0.001	0.001
me	0.004	0.02	0.9	0.07	0.001	0.001	0.001	0.001	0.001
a	0.001	0.001	0.001	0.004	0.33	0.3	0.36	0.001	0.001
joke	0.001	0.001	0.001	0.001	0.001	0.62	0.37	0.004	0.001

表 2 の例示において、出力トークンは単語単位であり、各入力フレームは、10ms の時間単位を有してもよい。表 2 を参照すると、出力トークン「tell」の場合、入力フレーム i_1 のアテンション重みが 0.9 として最も大きく、 i_2 のアテンション重みは 0.07 として 2 番目に大きく、 i_3 のアテンション重みは 0.02 として 3 番目に大きく、 $i_1 \sim i_3$ のアテンション重みを全て和すると、0.99 であり、出力トークン「tell」は入力フレーム $i_1 \sim i_3$ に集中して抽出されることが分かる。

【0083】

多重ヘッド基盤シーケンスツーシーケンスモデルでは、アテンションヘッド当たり 1 つのアテンション重み行列が作られ、各アテンション重み行列は暗黙的に各自の役割を果たすように学習され得る。例えば、 n 個のアテンションヘッドを有する多重ヘッド基盤シーケンスツーシーケンスモデルでは、 n 個のアテンション重み行列が作られてもよい。

【0084】

図 3A を参照すると、アテンション重み行列のアテンション重みは色を用いて表現されてもよい。例えば、2 つのアテンション重み行列 310、320 において、アテンション重みが大きいほど白色に近く表現され、反対に、アテンション重みが小さいほど黒色に近づいて表現される。2 つのアテンション重み行列 310、320 の横軸は、左側から右側に時間の流れによる入力フレームの順であり、縦軸は、上側から下側に時間の流れによる出力トークン（例えば、単語）の順である。

【0085】

複数のアテンション重み行列を用いて出力シーケンスを推正する場合、それぞれのアテンション重み行列の役割が異なってもよい。暗黙的に学習されたが、2 つのアテンション重み行列 310、320 は互いに異なる役割を行っているものと解釈できる。例えば、アテンション重み行列 310 は、音声のうち、実際発話のない黙音区間でアテンション重みが大きく示され、これは出力を構成するとき背景騒音を無視できるように助け、アテンション重み行列 320 は、各単語が実際発話された音声の位置でアテンション重みが高いことから、実際に認識する単語の音声部分に集中可能にする。

【0086】

前述したように、多重ヘッドアテンションモデルは、既存の単一アテンションモデルより性能が改善しているものの、可変の長さを出力するシーケンスツーシーケンスモデルの慢性的な短所である一部の出力をジャンピングしたり、以前の出力内容を繰り返すという問題がある。

【0087】

例えば、図 3 B に示すアテンション重み行列 3 3 0 は、認識を正常にできず、繰り返されるエラーを出力する場合のアテンション重み行列である。アテンション重み行列 3 3 0 は、アテンション重み行列 3 2 0 のようなアテンションヘッドで作られたアテンション重み行列であってもよいが、アテンション重み行列 3 2 0 は、正常に出力トークンを生成するときのアテンション重み行列であってもよく、アテンション重み行列 3 3 0 は、以前の出力内容を繰り返すエラーを出力する場合のアテンション重み行列であってもよい。アテンション重み行列 3 3 0 を参照すると、出力トークン軸である x 軸を基準にして、約 1 / 3 地点までは結果が正常に出力されたが、その後の数個の出力トークン（例えば、単語）が繰り返されて出力されていることが分かる。

【 0 0 8 8 】

一実施形態に係る補正装置は、上記のようなエラーを補正するために、特定のアテンションヘッドを選択し、このアテンションヘッド（以下、ターゲットアテンションヘッド）のアテンション重みを分析しエラー出力を抑制することで、多重ヘッドアテンションシーケンスツースーシーケンスモデルの性能を高めることができる。具体的なアテンション基盤シーケンスツースーシーケンスモデルの性能向上方法については、次の図 4 A ~ 図 7 を参照して説明される。

