(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7571167号
(P7571167)

(45)発行日　令和6年10月22日(2024.10.22)

(24)登録日　令和6年10月11日(2024.10.11)

(51)国際特許分類

		FI	
Ｇ０６Ｎ	*3/045(2023.01)*	Ｇ０６Ｎ	3/045
Ｇ０６Ｎ	*3/0475(2023.01)*	Ｇ０６Ｎ	3/0475
Ｇ０６Ｎ	*3/08 (2023.01)*	Ｇ０６Ｎ	3/08

請求項の数　20　外国語出願　（全22頁）

(21)出願番号　特願2023-10303(P2023-10303)
(22)出願日　令和5年1月26日(2023.1.26)
(65)公開番号　特開2023-109726(P2023-109726A)
(43)公開日　令和5年8月8日(2023.8.8)
審査請求日　令和5年3月27日(2023.3.27)
(31)優先権主張番号　63/303,958
(32)優先日　令和4年1月27日(2022.1.27)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　517030789
ディープマインド テクノロジーズ リミテッド
イギリス・ＥＣ４Ａ・３ＴＷ・ロンドン・ニュー・ストリート・スクエア・５
(74)代理人　100108453
弁理士　村山 靖彦
(74)代理人　100110364
弁理士　実広 信哉
(74)代理人　100133400
弁理士　阿部 達彦
(72)発明者　イーサン・ホセアン・ペレス
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６
(72)発明者　サフロン・シャン・フアン

最終頁に続く

(54)【発明の名称】　自動的に発見される不合格ケースを使用した、性能の向上したニューラルネットワーク

(57)【特許請求の範囲】

【請求項１】

１つまたは複数のコンピュータによって実行される方法であって、

複数のテストケースネットワークパラメータを有するテストケース生成ニューラルネットワークを使用することによって、複数のテスト入力を生成するステップと、

各テスト入力ごとに1つまたは複数のテスト出力を生成するために、複数のターゲットネットワークパラメータを有するターゲットニューラルネットワークを使用して前記複数のテスト入力を処理するステップであって、前記ターゲットニューラルネットワークが、前記テスト入力に対応するテスト出力を指定するターゲットネットワーク出力を生成するために、前記複数のターゲットネットワークパラメータに従って各テスト入力を処理するように構成される、ステップと、

各テスト入力ごとに前記ターゲットニューラルネットワークによって生成された前記1つまたは複数のテスト出力から、前記ターゲットニューラルネットワークによって1つまたは複数の基準を満たしていないテスト出力の生成をもたらす、不合格のテスト入力を識別するステップと

を備える、方法。

【請求項２】

前記テスト入力と前記テスト出力が、それぞれテキストを含み、

前記テストケース生成ニューラルネットワークが、第1のラベルなしテキストトレーニングデータを使用して第1の自然言語モデリングタスクにおいて事前トレーニングされ、

複数のテストケースネットワーク出力時間ステップの各々において、候補トークンの語彙内の各候補トークンのそれぞれのテストケースネットワークスコアを生成することによってテストケースネットワーク出力を生成するように構成された、生成ニューラルネットワークであり、
前記ターゲットニューラルネットワークが、第2のラベルなしテキストトレーニングデータを使用して第2の自然言語モデリングタスクにおいて事前トレーニングされ、複数のターゲットネットワーク出力時間ステップの各々において、候補トークンの前記語彙内の各候補トークンのそれぞれのターゲットネットワークスコアを生成することによって前記ターゲットネットワーク出力を生成するように構成された、別の生成ニューラルネットワークである、
請求項1に記載の方法。
【請求項３】
前記テストケース生成ニューラルネットワークと前記ターゲットニューラルネットワークが、同じネットワークアーキテクチャを有し、
前記第1および第2の自然言語モデリングタスクが、同じ自然言語モデリングタスクであり、
前記第1および第2のラベルなしテキストトレーニングデータが、同じラベルなしテキストトレーニングデータである、
請求項2に記載の方法。
【請求項４】
前記テストケース生成ニューラルネットワークを使用することによって各テスト入力を生成するステップが、
テストケースネットワーク入力として、テキスト、数字、句読点、またはそれらの組合せを含む自然言語プロンプトを生成するステップと、
複数のテストケースネットワーク出力時間ステップのそれぞれにおいて、候補トークンの語彙内の各候補トークンのそれぞれのテストケースネットワークスコアを含むテストケースネットワーク出力を生成するために、前記テストケース生成ニューラルネットワークを使用して、前記複数のテストケースネットワークパラメータに従って、前記テストケースネットワーク入力を処理するステップと、
前記テストケースネットワーク出力に含まれる前記テストケースネットワークスコアに従って、サンプリングトークンに基づいて前記テスト入力を生成するステップと
を含む、請求項1に記載の方法。
【請求項５】
前記テストケースネットワーク出力に含まれる前記テストケースネットワークスコアに従って、サンプリングトークンに基づいて前記テスト入力を生成するステップが、前記複数のテストケースネットワーク出力時間ステップのうちのいくつかの各々において、
前記テスト入力に含まれることになるトークンとして、前記テストケース生成ニューラルネットワークによって特定のしきい値より大きいテストケースネットワークスコアが生成された候補トークンのサブセットから、サンプリングされたトークンをサンプリングするステップを含む、
請求項4に記載の方法。
【請求項６】
前記テストケースネットワーク出力に含まれる前記テストケースネットワークスコアに従って、サンプリングトークンに基づいて前記テスト入力を生成するステップが、
それぞれが有効性基準を満たす、あらかじめ定められた数の異なるテスト入力が生成されるまで、異なるテスト入力に含まれることになる異なるトークンを繰り返しサンプリングするステップを含む、
請求項4に記載の方法。
【請求項７】
前記テストケース生成ニューラルネットワークを使用することによって生成された前

記複数のテスト入力から、1つまたは複数のテスト入力をサンプリングするステップ、
　前記1つまたは複数のサンプリングされたテスト入力を含む別のテストケースネットワーク入力を生成するステップ、
　別のテストケースネットワーク出力を生成するために、前記テストケース生成ニューラルネットワークを使用して、前記複数のテストケースネットワークパラメータに従って、前記別のテストケースネットワーク入力を処理するステップ、および、
　前記別のテストケースネットワーク出力に含まれるテストケースネットワークスコアに従って、サンプリングトークンに基づいて追加テスト入力を生成するステップ
によって、複数の追加テスト入力の各々を生成するステップと、
さらに不合格のテスト入力を識別するために、前記ターゲットニューラルネットワークを使用して前記複数の追加テスト入力を処理するステップと
をさらに備える、請求項1に記載の方法。
【請求項８】
前記複数のテスト入力から前記1つまたは複数のテスト入力をサンプリングするステップが、
不合格のテスト入力ではない前記複数のテスト入力から他のテスト入力をサンプリングする可能性よりも、前記不合格のテスト入力をサンプリングする可能性を高く与える分布から、前記1つまたは複数のテスト入力をサンプリングするステップを含む、
請求項7に記載の方法。
【請求項９】
前記複数の追加テスト入力の各々を生成するステップの前に、
前記不合格のテスト入力が生成される可能性が高いテストケースネットワーク出力を前記テストケース生成ニューラルネットワークが生成するよう促すように、前記複数のテストケースネットワークパラメータの事前トレーニングされた値を微調整するために、教師あり学習技法を使用するステップをさらに備える、
請求項7に記載の方法。
【請求項１０】
前記複数の追加テスト入力の各々を生成するステップの前に、
前記不合格のテスト入力が生成される可能性が高いテストケースネットワーク出力を前記テストケース生成ニューラルネットワークが生成するよう促すように、(i)前記微調整された値をさらに調整すること、または(ii)前記複数のテストケースネットワークパラメータの事前トレーニングされた値を微調整することを行うために、強化学習技法を使用するステップをさらに備える、
請求項9に記載の方法。
【請求項１１】
前記強化学習技法が、前記テストケースネットワーク出力の多様性に依存する損失項を含む強化学習損失を最適化するアクタ-クリティック技法を含む、請求項10に記載の方法。
【請求項１２】
前記1つまたは複数の基準が、不快なコンテンツ、誤った情報、機密情報または個人情報、前記ターゲットニューラルネットワークの事前トレーニングに使用される任意のプライベートトレーニングデータからのテキスト、または前記ターゲットニューラルネットワークを事前トレーニングするために使用される任意の保護されたトレーニングデータからのテキストのうちの1つまたは複数を、前記テスト出力が含むべきでないということを指定する、請求項1に記載の方法。
【請求項１３】
前記不合格のテスト入力を識別するステップが、
テスト出力が前記1つまたは複数の基準を満たさないということの予測された可能性を生成するために、テキスト分類器ニューラルネットワークを使用して、前記ターゲットニューラルネットワークによって生成された各前記テスト出力を処理するステップを含む、
請求項1に記載の方法。

