

進行中の作業

BERT はブラウザのユーザー インターフェイスの使い方を学習しますか？

統一されたビジョンと言語の BERT を使用したマルチステップ タスクの探索

梶崎太一、相澤晶子国立情報学研究所、

東京都千代田区、日本 総合研究大学院大学、神奈川県葉山市 {iki,aizawa}

@nii.ac.jp

概要

事前トレーニング済みのトランスフォーマーは、タスクにとらわれない表現により、統合されたマルチタスク モデルの優れた基盤となります。事前トレーニング済みのトランスフォーマーは、多くの場合、テキストからテキストへのフレームワークと組み合わされて、単一のモデルで複数のタスクを実行します。グラフィカル ユーザー インターフェイス (GUI) を介してタスクを実行することは、視覚と言語の入力を伴う複数ステップのタスクを含む、さまざまなタスクに対応するために行うことができるもう 1 つの方法です。ただし、事前にトレーニングされた Transformer と GUI による実行を組み合わせた論文はほとんどありません。このギャップを埋めるために、Web ページで実装された GUI を複数のステップで操作することによって、モデルがタスクを実行するフレームワークを調査します。ページ遷移のあるタスク ページのないタスク ページを開発し、フレームワークの BERT 拡張を提案します。これらのタスク ページを使用して BERT 拡張機能を共同でトレーニングし、次のことを確認しました。

(1) モデルは、ページ遷移がある場合とない場合の両方のタスク ページを使用することを学習しました。(2) ページ遷移のない 5 つのタスクのうち 4 つでは、モデルは、ブラウザーを使用しない元の BERT のパフォーマンスの 75% 以上を実行します。(3) モデルは目に見えないタスクでは効果的に一般化されませんでした。これらの結果は、GUI を介して BERT を多段階タスクに微調整できることを示唆しており、一般化可能性には改善の余地があります。コードはオンラインで入手可能になります¹

1 はじめに

以前の研究では、自然言語を処理するためのモデルを統一して、タスク固有の構造を削減することにより、学習した知識の伝達を促進しようと試みてきました。たとえば、ラドフォード等。(2018); De vlin等。(2019)は、一般的な構造を持つ言語モデルである Transformer (Vaswani et al., 2017)が有効であることを示唆しています。ラッフェル等。(2020)モデルが受け取って生成するタスクを問題に変換するテキストからテキストへのフレームワークを提案

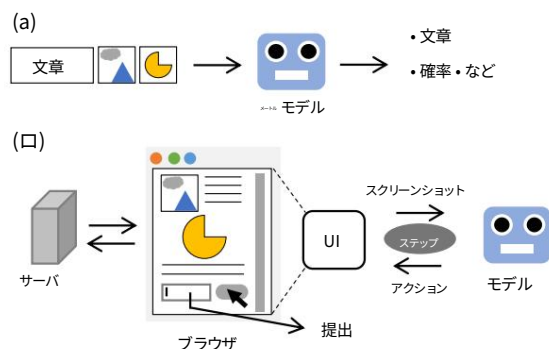


図 1: フレームワークの比較。(a) 典型的なフレームワークは、モデルが一連のテキストと画像を取得して出力を生成する単一ステップのタスクを想定しています。(b) WoB のようなフレームワークでは、構造化されたコンテンツ、ハイパーリンク、およびスクリプトを許可する Web ページとしてタスクを作成します。ページのデザインによって、メッセージを送信する方法が決まります (ボタンの選択やテキストの入力など)。モデルは、ブラウザー ユーザー インターフェイス (BUI) を使用して複数の手順でタスクを完了します。モデルはスクリーンショットを撮り、各ステップのアクションを出力します。(例: クリックまたはキーストローク)。

文章。チョラ。(2021)画像に対応するために、テキストからテキストへのフレームワークの入力を拡張しました。

ただし、統一モデルに関する既存の研究は限られています。まず、Cho らによって提案されたモデル。(2021)はテキストの線形シーケンスといくつかの画像を入力として使用しますが、そのモデルはレイアウトで入力を処理するには設計されていません。