【 0 0 8 9 】

図 4 A は、一実施形態に係る電子装置のブロック図である。

図 4 A を参照すると、一実施形態に係る電子装置 4 0 1 は、エンコーダー 4 2 0、デコーダー 4 3 0、ターゲットアテンションヘッド選択モジュール 4 6 0、及びエラー出力トークン検出及び補正モジュール 4 7 0 を含む。但し、図 4 A には、実施形態に係る構成要素のみが示されており、他の汎用的な構成要素がさらに含まれてもよい。また、図 4 A に示す電子装置 4 0 1 の各構成要素は、機能的及び論理的に分離できることを示すために別途に示し、物理的に必ずしも別途の構成要素、又は別途のコードに実現されることを意味するものではなく、実際の物理的環境では互いに統合される形態において実現され得ることに留意しなければならない。以下、電子装置 4 0 1 の各構成要素について説明する。

【 0 0 9 0 】

電子装置 4 0 1 は、特徴ベクトル抽出モジュール（図示せず）をさらに含む。特徴ベクトル抽出モジュールは、入力信号（例えば、音声信号）から特徴ベクトルを抽出する。入力信号は、複数のフレームごとの情報を含む信号であり、特徴ベクトルは、少なくとも 1 つのフレーム単位に抽出された情報のシーケンスであり、多次元のベクトルに表現されてもよい。以下、入力シーケンスは、入力信号と入力信号から抽出された特徴ベクトルを含んでもよく、入力フレーム単位で情報を含んでもよい。

【 0 0 9 1 】

図 4 A に示すエンコーダー 4 2 0 及びデコーダー 4 3 0 は、図 2 A を参照して説明したエンコーダー 2 2 0 及びデコーダー 2 3 0 に対応する。したがって、図 2 A と重複される説明は省略することにする。デコーダー 4 3 0 は、ステップごとに出力トークンを推正するたびに、どのような入力フレームにもっと集中することがよいかを決定するアテンションモジュール 4 4 0 を含む。具体的に、アテンションモジュール 4 4 0 は、デコーダー 4 3 0 がステップごとに出力トークンを推正するとき使用され、該当ステップで入力シーケンスを構成している各入力フレームのアテンション重みを決定するデコーダー 4 3 0 は、設定されたアテンション重みに基づいてステップごとに出力トークンを推正することができる。アテンションモジュール 4 4 0 は、多重ヘッドアテンションモジュールであってもよい。多重ヘッドアテンションモジュールは、様々な役割を果たすように学習された複数のアテンションレイヤーを用いて、全体ディメンションに対して 1 回のみのアテンションを適用することなく、全体ディメンションを複数に分けてアテンションを複数回適用させることができる。また、アテンションモジュール 4 4 0 は、全てのアテンションヘッドのアテンション重み行列を格納することができる。

【 0 0 9 2 】

電子装置 4 0 1 は、エンコーダー 4 2 0 及びデコーダー 4 3 0 を用いて入力シーケンス

10

20

30

40

50

４１０に対応する出力シーケンス４５０を出力する。但し、前述したように、予想とは異なる出力シーケンス４５０が出力されてもよい。

【００９３】

一実施形態に係る電子装置４０１は、ターゲットアテンションヘッド選択モジュール４６０を用いて複数のアテンションヘッドのうち、ターゲットアテンションヘッドを選択し、エラー出力トークン検出及び補正モジュール４７０を用いて選択されたターゲットアテンションヘッドに基づいて、出力シーケンスに含まれた出力トークンのうちエラー出力トークンを検出し、補正された出力シーケンス４８０を最終結果として出力することができる。

【００９４】

より具体的に、ターゲットアテンションヘッド選択モジュール４６０は、アテンションモジュール４４０に格納されている複数のアテンション重み行列のうち、ターゲットアテンションヘッドを選択してもよい。アテンションモジュール４４０は、エンコーダー４２０及びデコーダー４３０に正しい結果が出る例題を入力し、全てのアテンションヘッドに対するアテンション重み行列を格納することができ、ターゲットアテンションヘッド選択モジュール４６０は、格納されている全てのアテンションヘッドのうちシーケンスツーシーケンスモデルが行う目的に応じて、最も適切なアテンション重み行列を作り出すアテンションヘッドをターゲットアテンションヘッドとして選定することができる。ここで、アテンション重み行列を格納してターゲットアテンションヘッドを選定する動作について、エラー可否の判断を受けようとする入力シーケンス４１０が入力される以前に予め行われ得る。