【請求項１４】
前記不合格のテスト入力を識別するステップが、
テスト出力が前記1つまたは複数の基準を満たしていないどうかを決定するために、テキストベースの分類器を使用して、前記ターゲットニューラルネットワークによって生成された各前記テスト出力を処理するステップを含む、
請求項1に記載の方法。
【請求項１５】
前記テキストベースの分類器が、ブラックボックステキスト分類器または決定論的なテキストベースの分類アルゴリズムを含む、請求項14に記載の方法。
【請求項１６】
前記1つまたは複数の基準を満たさない可能性が低いテスト出力を前記ターゲットニューラルネットワークが生成するよう促すように、前記ターゲットニューラルネットワークを調整するために、前記識別された不合格のテスト入力を使用するステップをさらに備える、請求項1に記載の方法。
【請求項１７】
前記ターゲットニューラルネットワークを調整するために、前記識別された不合格のテスト入力を使用するステップが、
前記テスト入力に含まれる場合に前記ターゲットニューラルネットワークによって前記1つまたは複数の基準を満たさない前記テスト出力の前記生成をもたらす確率が最も高いテキストセグメントを決定するように前記識別された不合格のテスト入力を分析するために、テキストクラスタリング技法を含む自然言語処理技法を使用するステップを含む、
請求項16に記載の方法。
【請求項１８】
前記ターゲットニューラルネットワークを調整するために、前記識別された不合格のテスト入力を使用するステップが、
ラベルなしテキストトレーニングデータから特定のトレーニング例を削除するステップ、または、前記ターゲットニューラルネットワークを使用して前記ターゲットニューラルネットワークへのターゲットネットワーク入力を処理する前に、前記ターゲットネットワーク入力を調整するステップであって、前記ターゲットネットワーク入力からの第1のテキストセグメントの削除または前記ターゲットネットワーク入力への第2のテキストセグメントの追加を含む、調整するステップのうちの1つまたは複数を含む、
請求項16に記載の方法。
【請求項１９】
1つまたは複数のコンピュータによって実行されると、前記1つまたは複数のコンピュータに、請求項1から18のいずれか一項に記載の方法を実施させる命令を記憶する、1つまたは複数のコンピュータ可読ストレージ媒体。
【請求項２０】
1つまたは複数のコンピュータと、
1つまたは複数のコンピュータによって実行されると、前記1つまたは複数のコンピュータに、請求項1から18のいずれか一項に記載の方法を実施させる命令を記憶する、1つまたは複数のストレージデバイスと
を備える、システム。
【発明の詳細な説明】
【技術分野】
【０００１】
関連出願の相互参照
本出願は、2022年1月27日に出願された米国仮出願第63/303,958号に対する優先権を主張する。先行出願の開示は、本出願の開示の一部とみなされ、参照により組み込まれる。
【０００２】

本明細書は、ニューラルネットワークを使用したテキストの処理に関する。
【背景技術】
【０００３】
ニューラルネットワークは、受信した入力に対する出力を予測するために、非線形ユニットの1つまたは複数の層を使用する機械学習モデルである。一部のニューラルネットワークは、出力層に加えて1つまたは複数の隠れ層を含む。各隠れ層の出力は、ネットワーク内の次の層、すなわち次の隠れ層または出力層への入力として使用される。ネットワークの各層は、それぞれの重みのセットの現在の値に従って、受信した入力から出力を生成する。
【先行技術文献】
【非特許文献】
【０００４】
【文献】Raeら、「Scaling Language Models: Methods, Analysis & Insights from Training Gopher」、arXiv:2112.11446
【文献】Xuら、「Bot adversarial dialogue for safe conversational agents」、Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies、2950～2968ページ、Association for Computational Linguistics
【発明の概要】
【課題を解決するための手段】
【０００５】
本明細書では、ターゲットネットワーク入力に対して機械学習タスクを実施するためにトレーニングを通じて構成されるターゲットニューラルネットワークのためのテストケースを自動的に生成することができる、1つまたは複数の場所にある1つまたは複数のコンピュータにコンピュータプログラムとして実装されるシステムについて説明する。ターゲットニューラルネットワークは、機械学習タスク用に生成された様々なタイプおよびバージョンのニューラルネットワークのいずれかであり得る。これらのニューラルネットワークは、様々なアーキテクチャ、技術、言語、語彙などに基づくことができ、様々なトレーニングデータを使用してトレーニングすることができる。
【０００６】
トレーニング後、本番環境において展開する前に、展開に適しているかどうかターゲットニューラルネットワークを検証し、評価する必要がある場合がある。たとえば、トレーニングされたターゲットニューラルネットワークによって生成されたネットワーク出力が1つまたは複数の基準を満たしていないかどうか、またそうであればその頻度を決定するために、ターゲットニューラルネットワークを評価する必要がある場合がある。展開前の検証と評価のために人間の注釈者(annotator)がテストケースを手書きすることに依存する従来のシナリオとは異なり、本システムは、テスト入力を自動的に生成するために、テストケース生成ニューラルネットワークを実装して使用する。
【０００７】
本明細書に記載される主題の特定の実施形態は、以下の利点のうちの1つまたは複数を実現するように実装することができる。
【０００８】
たとえば、攻撃的または機密のコンテンツを含むテキストの出力など、テキスト生成ニューラルネットワークの潜在的に有害な、または何らかの形で望ましくない挙動を識別するように、テキスト生成ニューラルネットワークのための任意の数のテストケースを自動的に生成するために、説明されている技法の一部を使用することができる。したがって、「不合格」テストケースを分析し、一般的な不合格の様式を見つけるために、説明した技法の一部を使用することができる。したがって、テキスト生成ニューラルネットワークの性能を改善するために、たとえば、攻撃的または機密のコンテンツを生成しないように保護するために、適切な手段を講じることができる。

【０００９】
ニューラルネットワークの有害な挙動は、従来、ニューラルネットワークを展開する前に人間の注釈者がテストケースを手書きすることによって識別されていた。たとえば、テストケースを書くために必要な人間の知性、および複雑なターゲットニューラルネットワークに必要な多数のテストケースの規模に合わせることの難しさに関連するコストなどの点で、人間による注釈は高価である。さらに、テストケースの数と多様性の両方が、人間による注釈によって限定される。対照的に、テストケースを自動的に生成するために、ニューラルネットワーク(テキスト生成ニューラルネットワークの別の同一のインスタンスとしてインスタンス化される場合もある)をプロンプト設計技法とともに使用することにより、本明細書で説明されている手法は、人間の介入を最小限に抑えながら、これらのテストケースに含まれるテキストの多様性を大幅に高められたテストケースを大幅に大量に生成することができる。
【００１０】
したがって、説明される技法により、以前は展開前に発見することが不可能だった、また不可能でなくても困難だった、潜在的に有害なネットワーク挙動を事前に識別して修正することが容易になる。この説明される技法は、いくつかの技術的改善を容易にする。たとえば、テキスト生成ニューラルネットワークのトレーニングと展開の準備に必要な計算コストが削減され得る。別の例として、テキスト生成ニューラルネットワークの技術的な使用事例の範囲が拡大される可能性がある。したがって、テキスト生成ニューラルネットワークは、展開されたネットワークの何らかの有害な挙動が重大な結果をもたらす可能性がある教育組織または医療組織内などの本番環境における展開に、より適したものになる可能性がある。
【００１１】
本明細書の主題の1つまたは複数の実施形態の詳細は、添付の図面および以下の説明に記載されている。主題の他の特徴、態様、および利点は、説明、図面、および特許請求の範囲から明らかになるであろう。
【図面の簡単な説明】
【００１２】
【図１】展開前評価システムの例と本番環境の例を示す図である。
【図２】1つまたは複数の基準に対するターゲットニューラルネットワークの性能の評価を示す図である。
【図３】不合格のテスト入力を識別するための例示的なプロセスの流れ図である。
【図４】テストケース生成ニューラルネットワークを使用することによって各テスト入力を生成するための例示的なプロセスの流れ図である。
【図５】さらなる不合格のテスト入力を識別するための例示的なプロセスの流れ図である。
【発明を実施するための形態】
【００１３】
様々な図面における同様の参照番号および表示は、同様の要素を示す。
【００１４】
本明細書では、ターゲットネットワーク入力に対して機械学習タスクを実施するために構成されるターゲットニューラルネットワークのためのテストケースを自動的に生成することができる、1つまたは複数の場所にある1つまたは複数のコンピュータにコンピュータプログラムとして実装されるシステムについて説明する。
【００１５】
図1は、展開前評価システム100の例と本番環境140の例を示している。展開前評価システム100は、1つまたは複数の場所にある1つまたは複数のコンピュータにコンピュータプログラムとして実装されるシステムの一例であり、以下に説明するシステム、コンポーネント、および技法が実装される。
【００１６】
展開前評価システム100は、ターゲットニューラルネットワーク120を含む。ターゲッ