第 2 に、既存の統合モデルは単一ステップのタスクを前提としています。これらのモデルを複数タスク (1 つのドキュメントを読み、その後不足している情報を検索するなど)に適用する場合は、タスク固有の設計を完了する必要があります。後者の課題は、複数ステップのタスクでトランスフォーマーを使用する方法がまだ完全に確立されていないため、対処がより困難です。トランスフォーマーベースのモデルは、言語理解(Wang et al., 2019)、質問応答(Rajpurkar et al., 2016)、視覚的質問応答(Antol et al., 2015)、参照表現

¹リポジトリ: https://github.com/Alab-NII/bertbui_pub/

理解(Kazemzadeh et al., 2014)にもかかわらず、これらは 1 ステップのタスクです。

この研究では、タスクを実行する可能性を探るために、次の質問を調査します。言語モデルは、視覚的な入力として表現されるグラフィカル ユーザー インターフェイスを介してタスクを完了できますか？

グラフィカルユーザーインターフェイスの操作については、Shi et al. (2017)は、タスクが Web ページとして表現され、モデルがブラウザー UI を介してそれらのタスクを完了するWorld of Bits (WoB) プラットフォームを導入しました。

私たちは彼らの仕事に従い、言語モデルを同様のタスク フレームワークに適用します (図1)。

ブラウザ間の相互作用を定式化します

モデル (§ 3) を作成し、GLUE (Wang et al., 2019)、SQuAD (Rajpurkar et al., 2016)、VQA (Antol et al., 2015)などの既存のデータセットに基づいてタスク ページを作成します (§ 4)。私たちのタスクには、単一ページだけでなく、アクションの目的を多様化するためにページジャンプを必要とする複数ページのタスクも含まれます。

BERT (Devlin et al., 2019)拡張を導入し、単純な記憶メカニズムとアクションの事前トレーニングを行います (§ 5)。私たちの実験では、マルチタスク設定でモデルをトレーニングします。モデルがフレームワークで学習できるかどうかを検証し、

同じBERTに基づく他のモデルと比較してください。事前トレーニングと記憶メカニズムが効果的であることを示し、見えないタスクを解決するモデルの能力を分析します (§ 6)。

私たちの貢献:

- 言語モデルがその言語知識を保持したままGUI 操作にどの程度うまく転送されるかを評価するためのタスク セットを構築します。
- BERT 拡張機能を導入し、さまざまなタスク (GLUE、SQuAD、VQA、マルチページ タスク) を共同で学習する能力を実証しました。

2 関連作品

2.1 統一モデルの実行スタイル

統合モデルは、タスク固有の構造を削減して、さまざまなタスクを共同で学習することを促進し、学習した知識をタスク間で共有できるようにすることを目的としています。Lu et al., 2019; Li et al., 2020; Fan and Bansal, 2019; Chen et al., 2020; Su et al., 2020)。変圧器を使用した統合モデルは大きな注目を集めています。

統合トランスフォーマーは、タスク実行の観点から分類できます: タスク固有の head と text gen

2 それは一種のマルチタスク学習ですが(Caruana, 1997; Ruder, 2017)、中心的なタスクを持たないことがよくあります。

レーション スタイル。タスク固有のヘッド スタイルは、タスク間でほとんどのモデル ウェイトを共有し、各タスクにヘッドを提供します。ViLBERT-MT (Lu et al., 2020)と UniT (Hu and Singh, 2021)はこのスタイルを使用します。

テキスト生成スタイルでは、テキスト生成を使用して、タスク間の出力の違いを埋めます。

GPT-2 (Radford et al., 2019)と GPT-3 (Brown et al., 2020)は、プロンプトを変更することで、事前トレーニング済みの大規模なモデルがテキスト領域でマルチタスクできることを示しています。

T5 (Raffel et al., 2020)や、T5 を視覚に拡張したVL-T5 (Choi et al., 2021)もこのスタイルを採用しています。テキスト生成スタイルは、原則として、テキストからテキストへの形式で表現できるすべてのタスクに適用できます。私たちの WoB のようなフレームワークも非常に柔軟です。モデルは、一般的なアクションを介してタスクを定義する Web ページを操作します。その結果、モデル構造を維持しながら、ブラウザ画面でレンダリングできるものにタスクを拡張します。