【００９５】

前述したように、電子装置４０１は、補正装置を含み、補正装置は、ターゲットアテンションヘッド選択モジュール４６０及びエラー出力トークン検出及び補正モジュール４７０を含む。

【００９６】

図４Ｂは、他の実施形態に係る電子装置のブロック図である。

【００９７】

図４Ｂに示す電子装置４０２は、図４Ａに示す電子装置４０１のターゲットアテンションヘッド選択モジュール４６０の代りにアテンションガイドモジュール４９０を含んでもよい。図１～図４Ａに示す説明は図４Ｂにも適用可能であるため、重複される内容の説明は省略する。

【００９８】

図４Ａに示す電子装置４０１がアテンションモジュール４４０に格納されているアテンション重み行列のうち、ターゲットアテンションヘッドを選択することとは異なって、図４Ｂに示す電子装置４０２は、明示的に特定アテンションヘッドをターゲットアテンションヘッドとして活用するよう学習することができる。

【００９９】

より具体的に、アテンションガイドモジュール４９０は、任意に決定されたアテンションヘッドが希望する形態のアテンション重み行列を生成するように学習することができ、後で出力シーケンスのエラー可否を判断する時、エラー出力トークン検出及び補正モジュール４７０にターゲットアテンションレイヤーに学習したアテンションヘッドを知らせることができる。

【０１００】

前述したように、電子装置４０２は補正装置を含んでもよく、補正装置は、アテンションガイドモジュール４９０、及び、エラー出力トークン検出及び補正モジュール４７０を含んでもよい。

【０１０１】

図５は、一実施形態に係るアテンション基盤シーケンスツーシーケンスモデルの性能向上方法を説明するためのフローチャートである。

【0102】

図5を参照すると、ステップS510～S540は図1ないし図4Bを参照して前述した電子装置によって行うことができる。電子装置は、1つ以上のハードウェアモジュール、1つ以上のソフトウェアモジュール、又はこれらの様々な組み合わせによって実現され、前述したように、電子装置は、補正装置を含む。図1ないし図4Bの説明は図5にも適用され得るため、重複する内容の説明は省略する。

【0103】

ステップS510において、電子装置は、アテンション基盤シーケンスツリーモデルに基づいて入力シーケンスに対応する出力シーケンスを決定する。例えば、電子装置は、入力音声又は入力音声に対応する入力音声特徴ベクトルを受信してもよい。

10

【0104】

ステップS520において、電子装置は、複数のアテンションヘッドのうち、1つ以上のターゲットアテンションヘッドを選択する。多重ヘッドアテンション基盤シーケンスツリーモデルに基盤した電子装置は、アテンションレイヤー数とデコーダーレイヤー数の積分だけのアテンションヘッドを有してもよい。それぞれのアテンションヘッドが作り出すアテンション重み行列の役割が異なってもよい。電子装置は、互いに異なる役割を行うアテンション重み行列を作り出すアテンションヘッドのうち、該当する多重ヘッドアテンション基盤シーケンスツリーモデルが行う目的に最も適したアテンション重み行列を作り出す1つ以上のアテンションヘッドをターゲットアテンションヘッドとして選定してもよい。

20

【0105】

ターゲットアテンションヘッドを選択する方法は、大きく2つに区分される。第1に、電子装置に正しい結果が出る例題入力を入れ、全てのアテンションヘッドに対するアテンション重み行列を予め格納し、格納された全てのアテンションヘッドのうち、シーケンスツリーモデルが行う目的に応じて最も適切なアテンション重み行列を作り出すアテンションヘッドを求める方法である。第2に、任意に決定されたアテンションヘッドが希望する形態のアテンション重み行列を生成するように学習し、そのように学習されたアテンション重み行列をターゲットアテンションレイヤーとして選択する方法である。

【0106】

第1の方法によれば、電子装置は、複数のアテンションヘッドのうち、予め決定された目的に最も適するアテンション重み行列を生成するアテンションヘッドをターゲットアテンションヘッドとして選択する。

30

【0107】

一実施形態によると、電子装置は、予め決定された目的による予め決定された形態を有するガイド重み行列に基づいて、ターゲットアテンションヘッドを選択してもよい。例えば、音声認識は、入力シーケンス（例えば、入力音声）と出力シーケンス（例えば、認識された単語）を時間に応じてマッピングすれば単調性質を有することになる。この場合、電子装置は、複数のアテンションヘッドが作り出すアテンション重み行列に対して単調回帰分析（monotonic regression analysis）を行い、最も単調性質を満足するアテンションヘッドをターゲットアテンションヘッドとして選択する。この場合、ガイド重み行列は、単調行列であってもよい。