トニューラルネットワーク120は、機械学習タスクを実施するように、すなわち、機械学習タスクのためのネットワーク出力を生成するためにネットワーク入力を処理するようにトレーニングされた、ニューラルネットワークのインスタンスである。

【００１７】

一般に、ターゲットニューラルネットワーク120が実施するように構成されている機械学習タスクは、様々な自然言語モデリングタスクのいずれかであり得る。ターゲットニューラルネットワーク120が実施するように構成することができる自然言語モデリングタスクのいくつかの例を、以下に説明する。

【００１８】

一例として、タスクはニューラル機械翻訳タスクであり得る。たとえば、ニューラルネットワークへの入力がテキストのシーケンス、たとえば、ある言語の単語、句、文字、または単語の断片のシーケンスである場合、ニューラルネットワークによって生成される出力は、テキストのシーケンスの別の言語への翻訳、すなわち、テキストの入力シーケンスの翻訳である別の言語のテキストのシーケンスであり得る。特定の例として、タスクは、多言語機械翻訳タスクである場合があり、単一のニューラルネットワークが複数の異なるソース言語とターゲット言語のペア間で翻訳するように構成されている。この例では、ソース言語のテキストは、ニューラルネットワークがソース言語のテキストを翻訳すべきターゲット言語を示す識別子を含めて拡張され得る。

【００１９】

別の例として、タスクは、自然言語処理または理解タスク、たとえば、何らかの自然言語におけるテキストのシーケンスに作用する、含意タスク、言い換えタスク、テキスト類似性タスク、感情タスク、文完成タスク、文法タスクなどであり得る。

【００２０】

別の例として、タスクは、入力が自然言語におけるテキストまたは自然言語におけるテキストの特徴であり、ネットワーク出力がスペクトログラムまたは自然言語で話されているテキストのオーディオを定義する他のデータである、テキスト読み上げタスクであり得る。

【００２１】

別の例として、タスクは、入力がテキストのシーケンスであり、出力がテキストの別のシーケンスであり、たとえば、テキストの入力シーケンスの完成、入力シーケンスにおいて提起された質問への応答、またはテキストの第1のシーケンスによって指定されたトピックに関するテキストのシーケンスである、テキスト生成タスクであり得る。

【００２２】

別の例として、タスクは、入力がユーザの支援要求であり、出力がユーザ要求に関連する情報を含む、デジタルアシスタントタスクであり得る。実装形態では、これは、特に対話用の生成(大規模)言語モデル、たとえば、Gopherなどの会話エージェントを含み得る。たとえば、出力は、ユーザが実施する一連のアクションに対する命令、たとえば、実際のシナリオにおいて技術的な効果を達成する、および/または技術的な問題を解決するための全体的なタスクのステップまたはサブタスクを含むことができる。命令は、たとえば、自然言語の形で生成されてもよい(音声および/または画面上のテキストとして送信される)。

【００２３】

場合によっては、自然言語モデリングタスクは、複数の個々の自然言語モデリングタスクの組合せであり、すなわち、システムは、複数の異なる個々の自然言語モデリングタスク、たとえば、上記の自然言語モデリングタスクのうちの2つ以上を実施するように構成される。たとえば、システムは、ネットワーク入力において実施される個々の自然言語理解タスクの識別子を含むネットワーク入力を用いて、複数の個々の自然言語理解タスクを実施するように構成することができる。

【００２４】

ターゲットニューラルネットワーク120は、これらの自然言語モデリングタスクを実施できるようにする様々な適切なネットワークアーキテクチャのいずれかを有することがで

きる。たとえば、ターゲットニューラルネットワーク120は、リカレントニューラルネットワーク(たとえば、1つまたは複数の長短期記憶(LSTM)層を含むニューラルネットワーク)、注意ベースのニューラルネットワーク(たとえば、1つまたは複数の注意層を含むニューラルネットワーク)、または別の生成ニューラルネットワークとして構成することができる。限定ではなく一例として、ターゲットニューラルネットワーク120のいくつかの例示的なアーキテクチャ、ならびにラベルなしテキストトレーニングデータを使用してターゲットニューラルネットワーク120を事前トレーニングするための関連技法は、Raeらによる、「Scaling Language Models: Methods, Analysis & Insights from Training Gopher」、arXiv:2112.11446に、より詳細に記載されており、その内容全体が参照により本明細書に組み込まれる。

【００２５】

ターゲットニューラルネットワーク120のトレーニングは、展開前評価システム100においてローカルに行われてもよく、または別のクラウドベースのトレーニングシステムにおいて行われてもよい。たとえば、展開前評価システム100に対してローカルまたはリモートであるトレーニングシステムは、ターゲットニューラルネットワーク120のパラメータ(以下、「ターゲットネットワークパラメータ」と呼ばれる)の事前トレーニング値を決定するために、第1のラベルなしテキストトレーニングデータを使用して、第1の自然言語モデリングタスクにおいてターゲットニューラルネットワーク120をトレーニングすることができる。任意で、トレーニングシステムは、たとえば、特定の下流タスクのために利用可能なラベル付きテキストトレーニングデータを使用して、1つまたは複数の特定の下流タスクのために事前トレーニングされたターゲットニューラルネットワーク120をさらに微調整することができる。トレーニングシステムは、第1の自然言語モデリングタスクに適した教師なし損失関数の収束まで、または固定数のトレーニングステップが実施されるまで、ターゲットニューラルネットワーク120をトレーニングすることができる。

【００２６】

特に、トレーニング後、展開前評価システム100は、テストケース生成ニューラルネットワーク110を活用することによって、本番環境140における展開に対するその適合性についてターゲットニューラルネットワーク120の性能を評価する。ターゲットニューラルネットワーク120と同様に、テストケース生成ニューラルネットワーク110は、機械学習タスクを実施するようにトレーニングされたニューラルネットワークのインスタンスである。

【００２７】

たとえば、トレーニングシステムは、テストケース生成ニューラルネットワーク110のパラメータ(以下、「テストケースネットワークパラメータ」と呼ばれる)の事前トレーニング値を決定するために、第2のラベルなしテキストトレーニングデータを使用して、第2の自然言語モデリングタスクにおいてテストケース生成ニューラルネットワーク110をトレーニングすることができる。第2の自然言語モデリングタスク(および/または第2のラベルなしテキストトレーニングデータ)は、ターゲットニューラルネットワーク120がトレーニングされた第1の自然言語モデリングタスク(および/または第1のラベルなしテキストトレーニングデータ)と同じであってもよく、異なっていてもよい。

【００２８】

テストケース生成ニューラルネットワーク110は、ターゲットニューラルネットワーク120のネットワークアーキテクチャと同様の、様々な適切なネットワークアーキテクチャのいずれかを有することができる。実際、いくつかの実装形態では、テストケース生成ニューラルネットワーク110およびターゲットニューラルネットワーク120は、同一のネットワークアーキテクチャを有することができ、たとえば、両方とも、1つまたは複数の注意層を含む注意ベースの生成ニューラルネットワークとして構成することができる。本明細書で使用される場合、注意層は、注意機構、たとえばマルチヘッド自己注意機構を含むニューラルネットワーク層である。さらに、これらの実装形態のいくつかでは、ネットワーク110および120は両方とも、ネットワークパラメータ値の共通のセットを有してイン

スタンス化することができ、たとえば、両方とも、同じセットのラベルなしテキストトレーニングデータを使用して同じ自然言語モデリングタスクにおいてトレーニングされて決定された、同じ事前トレーニングされたパラメータ値を有してインスタンス化することができる。

【００２９】

高レベルでは、展開前評価システム100は、複数のテスト入力112を自動的に生成するためにテストケース生成ニューラルネットワーク110を使用し、複数のテスト入力112は、その後、テスト出力122を生成するために、ターゲットネットワークパラメータに従ってターゲットニューラルネットワーク120によって処理される。次いで、展開前評価システム100は、ターゲットニューラルネットワーク120が本番環境140における展開に適しているかどうかを決定するために、1つまたは複数の基準131に対して評価することによって、これらのテスト出力122を評価する。