このスタイルを BUI アクション スタイルと呼びます。

2.2 視覚と言語のタスク

World of Bit プラットフォーム(Shi et al., 2017)は、Web タスクの強化学習環境であり、私たちの仕事に最も関連しています。プラットフォーム上の MiniWoBベンチマークとその拡張機能である MiniWoB++ (Liu et al., 2018)は、Web タスクのモデルを研究するために広く使用されています。これらは、ボタンのクリックやテキストの入力などの単純なタスクから、フライトの予約などのより複雑なタスクで構成されています。

これらのベンチマークに関する以前の研究は、アーキテクチャ/ポリシー(Liu et al., 2018; Gur et al., 2018; Jia et al., 2019; Humphreys et al., 2022)、構文解析(Srivastava et al., 2020)、および環境生成(Gur et al., 2021)、タスク手順における LM の言語知識の使用を調査した研究はほとんどありません。

ドキュメント AIは、ドキュメントを自動的に読み取り、理解し、分析するための技術です(Cui et al., 2021)。私たちの仕事は、HTML 文書に関する研究に関連しています。田中ら。(2021);チェン等。(2021)は、Web ページ上の読解データセットを提案しました。呉ら。(2021);李等。(2021a)は、HTMLドキュメントの事前トレーニング済みモデルを提案しました。ドキュメントはさまざまな方法で処理されますが(スクリーンショットを使用したり、要素の階層を組み込んだりするなど)、以前の研究では視覚的にリッチなレイアウトが考慮されていました。私たちの焦点は、モデルとドキュメントの間の相互作用にあります。

UIモデリングは新たなトピックであり、Bai et al. (2021);彼等。(2021)より良い表現を得るために、モバイル デバイス用のUI モデルを事前にトレーニングしました。

アプリケーションの種類を予測したり、類似した UI コンポーネントを取得したりするなど、タスクを理解するという観点から UI について説明します。李等。(2021b)は、UI に関する質問に答えることができるマルチタスク UI モデルを提案しました。質問には「次の画面に移動」などのコマンドが含まれていますが、それらは単一ステップのコマンドに限定されています。対照的に、私たちのモデルはアクションを繰り返し生成することで UI を使用します。

視覚言語ナビゲーション (VLN) (Anderson et al., 2018; Das et al., 2018; Shridhar et al., 2020)は、部屋などの物理空間で指示に従うモデルを研究しています。VLN タスクは、V&L モデルを使用したアクション生成で進歩しました。

最近の研究では、事前にトレーニングされた LM を使用して構造をエンコードしました(Li et al., 2019b; Majumdar et al., 2020; Hong et al., 2021; Qi et al., 2021)。ただし、ターゲットが物理空間であるため、視覚入力に長いテキストが含まれることはめったにありません。長いテキストとアクションを伴うビューの組み合わせは、依然として課題です。

3 ブラウザUIによるタスク策定

この調査では、ブラウザという用語は、Web ページにアクセスするためのソフトウェアを指します。ブラウザは Web ページをレンダリングし、ハイパーリンクがクリックされると新しいページに移動し、ページ上のスクリプトを内部的に実行します。

私たちの定式化は、上で実行されるブラウザに焦点を当てています。パーソナルコンピュータ³。ブラウザの入力デバイスはマウスとキーボードであり、ブラウザはスクリーンショットを提供すると想定しています。各ステップで、モデルはスクリーンショットから Web ページの状態を部分的に観察してアクションを出力します。スクリーンショットでは、カーソル位置が点として描かれています。アクションをブラウザに適用し、ブラウザが内部計算 (レンダリング、ナビゲートなど)を完了するまで一定時間(~ 500ms)4待ちました。

続いて、次のスクリーンショットを撮ります。結論として、ページ s_i の可視領域のスクリーンショットとステップ i でのモデルのアクション a_i を想定すると、モデルは s_{i+1} から a_i を予測します: $a_i = \text{Model}(s_0, \dots, s_i + 1 = \text{Browser}(a_i))$

$$s_i, a_0, \dots, a_{i-1}),$$

私たちの定式化はマルコフを想定していないため、上記の式で以前のすべてのスクリーンショットとアクションを考慮します。モデルはそれらすべてを使用する必要はありません。

³Firefox (<https://www.mozilla.org/en-US/firefox/>) ブラウザに採用された Selenium (<https://www.selenium.dev/>)、ブラウザ操作の自動化ツールである を使用して、モデルのアクションをブラウザに適用しました。