40

【0108】

図6は、一実施形態に係る予め決定された形態を有するガイド重み行列を説明するための図である。

【0109】

図6を参照すると、ガイド重み行列600は、アテンション基盤シーケンスツリーモデルが単調性質を有するように誘導するアテンション重みを有し得る。ガイド重み行列600のアテンション重みは、下記の数式（2）のように決定される。

【数2】

50

$$W[i, j] = \exp\left(-\frac{\left(\frac{j}{J} - \frac{(i + s_b)^2}{I + s_b + s_e}\right)}{2g^2}\right) \quad (2)$$

数式(2)において、 i は出力シーケンスの長さ、 j は入力シーケンスの長さ、 s_b は開始シフト(start shift)、 s_e は終了シフト(end shift)、 g はディフュージョン比率(diffusion ratio)を意味する。図面610を参照すると、ガイド重み行列600の色はアテンション重みを意味する。

【0110】

また、図5を参照すると、第1方法の他の例として、翻訳、要約など入力を多様に見なければならないタスクの場合、ガイド重み行列のガイドアテンション重みは予め決定された形態を有するものではない、出力トークンを生成するたびに以前とは異なるアテンション重みを有するよう構成される。このような場合、電子装置は、複数のアテンションヘッドが作り出すアテンション重み行列のエントロピー(entropy)に基づいてターゲットアテンションヘッドを選択することができる。電子装置は、アテンション重み行列のうち、エントロピーが最大のアテンション重み行列を作り出すアテンションヘッドを前記ターゲットアテンションヘッドとして選択してもよい。例えば、電子装置は、アテンション重み行列のうち、アテンション重み行列の各行の分布間の距離が最も遠いアテンション重み行列を作り出すアテンションヘッドをターゲットアテンションヘッドとして選択してもよい。具体的に、電子装置は、カルバック・ライブラー情報量(Kullback-Leibler divergence)を算出し、カルバック・ライブラー情報量が最大のアテンション重みを作り出すアテンションヘッドをターゲットアテンションヘッドとして選択してもよい。当該分野で通常の知識を有する者であれば、カルバック・ライブラー情報量を明確に理解できるため、より詳細な説明は省略する。

【0111】

図7は、一実施形態に係る出力トークンを生成するたびに以前とは異なるアテンション重みを有するよう構成されたガイド重み行列を説明するための図である。

【0112】

図7を参照すると、ガイド重み行列700は、出力トークンを生成するたびに以前とは異なるアテンション重みを有するよう構成されてもよい。図7において、 x 軸は出力シーケンス、 y 軸は入力シーケンスに対応する。 t 出力トークンを予測するとき、以前出力トークンを予測するとき、比重の低かった6番目の入力フレームにさらに大きいアテンション重みを有してもよい。

【0113】

また、図5を参照すると、2番目の方法によれば、電子装置は、任意に決定されたアテンションヘッドが希望する形態のアテンション重み行列を生成するように学習し、そのように学習されたアテンション重み行列をターゲットアテンションレイヤーとして選択してもよい。電子装置は、ターゲットアテンションヘッドに対応するターゲットアテンション重み行列になるよう学習された特定のアテンション重み行列を作り出すアテンションヘッドをターゲットアテンションヘッドとして選択してもよい。

【0114】

一例として、特定のアテンション重み行列は、予め決定された目的による予め決定された形態を有するガイド重み行列に基づいて学習されてもよい。予め決定された目的による予め決定された形態を有するガイド重み行列は、1番目の方法により前述したガイド行列である。例えば、電子装置は、図6に示すガイド行列600に基づいて特定のアテンション重み行列が単調性質を有するよう学習してもよい。

【0115】

他の例として、翻訳、要約など入力を多様に見なければならないタスクの場合、特定のアテンション重み行列は、ステップごとにアテンション重みの分布が異なるように学習されてもよい。特定のアテンション重み行列は、以前ステップのアテンション重みの累積和に基づいて、現在ステップのアテンション重みが決定されるように学習されてもよい。例えば、図7を参照すると、出力トークンを予測するステップごとに以前ステップのアテンション重み累積和を算出し、以前ステップのアテンション重み累積和が小さい箇所に高いアテンション重みを有するよう特定のアテンション重み行列を学習することができる。