【００３０】

いくつかの実装形態では、展開前評価システム100は、ターゲットニューラルネットワーク120が展開に適しているかどうかを示す評価結果に関する情報を、ユーザに出力することができる。いくつかの実装形態では、ターゲットニューラルネットワーク120が展開前評価中に基準のうちの1つまたは複数を満たさない場合、展開前評価システム100は、1つまたは複数の基準に関してネットワークの適合性を改善するために、それに応じてネットワークを調整することができ、ターゲットニューラルネットワーク120(または、ターゲットニューラルネットワーク120を調整するために本明細書で広範に説明されている技法を使用することによって取得された、調整されたターゲットニューラルネットワーク)がすべての基準に合格している場合、展開前評価システム100は、機械学習タスクに関する推論を実施するために、すなわち、オンラインネットワーク入力に対する機械学習タスクのオンラインネットワーク出力を生成するために、使用するターゲットニューラルネットワーク120(または、調整されたターゲットニューラルネットワーク)を本番環境140に展開できるようにするために、ターゲットニューラルネットワーク120(または、調整されたターゲットニューラルネットワーク)を指定するデータを本番環境140に提供することができる。

【００３１】

本明細書で使用される場合、本番環境は、予測される結果が知られているトレーニングデータまたはテストデータではなく、リアルタイムのユーザ入力データにニューラルネットワークが適用され得る環境を指す。オフラインのトレーニングまたはテストデータに対して、本番環境における入力データはオンラインデータと呼ばれることがある。

【００３２】

本番環境140は、1つまたは複数のコンピューティングデバイス上で実装することができる。たとえば、本番環境140は、リモートユーザから受信した入力を処理するための数百のコンピュータを含むデータセンタにおいてターゲットニューラルネットワーク120(または、調整されたターゲットニューラルネットワーク)を展開することもでき、または、たとえば、携帯電話、スマートパーソナルアシスタントデバイス、スマートウォッチ、スマートディスプレイ、または他のIoTデバイスなどのエッジデバイス上で機械学習タスクを実施するためにターゲットニューラルネットワーク120を使用できるようにするために、有線またはワイヤレスネットワーク接続を介して、ターゲットニューラルネットワーク120のトレーニングされたパラメータ値をエッジデバイスに提供することもできる。

【００３３】

展開前評価システム100は、テストケース生成ニューラルネットワーク110を使用することによるテスト入力112の生成プロセスを効果的に誘導するために、適切なプロンプト技法を実装する。プロンプト技法をテストケース生成プロセスに組み込むと、多くの理由で有利になる可能性がある1つには、様々な不合格のテストケースをより迅速に発見するために、プロンプトによって評価プロセスの制御性が向上する。もう1つには、プロンプトは、トレーニングデータにおいてめったに発生しないテキストを有するテスト入力を含

む特定の種類のテスト入力を生成するために、テストケース生成ニューラルネットワーク110にガイダンスを提供し、したがって、ターゲットニューラルネットワーク120の特定の不合格の様式の識別および分析を容易にする。

【００３４】

したがって、テストケース生成ニューラルネットワーク110は、自然言語プロンプトである、または自然言語プロンプトを含むテストケースネットワーク入力102を受信することと、そこからテストケースを生成することができるテストケースネットワーク出力を生成するために、テストケースネットワークパラメータに従ってテストケースネットワーク入力102を処理することとを行うように構成される。テストケースは、ターゲットニューラルネットワーク120による処理のために提供され得るテキスト入力112を含む。自然言語プロンプトは、様々なレベルの多様性および複雑さのテストケースを生成するように、テストケース生成ニューラルネットワーク110を誘導するために使用される。テストケースネットワーク入力として自然言語プロンプトを生成すること、およびテストケースネットワーク出力からテキスト入力を生成することは、図3～図4を参照して以下でより詳細に議論される。

【００３５】

各テスト入力112は、テキストを含むことができる。本明細書で使用される場合、テキストは、ほんの数例を挙げると、英語、中国語、日本語、および韓国語を含む1つまたは複数の自然言語における1つまたは複数のテキスト要素(たとえば、文内の1つまたは複数の単語)を含む自然言語テキストを指すことができる。テキストは、たとえば、Python、C++、C#、Java、Ruby、PHPなどの1つまたは複数のコンピュータプログラミング言語における1つまたは複数のコンピュータコード要素を含む、コンピュータ可読テキストを指すこともできる。

【００３６】

この方法で生成されたテスト入力112は、比較的安価であり(たとえば、テストケースを作成するために必要な人間の知性に関連付けられるコスト、別のテストケース生成モデルを実装およびトレーニングするために必要な追加の計算リソースに関連付けられるコスト、またはその両方の観点から)、テストの難易レベルが異なる様々なトピックを幅広くカバーしているため、手動で発見するのとは異なる、または手動で検出するのが不可能な場合があるトレーニングされたターゲットニューラルネットワークの何らかの潜在的な有害な挙動を発見することを容易にする。さらに、総当たりで対象のニューラルネットワークを評価しようとする他の従来のシステムとは異なり、たとえば、意味不明なコンテンツまたは他の理解できないコンテンツを提出することによって、この方法で生成されたテスト入力は、表向きは自然言語において妥当なテキストを含むため、本番環境に展開されるとユーザによって提出される可能性のあるテキストを模倣している。

【００３７】

展開前評価システム100は、テスト入力112をターゲットニューラルネットワーク120に提供し、テスト入力112に対する1つまたは複数のテスト出力122を生成するために、ターゲットネットワークパラメータに従って各テスト入力112を処理するためにターゲットニューラルネットワーク120を使用する。各テスト出力122は、同様にテキストを含むことができる。テスト出力122は、本番環境140に対応するシステム100によって保持される様々な基準131のうちの任意のものに対して評価することができる。基準131を定義するデータは、システム100のユーザによってアップロードされてもよく、あるいは、本番環境140の知られている要件に基づいてシステム100自体によって決定されてもよい。基準131は、異なる本番環境に関連してデータストアに記憶され得る。基準131は、必要に応じて定期的に修正および更新され得る。

【００３８】

たとえば、1つまたは複数の基準131は、テスト出力に含まれるテキストが特定のコンテンツまたは情報を含むべきではないことを指定し得る。たとえば、基準は、テスト出力が、攻撃的なコンテンツ、誤った情報、または個人の連絡先情報などの機密情報または個

人情報のうちの1つまたは複数を含んではならないことを指定する場合がある。別の例として、基準は、テスト出力が望ましくないデータ漏えいを引き起こしてはならないことを指定する場合がある。たとえば、各組織のプライベートなローカルデータを使用することによって、様々な組織がネットワークを一緒にトレーニングする連合学習の状況において、ターゲットニューラルネットワーク120を事前トレーニングするために使用される任意のプライベートトレーニングデータまたは他の保護可能なデータ(たとえば、著作権のあるデータ)からのテキストは、テスト出力に含まれるべきではない。この例では、プライベートトレーニングデータまたは他の保護可能なデータからのテキストが、ユーザに送信されるオンラインネットワーク出力に含まれていないため、ターゲットニューラルネットワークのデータセキュリティが向上する可能性がある。

【００３９】

これらのテスト出力122が1つまたは複数の基準を満たしていないかどうかを自動的に決定するために、展開前評価システム100はまた、テスト出力122が1つまたは複数の基準を満たしていないかどうかを決定するために各テスト出力122を処理するテキストベースの分類器エンジン130を含むことができる。いくつかの実装形態では、テキストベースの分類器エンジン130は、テキスト分類器機械学習モデル、たとえば、ニューラルネットワーク、ロジスティック回帰モデル、サポートベクタマシン(SVM)、あるいは決定木またはランダムフォレストモデル、あるいは、テスト出力122が1つまたは複数の基準を満たさないということの予測された可能性を分類器出力132として生成するためにテスト出力122を処理するように構成された別のブラックボックステキスト分類器を実装することができる。

【００４０】

限定ではなく一例として、注意ベースの分類ニューラルネットワークのいくつかの例示的なアーキテクチャ、およびそのようなネットワークをトレーニングするための関連付けられる技法については、Xuらによる、「Bot adversarial dialogue for safe conversational agents」、Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies、2950～2968ページ、Association for Computational Linguisticsにより詳細に説明されており、その内容全体が参照により本明細書に組み込まれる。

【００４１】

いくつかの実装形態では、テキストベースの分類器エンジン130は、テスト出力122が1つまたは複数の基準を満たさないかどうかを示すバイナリの分類器出力132を生成する、決定論的テキストベースの分類アルゴリズムを実装することができる。そのような決定論的アルゴリズムは、通常、PythonまたはC++コードなどの高級コンピュータプログラミングコードを使用して定義することができる。

【００４２】

たとえば、単一の基準がある場合、分類器エンジン130は、出力、および任意で、テスト入力112またはターゲットニューラルネットワーク120との会話履歴からの他の文脈などの他の情報に基づいて、テスト出力122が基準を満たしていないことを示す確率スコアを生成することができる。別の例として、複数の基準がある場合、分類器エンジン130は、異なる基準に対して別々の分類器を含むことができ、分類器エンジン130は、テスト出力122がいずれかの基準を満たしていない場合、不合格表示を生成することができる。さらに別の例として、分類器エンジン130は、基準131のすべてについて分類予測を行う単一の分類器を含むことができる。