⁴内部サーバーが使用されました。外部サーバーへのアクセス追加の時間が必要になる場合があります。

スコープ アクション名 説明 mouse

MOVETO(x, y) カーソルを (x, y) に移動 mouse CLICK 右クリック。

SPACE タイトル(単語)タイプの単語の文字 SPACE タイプのボックススペース キー ENTER タイプの Enter を左のビューに移動します RIGHT ビューを右のビューに移動します UP ビューを上ビューに移動し 移動 DOWN ビューを下に

表 1: 定義されたアクション。MOVETO および TOKEN は、括弧内に指定された引数を取ります。

固定サイズのスクリーンショット。可変サイズまたはスケールのスクリーンショットを使用してページ全体を入力する代わりに、固定サイズのスクリーンショットを使用して、モデルに表示領域を移動するアクションを与えます。このようなアクションは、追加のパーツを動的にロードし、予期しない長い入力を回避するページに適しています。

行動。表1は、定義されたアクションを示しています。アクションには、マウスの使用、キーストローク、および可視領域の移動が含まれません。キーストロークの単位は、モデルの語彙です。モデルは、ステップごとに 1 つのアクションを選択します。したがって、タスクでテキストボックスに文を入力する必要がある場合、モデルはカーソルをテキストボックスに移動し(MOVETO)、テキストボックスをクリックし(CLICK)、トークンを入力します (TOKEN)。

WoBの違い。WoB ではスクリーンショット、HTML ページのドキュメント オブジェクト モデル (DOM)を入力に使用しますが、私たちの定式化ではスクリーンショットのみを使用します。WoB に続く作品(Liu et al., 2018; Gur et al., 2018; Jia et al., 2019)では click(element) などの DOM 固有のアクションを使用していますが、元の WoB に沿った低レベルのアクションを使用しています。

さらに、アクションの定式化は、次の点で WoB の定式化とは異なります。まず、定式化では、テキストを入力する単位として、文字の代わりに言語モデルのサブワードを使用します。次に、Ctrl キーなどの特殊キーを使用しません。3 つ目は、マウス操作にドラッグ、ドロップ、または左クリックを使用しないことです。最後の 2 つの制限に注意してください。

簡単にするためだけです。

4 つのタスク ページ

このセクションでは、タスク用に作成した Web ページについて説明します。ページのレイアウトに制限はありませんが、わかりやすくするために、説明書、主な内容、および回答フォームを含むレイアウトを使用しました。タスクの例には 1 つの答えがあると仮定しました。図2は、作成したタスク ページをまとめたものです。このセクションでは、タスク ページの種類と、トレーニング用のゴールド アクションを取得する方法について説明します。