【0116】

第2方法において、特定のアテンション重み行列を学習する方法として、ガイド重み行列をバイアスのように活用し、人工神経網学習過程で自然にガイド重み行列と類似形態のアテンション重みを生成するよう学習したり、明示的にカルバック・ライブラー情報量のような損失関数（loss function）を人工神経網学習損失として追加してもよい。

【0117】

ステップS530において、電子装置は、出力シーケンスを構成する出力トークンのうち、ターゲットアテンションヘッドに基づいて、1つ以上のエラー出力トークンを検出する。電子装置は、出力シーケンスのエラー可否を判断する前に、予めターゲットアテンションヘッドを決定してもよい。電子装置は、決定されたターゲットアテンションヘッドに基づいて、1つ以上のエラー出力トークンを検出してもよい。

【0118】

一例として、電子装置は、ターゲットアテンションヘッドのエラー可否の判断対象の入力シーケンスと出力シーケンスとの間のアテンション重み行列（以下では、対象アテンション重み行列と称する）を取得することができる。電子装置は、対象アテンション重み行列のアテンション重みとガイド重み行列のアテンション重みとの差が、第1閾値以上である、1つ以上のエラーアテンション重みを検出することができる。例えば、電子装置は、対象アテンション重み行列のアテンション重みとガイド重み行列のアテンション重みのカルバック・ライブラー情報量の差が、第1閾値以上である、1つ以上のエラーアテンション重みを検出してもよい。電子装置は、エラーアテンション重みに対応する出力トークンを前記エラー出力トークンとして決定してもよい。

【0119】

他の例として、翻訳、要約などの入力を様々に見なければならないタスクの場合、電子装置は、対象アテンション重み行列のアテンション重みのうち、以前ステップのアテンション重みと第2閾値以上の類似度を有する1つ以上のエラーアテンション重みを検出してもよい。

【0120】

ステップS540において、電子装置は、エラー出力トークンに基づいて、出力シーケンスを補正する。一実施形態に係る電子装置は、一旦、出力シーケンスの全体を決定した後で、エラー出力トークンを検出し、エラー出力トークンに基づいて、出力シーケンスを補正してもよい。この場合、電子装置は、エラー出力トークンを出力シーケンスから取り除くことで、出力シーケンスのエラーを補正することができる。

【0121】

又は、他の実施形態に係る電子装置は、リアルタイムで出力トークンを予測するステップごとにエラー出力トークンを検出し、エラー出力トークンに基づいて、出力シーケンスを補正してもよい。エラー出力トークンを修正しない場合、電子装置は、エラー出力トークンに基づいて、次の出力トークンを決定するしかない。しかし、電子装置は、エラー出力トークンを出力シーケンスから取り除くだけでなく、エラー出力トークンを取り除いた他の出力トークン候補のうち、次の入力トークンを決定することができる。一実施形態によると、エラー出力トークンを出力したステップの入力トークンを、次の入力トークンとして決定してもよい。「tell me a joke」という文章に対応する音声が入力される場合、場合に依じて、「tell」の後に「me a」をジャンピングし「jok

10

20

30

40

50

e」を出力する現象が発生する場合、「j o k e」を次の入力トークンとして入力せず、「t e l l」をもう1回入力トークンとして入力する。又は、エラー出力トークンを除外し、最も高い確率を有する出力トークン候補を次の入力トークンとして決定することができる。

【0 1 2 2】

一実施形態に係る電子装置は、下記の表3のように、一単語を繰り返す現象を補正したことが分かる。

【表3】

表3

正解	corollary hence it follows that god is not only the cause of things coming into existence but also of their continuing in existence that is in scholastic phraseology god is cause of the being of things <i>essendi rerum</i>
従来認識結果	corollary hence it follows that god is not only the cause of things coming into existence but also of their continuing in existence that is in scholastic phraseology god is cause of the being of things as sindi rerum god is cause of the being of things as sindi rerum god is cause of the being of things as sindi rerum god is cause of the being of things acindi rerum god is cause of the being of the being of things as sindi
一実施形態に係る 電子装置認識結果	corollary hence it follows that god is not only the cause of things coming into existence but also of their continuing in existence that is in scholastic phraseology god is cause of the being of things as sindi re