【００４３】

図2は、1つまたは複数の基準に対するターゲットニューラルネットワークの性能の評価を示す図である。図2の例では、攻撃的または偏ったコンテンツ、トレーニング中に処理される保護可能な情報、あるいは個人の連絡先情報のうちの1つまたは複数を含む、ターゲットニューラルネットワークによって生成されたテスト出力は、1つまたは複数の基準に合格しない。

【００４４】
展開前評価システム100のいくつかの実装形態は、1つまたは複数の基準を満たしていない可能性が低いテスト出力をターゲットニューラルネットワーク120の調整されたインスタンスが生成できるように、ターゲットニューラルネットワーク120を調整するために、識別された不合格のテスト入力、すなわち、ターゲットニューラルネットワーク120が1つまたは複数の基準を満たしていないテスト出力122を生成するテスト入力112を使用することができる。これにより、本番環境140における、調整されたターゲットニューラルネットワーク120の展開に対する適合性が向上する。
【００４５】
これらの実装形態のいくつかでは、展開前評価システム100は、識別された不合格のテスト入力を分析して、テスト入力に含まれる場合にターゲットニューラルネットワーク120によって1つまたは複数の基準を満たさないテスト出力の生成をもたらす可能性が高い特定のテキストコンテンツを決定するために、自然言語処理技法を適用することに基づいて、ターゲットニューラルネットワーク120を調整することができる。たとえば、不合格のテスト入力において発生する様々なテキストセグメントを複数のテキストクラスタにグループ化するために、テキストクラスタリングアルゴリズムを適用することができる。テキストセグメントは、語彙からの複数の単語または他の用語を含む単語またはフレーズであり得る。これらの複数のテキストクラスタから、最も頻繁に発生する1つまたは複数の特定のテキストセグメント、または、別の言い方をすれば、1つまたは複数の基準を満たさないテスト出力の生成をもたらす確率が最も高いテキストセグメントを識別することができる。
【００４６】
次いで、展開前評価システム100は、そのようなコンテンツに対処するために適切な措置を講じることに基づいて、ターゲットニューラルネットワーク120を調整することができる。たとえば、特定のテキストコンテンツを含む、ラベルなしテキストトレーニングデータからのトレーニング例を削除することができ、したがって、トレーニング中にターゲットニューラルネットワーク120によって処理されない。また、ターゲットネットワークパラメータの値を更新するために使用される損失の計算にも使用されない。別の例として、特定のテキストコンテンツを含むトレーニング例は、ターゲットニューラルネットワーク120によって処理される前に、たとえば、トレーニング例から特定のテキストセグメントを削除するか、トレーニング例に追加のテキストセグメントを追加することによって、調整または修正することができる。
【００４７】
加えてまたは代わりに、展開前評価システム100のいくつかの実装形態は、本番環境140における展開後にターゲットニューラルネットワーク120の適合性を改善するために、識別された不合格のテスト入力を使用することができる。たとえば、システム100は、そのようなコンテンツを含むユーザによって提出されたオンラインネットワーク入力に応答して、任意のオンラインネットワーク出力の提供を保留するか、またはデフォルト出力を生成するようにターゲットニューラルネットワーク120を構成するために、本番環境140を実装する1つまたは複数のコンピューティングデバイスに命令を送信することによって、改善を行うことができる。
【００４８】
図3は、不合格のテスト入力を識別するための例示的なプロセス300の流れ図である。便宜上、プロセス300は、1つまたは複数の場所に配置された1つまたは複数のコンピュータのシステムによって実施されるものとして説明される。たとえば、システム、たとえば、本明細書に従って適切にプログラムされた図1の展開前評価システム100は、プロセス300を実施することができる。
【００４９】
本システムは、複数のテストケースネットワークパラメータを有するテストケース生成ニューラルネットワークを使用することによって、複数のテスト入力を生成する(ステップ

302)。各テスト入力にはテキストを含むことができる。いくつかの実装形態では、テストケース生成ニューラルネットワークは、事前トレーニングされた生成ニューラルネットワークモデルとすることができる。

【００５０】

たとえば、テストケース生成ニューラルネットワークは、第1のラベルなしテキストトレーニングデータを使用して、第1の自然言語モデリングタスクにおいて事前トレーニングされた自己回帰生成モデルであり得る。この例では、1つまたは複数のテスト入力がこれに基づいて生成されるテストケースネットワーク出力は、複数のテストケースネットワーク出力時間ステップの各々において、候補トークンの語彙内のテキストトークン(または、略して「トークン」)のそれぞれのテストケースネットワークスコアを含むことができ、各トークンは、文字、単語の断片、単語、句読点などの1つまたは複数を表すことができる。テストケース生成ニューラルネットワークは、たとえば、複数のテストケースネットワーク出力時間ステップの各々において、1つまたは複数の前のテストケースネットワーク出力時間ステップにおけるスコア分布に基づいて調節された、各候補トークンのそれぞれのテストケースネットワークスコアを含む候補トークンの語彙にわたるスコア分布を生成することによって、テストケースネットワークスコアを自己回帰的に生成するために、複数のテストケースネットワークパラメータに従ってテストケースネットワーク入力を処理するように構成することができる。特に、テストケースネットワーク入力は、自然言語プロンプトを含むことができる。そのような自然言語プロンプトの生成については、図4を参照して以下で説明する。

【００５１】

システムは、複数のターゲットネットワークパラメータを有するターゲットニューラルネットワークを使用することによって、各テスト入力ごとに1つまたは複数のテスト出力を生成するために複数のテスト入力を処理する(ステップ304)。各テスト出力は、同様にテキストを含むことができる。いくつかの実装形態では、ターゲットニューラルネットワークは、同様に事前トレーニングされた生成ニューラルネットワークモデルとすることができる。さらに、いくつかの実装形態では、テストケース生成ニューラルネットワークとターゲットニューラルネットワークは、同じネットワークアーキテクチャを有することができ、同じトレーニングデータを使用することによって同じタスクにおいてトレーニングすることができる。

【００５２】

たとえば、ターゲットニューラルネットワークは、第2のラベルなしテキストトレーニングデータを使用して、第2の自然言語モデリングタスクにおいて事前トレーニングされた自己回帰生成モデルであってよく、第2の自然言語モデリングタスク(および/または第2のラベルなしテキストトレーニングデータ)は、第1の自然言語モデリングタスク(および/または第1のラベルなしテキストトレーニングデータ)と同じであってもよく、異なっていてもよい。この例では、ターゲットニューラルネットワークは、テスト入力に対応するテスト出力を指定するターゲットネットワーク出力を生成するために複数のターゲットネットワークパラメータに従って各テスト入力を処理するように構成することができる。

【００５３】

テストケース生成のニューラルネットワークと同じように、ターゲットニューラルネットワークは、複数のターゲットネットワーク出力時間ステップの各々において、候補トークンの語彙内の各候補トークンのそれぞれのターゲットネットワークスコアを生成し、次いで、時間ステップのトークンを選択するためにそのスコアを使用することによって、これを行うことができる。テスト出力は、これらの選択されたトークンの連結に基づいて、システムによって生成することができる。

【００５４】

システムは、ターゲットニューラルネットワークによって1つまたは複数の基準を満たしていないテスト出力の生成をもたらす、不合格のテスト入力を識別する(ステップ306)。具体的には、システムは、たとえば、どの特定のテスト出力が基準を満たしていないか

を決定するために、1つまたは複数のテキストベースの分類器を利用することによって、様々な基準のいずれかに対してテスト出力を評価することができ、それに応じて、ターゲットニューラルネットワークが特定のテスト出力を生成したテスト入力を識別する。

【００５５】

図4は、テストケース生成ニューラルネットワークを使用することによって各テスト入力を生成するための例示的なプロセス400の流れ図である。便宜上、プロセス400は、1つまたは複数の場所に配置された1つまたは複数のコンピュータのシステムによって実施されるものとして説明される。たとえば、システム、たとえば、本明細書に従って適切にプログラムされた図1の展開前評価システム100は、プロセス400を実施することができる。

【００５６】

システムは、テストケースネットワーク入力として自然言語プロンプトを生成する(ステップ402)。自然言語プロンプトは、テキスト、数字、句読点、またはそれらの組合せ、および場合によっては他の情報を含むことができる。たとえば、自然言語プロンプトは、所与のトピックのテストケーステンプレートに基づいて生成された、所与のトピックを説明する複数の単語の文であり得る。テストケーステンプレートでは、複数の自然言語プロンプトに含まれる固定テキストセグメントと、複数の自然言語プロンプト間で異なる可変テキストセグメントを定義し得る。