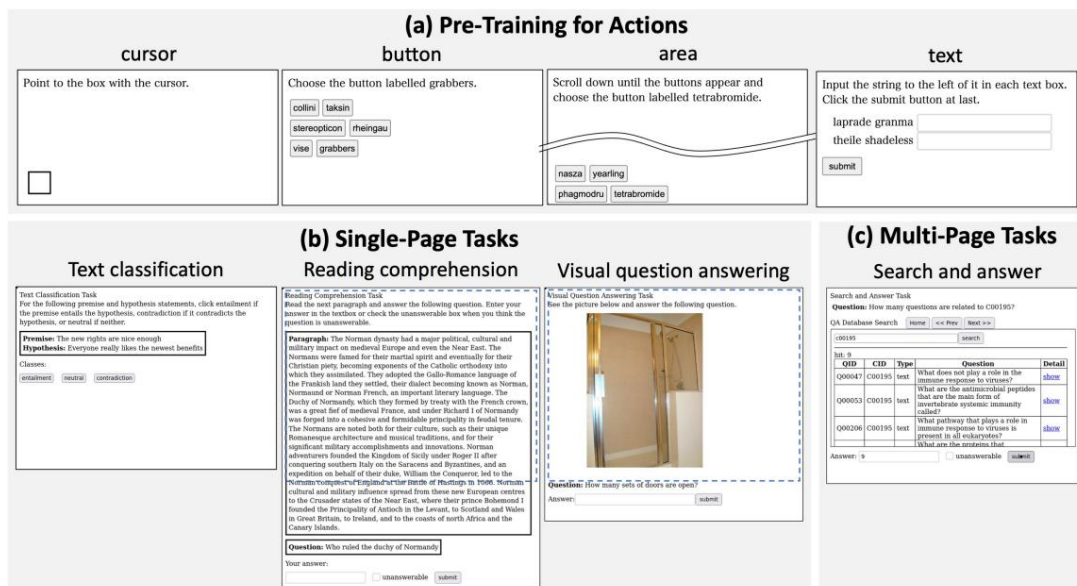


図 2: 3 種類のタスク ページと例。(a) アクションの事前トレーニング。area タスクでは、ボタンが表示されるまでモデルをスクロールする必要があるように、命令とボタンの間に空白が存在します。(口) 単一ページのタスク。青い点線で囲まれた四角形は、最初の可視領域を表します。(c) 複数ページのタスク。モデルは、外側のページに埋め込まれた子フレーム内でページ遷移を行うことができます。

4.1 タスクページの種類

(a) 行動のための事前訓練 (PTA)。クリック可能なボタンの使用など、インターフェイスの使用に関する事前知識は、モデルがそのような知識と推論 (読解など) を同時に学習する状況を回避することによって、BUI でのタスクのより効率的な学習を支援する可能性があります。アクションの事前トレーニングを導入しました。これは、カーソルの移動、ボタンのクリック、表示領域の移動、およびテキストの入力に焦点を当てた一連の小さなタスクです。図2のように、PTA課題では画面上部に指示が書かれており、その指示に従えばモデルは成功します。テンプレートを使用してタスク インスタンスを生成しました (付録C.2 を参照)。

(b) 単一ページのタスク。モデルが BUI の従来のタスクをどの程度解決できるかを評価するために、既存のデータセットに基づいてこのタイプのタスクを作成しました。自然な言語理解には GLUE (Wang et al., 2019) と SQuAD (Rajpurkar et al., 2016, 2018) を使用し、視覚的グラウンディングには VQA (Antol et al., 2015) を使用しました。このタイプのタスク ページには、ページのスクロールと回答の送信が含まれます。

の形式に一致する回答フォームを選択しました

それらのデータセット。GLUE (分類) にはボタンを使用し、SQuAD と VQA (質問応答) にはテキスト ボックスを使用しました。成功の条件は、元のデータセットの正しい答えを提出することです。

(c) 複数ページのタスク。このタイプは、手続的なタスクにより焦点を当てるためにページ遷移を導入します

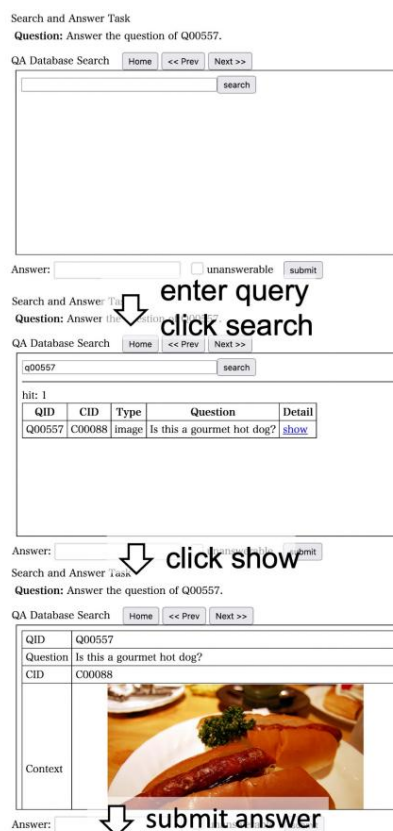


図 3: SA-A タスクのページの例。すべての SA タスクは、ページデザインを共有します。タスク ページには、最初のページ、検索結果、および詳細ページが含まれます。クエリによって結果が変わります。モデルは、質問に答えるためにこれらのページ間をジャンプする必要があります。

その BUI が有効にします。Search and Answer (SA) タスクを設計しました (図3)。タスクは、SQuAD と VQA からコンテキスト (段落または画像) と質問をサンプリングして、質問応答タスクのデータベースを作成しました。コンテキストと質問に一意の ID を割り当てました。SA のタスク ページは、検索 UI で照会できるデータベースの 1 つにリンクされています。タスクの目的は、検索 UI を使用してデータベースに関する質問に回答することです。モデルがさまざまな質問を処理できるかどうかを検証するために、4 つのグループを用意しました。

- SA-H: CID に関連する質問はいくつありますか? 特定のコンテキスト ID (CID) を照会し、ヒット数に答える必要があります。
- SA-Q: QID の問題は何ですか? 特定の質問 ID (QID) に対応する質問の識別が必要です。
- SA-QID: QUESTION の QID は?