実施形態に係る方法は、様々なコンピュータ手段を介して実施されるプログラム命令の形態で具現化され、コンピュータ読み取り可能な記録媒体に記録される。記録媒体は、プログラム命令、データファイル、データ構造などを単独又は組み合わせて含む。記録媒体及びプログラム命令は、本発明の目的のために特別に設計して構成されたものでよく、コンピュータソフトウェア分野の技術を有する当業者にとって公知のものであり使用可能なものであってもよい。コンピュータ読み取り可能な記録媒体の例として、ハードディスク、フロッピー（登録商標）ディスク及び磁気テープのような磁気媒体、CD-ROM、DVDのような光記録媒体、プロプティカルディスクのような磁気－光媒体、及びROM、

R A M、フラッシュメモリなどのようなプログラム命令を保存して実行するように特別に構成されたハードウェア装置を含む。プログラム命令の例としては、コンパイラによって生成されるような機械語コードだけでなく、インタプリタなどを用いてコンピュータによって実行される高級言語コードを含む。ハードウェア装置は、本発明に示す動作を実行するために1つ以上のソフトウェアモジュールとして作動するように構成してもよく、その逆も同様である。

【0123】

ソフトウェアは、コンピュータプログラム、コード、命令、又はそのうちの1つ以上の組合せを含み、希望通りに動作するよう処理装置を構成したり、独立的又は結合的に処理装置を命令することができる。ソフトウェア及び／又はデータは、処理装置によって解釈されたり、処理装置に命令又はデータを提供するために、いずれかの種類の機械、構成要素、物理的装置、仮想装置、コンピュータ格納媒体又は装置、もしくは、送信される信号波に永久的又は一時的に具体化することができる。ソフトウェアは、ネットワークに連結されたコンピュータシステム上に分散され、分散した方法で格納されたり、実行され得る。ソフトウェア及びデータは、1つ以上のコンピュータで読出し可能な記録媒体に格納され得る。

【0124】

上述したように、実施形態がたとえ限定された図面によって説明されても、当技術分野で通常の知識を有する者であれば、上記の説明に基づいて、様々な技術的な修正及び変形を適用することができる。例えば、説明された技術が、説明された方法と異なる順で実行され得るし、及び／又は説明されたシステム、構造、装置、回路などの構成要素が、説明された方法と異なる形態で結合又は組み合わせられてもよいし、他の構成要素又は均等物によって置き換え又は置換されたとしても適切な結果を達成することができる。

【0125】

したがって、本発明の範囲は、開示された実施形態に限定されて定められるものではなく、特許請求の範囲及び特許請求の範囲と均等なものなどによって定められるものである。

【符号の説明】

【0126】

- 110 ユーザ端末 120 電子装置
- 210 入力データ
- 215 入力特徴ベクトル
- 220 エンコーダー
- 225 符号化された特徴
- 230 デコーダー
- 240 エンコーダー
- 250 デコーダー
- 310、320、330 アテンション重み行列
- 401、402 電子装置