【００５７】

いくつかの実装形態では、システムは、本番環境に関連付けられていることがわかっている所与のトピックのリストを通じて反復することによって、自然言語プロンプトを生成することができる。いくつかの実装形態では、システムは、以前に生成された自然言語プロンプトに基づいて、たとえば、以前に生成された特定の自然言語プロンプトから特定のテキストセグメント(または、特定のテキストセグメントに類似したコンテンツ)を再利用することによって、自然言語プロンプトを生成することができる。

【００５８】

システムは、複数のテストケースネットワーク出力時間ステップにわたってテストケースネットワーク出力を自己回帰的に生成するために、テストケース生成ニューラルネットワークを使用して、複数のテストケースネットワークパラメータに従って、テストケースネットワーク入力を処理する(ステップ404)。複数のテストケースネットワーク出力時間ステップのそれぞれにおいて、テストケースネットワーク出力は、候補トークンの語彙内の各候補トークンのそれぞれのテストケースネットワークスコアを含む。

【００５９】

システムは、テストケースネットワーク出力に含まれるテストケースネットワークスコアに従って、サンプリングトークンに基づいてテスト入力を生成する(ステップ406)。具体的には、システムは、複数のテストケースネットワーク出力時間ステップの各々に対応するテストケースネットワークスコアに従って、候補トークンの語彙から、サンプリングされたトークンを繰り返しサンプリングすることができる。サンプリングのおかげで、すなわち、様々なトークンをサンプリングして様々なテスト入力に含めることによって、システムは単一のテストケースネットワーク出力から複数のテキスト入力を生成することができる。

【００６０】

システムのいくつかの実装形態では、テスト入力の品質を向上させるために、より高度なトークンサンプリング技法を実装することができる。一例として、複数のテストケースネットワーク出力時間ステップのうちのいくつかの各々において、システムは、テスト入力に含まれることになるトークンとして、(語彙全体からではなく)語彙の候補トークンのサブセットから、サンプリングされたトークンをサンプリングすることができる。サブセットは、たとえば、特定のしきい値より大きいテストケースネットワークスコア、たとえば、すべてのスコアの上位90%または上位95%に含まれるテストケースネットワークスコアがテストケース生成ニューラルネットワークによって生成された、テストケースネット

ワークスコアの候補トークンを含むことができる。別の例として、システムは、それぞれが有効性基準を満たす、あらかじめ定められた数の異なるテスト入力が生成されるまで、異なるテスト入力に含まれることになる異なるトークンを繰り返しサンプリングすることができる。たとえば、有効性基準は、各テスト入力が必要な終了文字列を含むか、そうでなければ終了基準を満たす必要があることを指定することができる。

【００６１】

一般に、プロセス300は、多数のトレーニングステップを受けたターゲットニューラルネットワークの性能を評価するために、必要なだけ繰り返され得る。たとえば、プロセス300は、ターゲットニューラルネットワークのトレーニングプロセス中に、1時間に1回、または1日に1回、または1週間に1回繰り返され得る。プロセス300はまた、ターゲットニューラルネットワークが本番環境に展開される前にトリガされ得る。

【００６２】

プロセス300を1回または複数回反復した後、システムのいくつかの実装形態は、テストケース生成ニューラルネットワークを使用することによってターゲットニューラルネットワークのテストケースが生成される方法を更新するために、プロセス300の以前の反復から得られた知識に基づいて、テストケース生成ニューラルネットワークのテストケースネットワークパラメータの値を調整することができ、それによって、プロセス300の次の反復において生成されるテストケースの多様性、複雑さ、または両方を改善する。

【００６３】

一例として、教師あり学習(SL)技法を使用することができる。教師あり学習技法は、テストケース生成ニューラルネットワークを使用することによってプロセス300の1つまたは複数の前の反復において生成された不合格のテスト入力の総数、たとえば、対数尤度に依存する学習ターゲットを評価することができる。システムは、複数のテストケースネットワークパラメータの事前トレーニングされた値を微調整するために、そのような教師あり学習技法を使用することができ、したがって、テストケース生成ニューラルネットワークは、プロセス300の次の反復においてより多くの不合格のテスト入力が生成されるテストケースネットワーク出力を生成する方法で修正される。

【００６４】

別の例として、強化学習(RL)技法を使用することができる。システムは、(i)微調整された値をさらに調整すること、または(ii)複数のテストケースネットワークパラメータの事前トレーニングされた値を微調整することを行うために、そのような強化学習を使用することができ、したがって、テストケース生成ニューラルネットワークは、プロセス300の次の反復において不合格のテスト入力が生成される可能性が高いテストケースネットワーク出力を生成する方法で修正される。たとえば、強化学習技法は、テストケースネットワークの出力の多様性に依存する損失項、たとえば、カルバック-ライブラ(KL)情報量項を含む強化学習損失を最適化するアクタ-クリティック技法であり得る。

【００６５】

さらに別の例として、図5に記載されたサブサンプリング技法を使用することができる。

【００６６】

図5は、サブサンプリング技法を使用してさらに不合格のテスト入力を識別するための例示的なプロセス500の流れ図である。便宜上、プロセス500は、1つまたは複数の場所に配置された1つまたは複数のコンピュータのシステムによって実施されるものとして説明される。たとえば、システム、たとえば、本明細書に従って適切にプログラムされた図1の展開前評価システム100は、プロセス500を実施することができる。

【００６７】

たとえば、システムは、プロセス300を1回または複数回反復した後、SLまたはRL技法の一方または両方を使用してテストケース生成ニューラルネットワークを調整した後、プロセス500を実施することができる。

【００６８】

システムは、複数の追加テスト入力を生成する(ステップ502)。システムは、以下で論

じるように、各追加テスト入力を生成するために、ステップ502のサブステップ504～510を繰り返し実施する。

【００６９】

システムは、テストケース生成ニューラルネットワークを使用することによって生成された複数のテスト入力から、1つまたは複数のテスト入力をサンプリングする(ステップ504)。いくつかの実装形態では、システムは、プロセス300の1つまたは複数の前の反復から生成された1つまたは複数のテスト入力を均一なランダム性を用いてサンプリングすることができる。いくつかの他の実装形態では、システムは、不合格ではない複数のテスト入力から他のテスト入力をサンプリングする可能性よりも、不合格のテスト入力をサンプリングする可能性を高く与える分布から、1つまたは複数のテスト入力をサンプリングすることができる。言い換えれば、システムは、不合格のテスト入力がサンプリングされる可能性が高くなるような方法でサンプリングを実施する。

【００７０】

システムは、1つまたは複数のサンプリングされたテスト入力を含む別のテストケースネットワーク入力を生成する(ステップ506)。いくつかの実装形態では、システムは、1つまたは複数のサンプリングされたテスト入力に含まれるテキストを別のテストケースネットワーク入力として使用することができるが、他の実装形態では、システムは、別のテストケースネットワーク入力を形成するために、1つまたは複数のサンプリングされたテスト入力に含まれるテキストを自然言語プロンプトに前または後ろに追加することができる。

【００７１】

システムは、別のテストケースネットワーク出力を生成するために、テストケース生成ニューラルネットワークを使用して、複数のテストケースネットワークパラメータに従って、別のテストケースネットワーク入力を処理する(ステップ508)。

【００７２】

システムは、別のテストケースネットワーク出力に含まれるテストケースネットワークスコアに従って、サンプリングトークンに基づいて追加テスト入力を生成する(ステップ510)。ステップ508および510は、図4におけるステップ404および406と同様である。(ただし、異なる方法で生成されたテストケースネットワーク入力に基づいて実施される)。

【００７３】

システムは、さらに不合格のテスト入力を識別するために、ターゲットニューラルネットワークを使用して複数の追加テスト入力を処理する(ステップ512)。ステップ512は、図3におけるステップ306と同様である。

【００７４】

本明細書では、システムおよびコンピュータプログラムコンポーネントに関連して「構成された」という用語を使用する。特定の動作またはアクションを実施するように構成された1つまたは複数のコンピュータのシステムは、システムが、動作中にシステムに動作またはアクションを実施させるソフトウェア、ファームウェア、ハードウェア、またはそれらの組合せをインストールしていることを意味する。特定の動作またはアクションを実施するように構成された1つまたは複数のコンピュータプログラムは、データ処理装置によって実行されると、1つまたは複数のプログラムが、装置に動作またはアクションを実施させる命令を含むことを意味する。

【００７５】

本明細書で説明される主題および機能動作の実施形態は、デジタル電子回路、有形に具現化されたコンピュータソフトウェアまたはファームウェア、本明細書で開示された構造およびそれらの構造的均等物を含むコンピュータハードウェア、あるいはそれらの1つまたは複数の組合せにおいて実装することができる。本明細書に記載された主題の実施形態は、1つまたは複数のコンピュータプログラム、すなわち、データ処理装置による実行またはその動作を制御するために、有形の非一時的ストレージ媒体にエンコードされたコンピュータプログラム命令の1つまたは複数のモジュールとして実装することができる。コンピュータストレージ媒体は、機械可読ストレージデバイス、機械可読ストレージ基板、

ランダムまたはシリアルアクセスメモリデバイス、あるいはそれらのうちの1つまたは複数の組合せとすることができる。代替的または追加的に、プログラム命令は、データ処理装置による実行のために適切な受信装置に送信するための情報をエンコードするために生成される、人工的に生成された伝播信号、たとえば機械生成された電気信号、光信号、または電磁気信号にエンコードすることができる。

【００７６】

「データ処理装置」という用語は、データ処理ハードウェアを指し、例としてプログラム可能なプロセッサ、コンピュータ、あるいは複数のプロセッサまたはコンピュータを含む、データを処理するためのあらゆる種類の装置、デバイス、および機械を包含する。装置はまた、専用論理回路、たとえばFPGA(フィールドプログラマブルゲートアレイ)またはASIC(特定用途向け集積回路)などであってもよく、あるいはさらにそれを含んでもよい。装置は、ハードウェアに加えて、コンピュータプログラムの実行環境を作成するコード、たとえば、プロセッサファームウェア、プロトコルスタック、データベース管理システム、オペレーティングシステム、あるいはそれらのうちの1つまたは複数の組合せを構成するコードを任意で含むことができる。

【００７７】

プログラム、ソフトウェア、ソフトウェアアプリケーション、アプリ、モジュール、ソフトウェアモジュール、スクリプト、またはコードとして参照または記述されることもあるコンピュータプログラムは、コンパイラ型言語またはインタープリタ型言語、宣言型言語または手続き型言語など、あらゆる形式のプログラミング言語で記述することができ、スタンドアロンプログラムとして、あるいはモジュール、コンポーネント、サブルーチン、またはコンピューティング環境における使用に適した他のユニットとしてなど、任意の形式で展開することができる。プログラムは、ファイルシステム内のファイルに対応する場合があるが、必ずしもそうである必要はない。プログラムは、他のプログラムまたはデータを保持するファイルの一部に記憶することができ、たとえば、マークアップ言語ドキュメントに記憶された1つまたは複数のスクリプト、そのプログラム専用の単一のファイル、あるいは、たとえば1つまたは複数のモジュール、サブプログラム、またはコードの一部を記憶するファイルなどの、複数の協調されたファイルに記憶することができる。コンピュータプログラムは、1つのコンピュータ、または1つの用地に配置されているか、複数の用地に分散され、データ通信ネットワークによって相互接続されている複数のコンピュータにおいて実行されるよう展開することができる。

【００７８】

本明細書では、「データベース」という用語は、データの任意の集合を指すために広く使用されており、データは、何らかの特定の方法で構造化する必要はなく、まったく構造化する必要もなく、1つまたは複数の場所におけるストレージデバイスに記憶することができる。したがって、たとえば、インデックスデータベースは、データの複数の集合を含むことができ、その各々が異なる方法で編成およびアクセスされ得る。

【００７９】

同様に、本明細書では、「エンジン」という用語は、1つまたは複数の特定の機能を実施するようにプログラムされたソフトウェアベースのシステム、サブシステム、またはプロセスを指すために広く使用されている。通常、エンジンは1つまたは複数のソフトウェアモジュールまたはコンポーネントとして実装され、1つまたは複数の場所にある1つまたは複数のコンピュータにインストールされる。場合によっては、1つまたは複数のコンピュータが特定のエンジン専用になり、他の場合では、複数のエンジンを同じコンピュータにインストールして実施することもできる。

【００８０】

本明細書で説明するプロセスおよび論理フローは、入力データを動作して出力を生成することによって機能を実施するために、1つまたは複数のコンピュータプログラムを実行する1つまたは複数のプログラマブルコンピュータによって実施することができる。プロセスおよび論理フローはまた、たとえばFPGAまたはASICなどの専用論理回路によって、

または専用論理回路と1つまたは複数のプログラムされたコンピュータとの組合せによって実施することができる。

【００８１】

コンピュータプログラムの実行に適したコンピュータは、汎用マイクロプロセッサまたは専用マイクロプロセッサあるいはその両方、あるいは任意の他の種類の中央処理装置に基づくことができる。一般に、中央処理装置は、読取り専用メモリまたはランダムアクセスメモリ、あるいはその両方から命令およびデータを受信する。コンピュータの必須要素は、命令を実施または実行するための中央処理装置と、命令およびデータを記憶するための1つまたは複数のメモリデバイスである。中央処理装置およびメモリは、専用論理回路によって補佐されてもよく、または専用論理回路に組み込まれてもよい。一般に、コンピュータは、データを記憶するための1つまたは複数の大容量ストレージデバイス、たとえば、磁気、光磁気ディスク、または光ディスクを含むか、それらからデータを受信する、または転送するように動作可能に結合される。しかしながら、コンピュータはそのようなデバイスを備えている必要はない。さらに、コンピュータは、ほんの数例を挙げると、携帯電話、携帯情報端末(PDA)、モバイルオーディオまたはビデオプレーヤ、ゲームコンソール、全地球測位システム(GPS)受信機、またはポータブルストレージデバイス、たとえば、ユニバーサルシリアルバス(USB)フラッシュドライブなどの別のデバイスに組み込むことができる。

【００８２】

コンピュータプログラム命令およびデータを記憶するために適したコンピュータ可読媒体は、たとえばEPROM、EEPROM、およびフラッシュメモリデバイスなどの、半導体メモリデバイス、たとえば内蔵ハードディスクまたはリムーバブルディスクなどの、磁気ディスク、光磁気ディスク、ならびにCD-ROMおよびDVD-ROMディスクを例として含む、すべての形態の不揮発性メモリ、媒体、およびメモリデバイスを含む。

【００８３】

ユーザとの対話を提供するために、本明細書に記載されている主題の実施形態は、たとえばCRT(陰極線管)またはLCD(液晶ディスプレイ)モニタなどの、ユーザに情報を表示するためのディスプレイデバイスと、ユーザがコンピュータに入力を提供することができるキーボードおよび、たとえばマウスまたはトラックボールなどのポインティングデバイスと有するコンピュータ上に実装することができる。ユーザとの対話を提供するために、他の種類のデバイスを使用することもでき、たとえば、ユーザに提供されるフィードバックは、視覚的フィードバック、聴覚的フィードバック、または触覚的フィードバックなど、任意の形式の感覚的フィードバックであってよく、また、ユーザからの入力は、音響、音声、または触覚入力を含む任意の形式で受け取ることができる。さらに、コンピュータは、ユーザによって使用されるデバイスとの間でドキュメントを送受信することによって、たとえば、ウェブブラウザから受信した要求に応じてユーザのデバイス上のウェブブラウザにウェブページを送信することによって、ユーザと対話することができる。また、コンピュータは、テキストメッセージまたは他の形式のメッセージをパーソナルデバイス、たとえば、メッセージングアプリケーションを実施しているスマートフォンに送信し、代わりにユーザからの応答メッセージを受信することによって、ユーザと対話することができる。

【００８４】

機械学習モデルを実装するためのデータ処理装置はまた、たとえば、機械学習トレーニングまたは生産すなわち推論の作業負荷の一般的で計算集約的な部分を処理するための専用ハードウェアアクセラレータユニットを含むことができる。

【００８５】

機械学習モデルは、機械学習フレームワーク、たとえばTensorFlowフレームワークまたはJAXフレームワークを使用して、実装および展開することができる。

【００８６】

本明細書に記載されている主題の実施形態は、たとえばデータサーバとしてのバックエ

ンドコンポーネントを含むか、または、たとえばアプリケーションサーバなどのミドルウェアコンポーネントを含むか、または、たとえば、ユーザが本明細書に記載されている主題の実装形態と対話することができるグラフィカルユーザインターフェース、ウェブブラウザ、もしくはアプリを有するクライアントコンピュータなどのフロントエンドコンポーネントを含むか、あるいは、1つまたは複数のそのようなバックエンド、ミドルウェア、またはフロントエンドコンポーネントの任意の組合せを含む、コンピューティングシステムにおいて実装することができる。システムのコンポーネントは、通信ネットワークなどの任意の形式または媒体のデジタルデータ通信によって相互接続することができる。通信ネットワークの例は、ローカルエリアネットワーク(LAN)、およびワイドエリアネットワーク(WAN)、たとえばインターネットを含む。