10

20

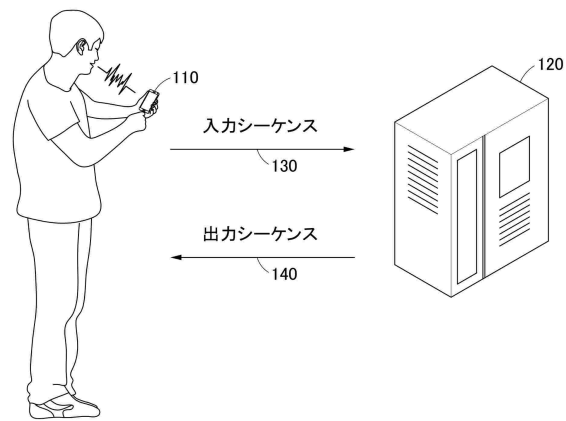
30

40

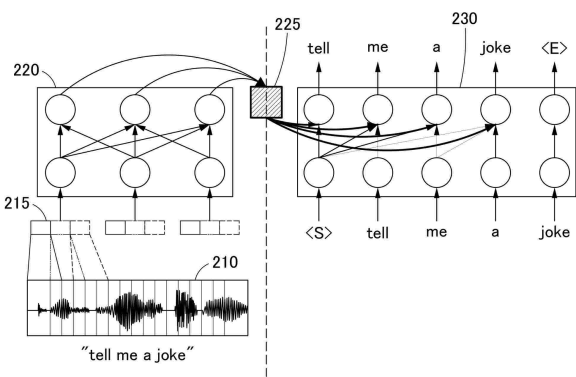
50

【図面】

【図 1】

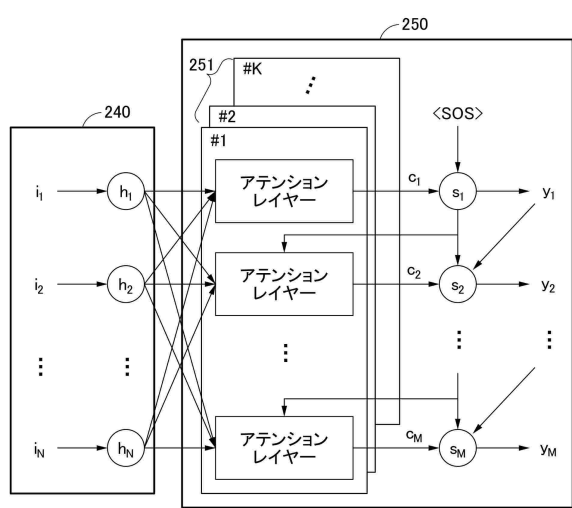


【図 2 A】

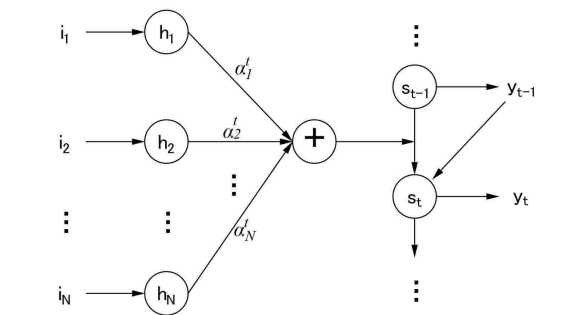


10

【図 2 B】



【図 2 C】



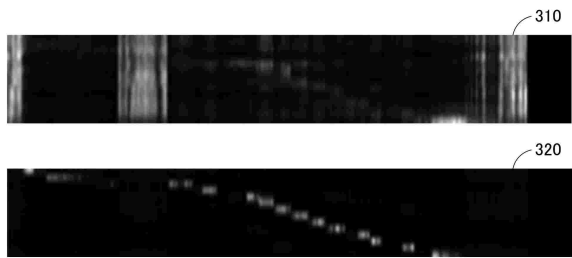
20

30

40

50

【図 3 A】

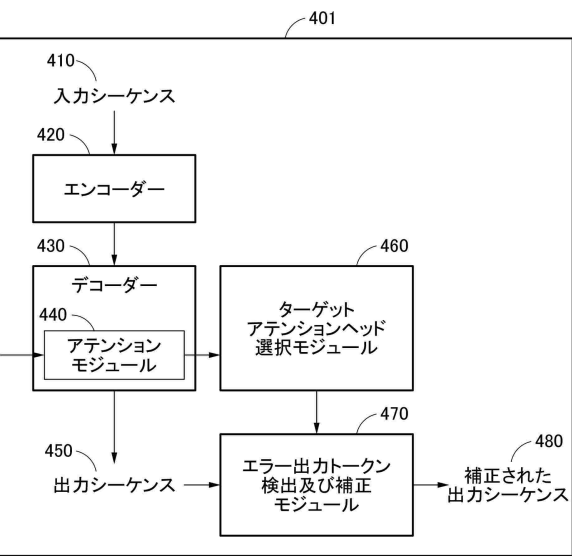


【図 3 B】

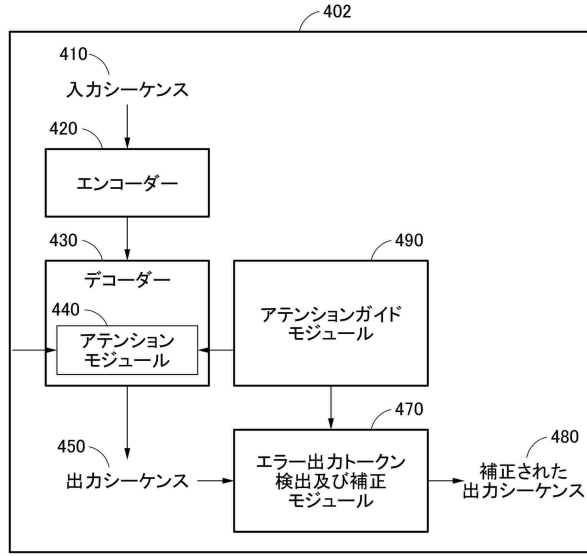


10

【図 4 A】



【図 4 B】



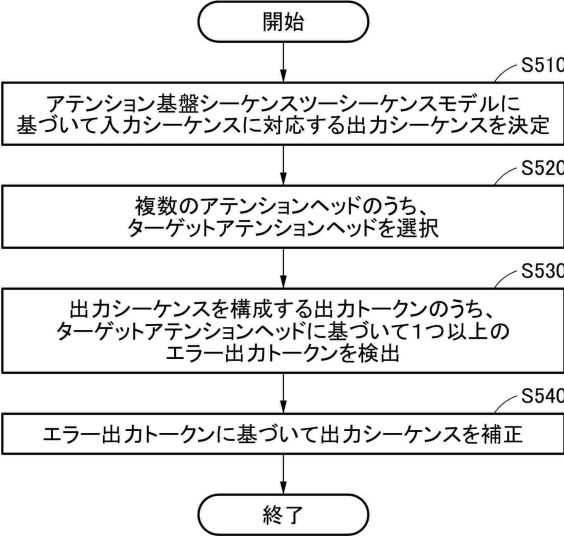
20

30

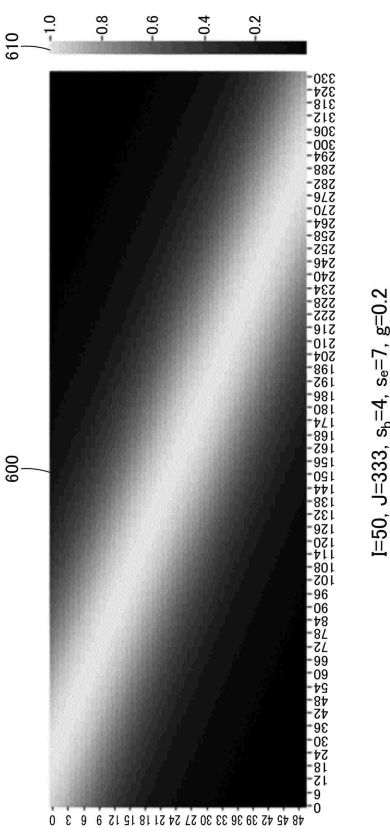