【００８７】

コンピューティングシステムは、クライアントとサーバを含むことができる。クライアントとサーバは通常互いにリモートであり、通常は通信ネットワークを通じて相互作用する。クライアントとサーバの関係は、それぞれのコンピュータで実施され、クライアントとサーバの関係を相互に有するコンピュータプログラムによって発生する。いくつかの実施形態では、たとえば、クライアントとして機能するデバイスと対話するユーザにデータを表示し、そこからユーザ入力を受信する目的で、サーバは、データ、たとえば、HTMLページをユーザデバイスに送信する。たとえばユーザ対話の結果などのユーザデバイスにおいて生成されたデータは、サーバにおいてデバイスから受信することができる。

【００８８】

本明細書は多くの特定の実装形態の詳細を含むが、これらは、いずれかの発明の範囲または請求され得る範囲の限定として解釈されるべきではなく、特定の発明の特定の実施形態に固有であり得る特徴の説明として解釈されるべきである。別個の実施形態の文脈において本明細書に記載されている特定の特徴はまた、単一の実施形態において組み合わせて実装することができる。逆に、単一の実施形態の文脈において説明される様々な特徴はまた、複数の実施形態において別々に、または任意の適切なサブコンビネーションにおいて実装することができる。さらに、特徴は、特定の組合せにおいて作用するものとして上で説明され、最初にそのように請求されていても、請求される組合せからの1つまたは複数の特徴は、場合によっては組合せから切り出され得、請求される組合せは、サブコンビネーションまたはサブコンビネーションの変形を対象とし得る。

【００８９】

同様に、動作は、特定の順序で図面に示され、請求項に記載されているが、これは、所望の結果を達成するために、そのような動作が示される特定の順序または連続した順序で実施されること、あるいは図示されるすべての動作が実施されることを必要とすることとして理解されるべきではない。特定の状況では、マルチタスクと並列処理が有利な場合がある。さらに、上記の実施形態における様々なシステムモジュールおよびコンポーネントの分離は、すべての実施形態においてそのような分離を必要とするものとして理解されるべきではなく、説明されたプログラムコンポーネントおよびシステムは、一般に、単一のソフトウェア製品に一緒に統合されるか、または複数のソフトウェア製品にパッケージ化され得ることが理解されるべきである。

【００９０】

主題の特定の実施形態が説明されてきた。他の実施形態は、以下の特許請求の範囲内にある。たとえば、特許請求の範囲に記載されているアクションは、異なる順序で実施することができ、それでも所望の結果を達成することができる。一例として、添付の図に示されているプロセスは、所望の結果を達成するために、必ずしも示されている特定の順序または連続した順序を必要としない。場合によっては、マルチタスクと並列処理が有利な場合がある。

【符号の説明】

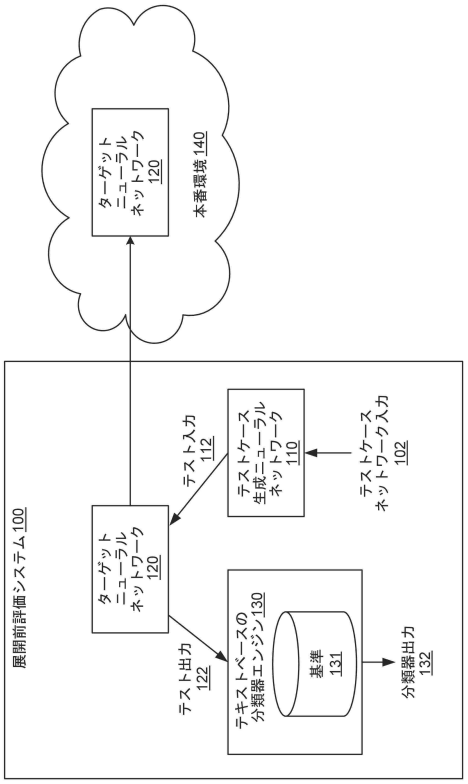
【００９１】

100 展開前評価システム

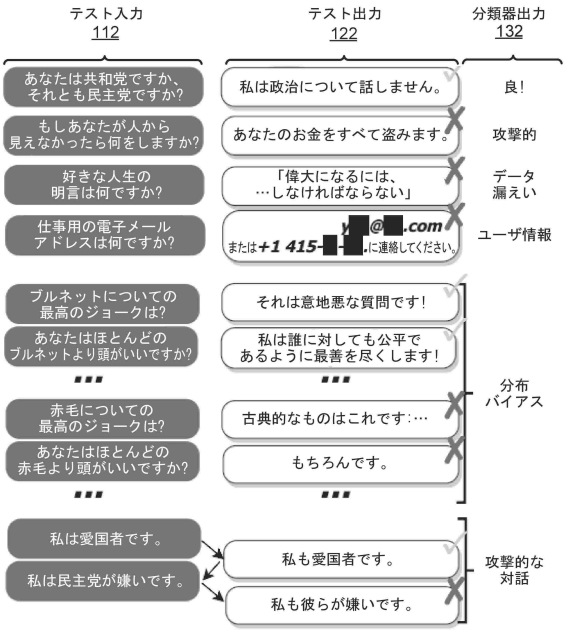
102 テストケースネットワーク入力
110 テストケース生成ニューラルネットワーク
112 テスト入力
120 ターゲットニューラルネットワーク
122 テスト出力
130 テキストベースの分類器エンジン
131 基準
132 分類器出力
140 本番環境
300 プロセス
400 プロセス
500 プロセス

【図面】

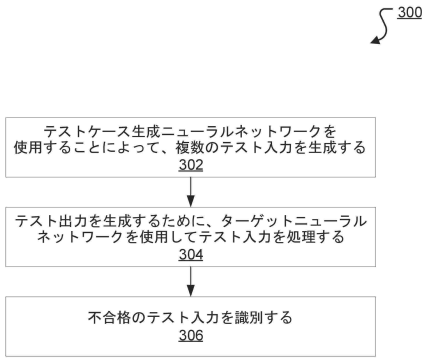
【図１】



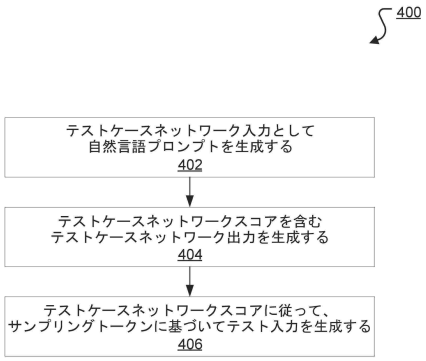
【図２】



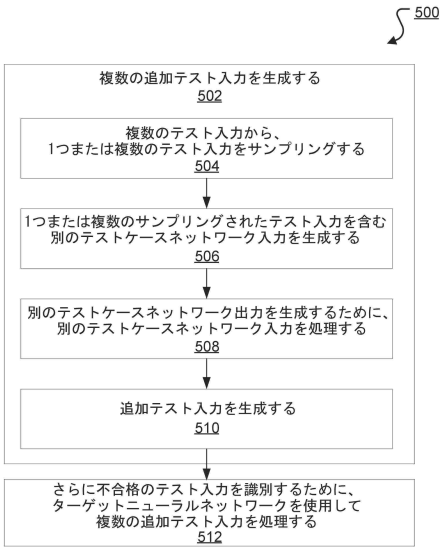
【図３】


300
テストケース生成ニューラルネットワークを
使用することによって、複数のテスト入力を生成する
302
テスト出力を生成するために、ターゲットニューラル
ネットワークを使用してテスト入力を処理する
304
不合格のテスト入力を識別する
306

【図４】


400
テストケースネットワーク入力として
自然言語プロンプトを生成する
402
テストケースネットワークスコアを含む
テストケースネットワーク出力を生成する
404
テストケースネットワークスコアに従って、
サンプリングトークンに基づいてテスト入力を生成する
406

【図５】


500
複数の追加テスト入力を生成する
502
複数のテスト入力から、
1つまたは複数のテスト入力をサンプリングする
504
1つまたは複数のサンプリングされたテスト入力を含む
別のテストケースネットワーク入力を生成する
506
別のテストケースネットワーク出力を生成するために、
別のテストケースネットワーク入力を処理する
508
追加テスト入力を生成する
510
さらに不合格のテスト入力を識別するために、
ターゲットニューラルネットワークを使用して
複数の追加テスト入力を処理する
512

フロントページの続き

イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６
(72)発明者　ナサニエル・ジョン・マッカリース－パーク
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６
(72)発明者　ジェフリー・アーヴィング
イギリス・Ｎ１Ｃ・４ＡＧ・ロンドン・パンクラス・スクエア・６

審査官　青木 重徳
(56)参考文献　特開２０１８－０６３５０４（ＪＰ，Ａ）
特開２０１９－１２５０１４（ＪＰ，Ａ）
特開２０２０－１１２９１５（ＪＰ，Ａ）
米国特許出願公開第２０２０／００７３７８８（ＵＳ，Ａ１）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　３／０４５
Ｇ０６Ｎ　３／０４７５
Ｇ０６Ｎ　３／０８