40

50

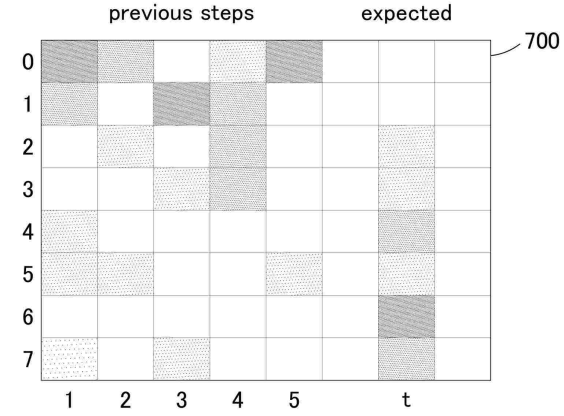
【図 5】



【図 6】



【図 7】



10

20

30

40

50

フロントページの続き

弁理士 宮崎 修

(72)発明者 李 敏重

大韓民国京畿道水原市靈通区三星路130 三星綜合技術院内

審査官 千葉 久博

(56)参考文献 特開2019-159058 (JP, A)

米国特許出願公開第2020/0151544 (US, A1)

中国特許出願公開第109829172 (CN, A)

Jinxi Guo, 外2名, "A SPELLING CORRECTION MODEL FOR END-TO-END SPEECH RECOGNITION", ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019年05月17日, p.5651-5655

(58)調査した分野 (Int.Cl., DB名)

G06N 20/00

G06N 3/08

G10L 15/16