特定の質問に対応する QID の識別が必要です。

- SA-A: QID の質問に教えてください。特定の QID に対応する質問に答える必要があります。

SA-H、-Q、-QID は検索結果から直接回答できますが、SA-A はモデルが詳細ページを表示して回答を生成する必要があります。付録 C.3 に詳細が記載されています。

4.2 ゴールド シーケンスの作成教師あ

り学習を使用して、提案された BUI モデルをトレーニングしました。アクション スクリーンショット ペアのゴールド シーケンスを記録するために、各タスクのルールを手動で作成し、ルールに従ってブラウザにロードされたタスク ページを操作しました。タスク ページのコンテンツを一度識別し、回答を送信するためのアクションを実行するようにルールを設計しました。

個々のルールは次のとおりです。ご了承ください
各ルールはさらにアクションに分類されます。

- 接着剤。ボタンが表示されるまで下にスクロールします (すべてのコンテンツを表示するため)。正しいボタンをクリックしてください。

SQuAD と VQA。送信フォームが表示されるまで下にスクロールします。

質問に答えられない場合は、答えられないにチェックを入れます。それ以外の場合は、テキスト ボックスに答えを入力します。送信ボタンをクリックします。

- SA-H。命令で CID を照会します。送信
ヒットレコードの数。
- SA-Q。命令で QID を照会します。クエリに関連するレコード
が表示されるまで下にスクロールします。
対応する質問を送信します。

- SA-QID。SA-Q ルールの質問と QID を入れ替えました。アクションの数を減らすために、質問の最初の 3 つのトークンのみを使用しました。
- SA-A。命令で QID を照会します。「表示」リンクをクリックして、コンテキストを表示します。コンテキスト全体を表示して、QID に対応する質問に対する回答を生成します。答えを提出してください。

5 プイバート

このセクションでは、事前トレーニング済みの BERTsmall を拡張してブラウザ UI を操作する方法について説明します。BUI セットアップで必要な複数の長いシーケンスのため、標準サイズの事前トレーニング済みプロフェッショナルモデル (例: Lay outLM) の代わりに小さな言語モデルが使用されました。図4 に示すように、視覚、言語、記憶、およびいくつかの拳補助が融合 BERT で融合され、次のものが得られます。

アクション。事前トレーニング済みの BERT の重みに基づいて融合 BERT の重みを初期化し、PTA タスクを使用してモデルを事前トレーニングしました。

5.1 ビジョン入力

Huang et al. に似た、事前に訓練された畳み込みニューラルネットワークのグリッド機能を使用しました。 (2020) 、速度とデータ量を考慮して、各ステップでスクリーンショットをエンコードし、その後フリーズを付けました。

トレーニング済みの ResNet (He et al., 2016) トレー

ニング可能な完全に接続されたレイヤー。

5.2 言語入力

単語を別のモダリティとして扱うために、スクリーンショットから単語を検出し、

BERT uncased tok enizer6 を使用して単語をサブ単語に変換します。単語ページの単語を span タグを挿入するこ

付録 D.1 に検出例を示します。

このエミュレーションはテキスト コンテンツとラで動作しますが、ボタンのベル、テキストボックス内のテキストをキャプチャしません。この制限にもかかわらず、このエミュレーションを使用して計算時間を短縮し、検出エラーを考慮する必要を回避します。モデルは、最後のアクション トークンを通じて入力を認識できることに注意してください。

場所の埋め込み。スクリーンショットで場所を示すために、各サブワード埋め込みに場所埋め込みディンディングを追加しました。ワードバウンディングボックス7

⁵ PyTorch Vision にバンドルされた事前トレーニング済みの ResNet18。

⁶ Huggingface のトークナイザー (<https://github.com/huggingface/tokenizers>) の BertWordPieceTokenizer を使用しました。事前に訓練された BERT の語彙を使用します。

⁷ (中心 x、中心 y、幅、および高さ)。すべての要素はスクリーンショットの幅または高さによって正規化されます。

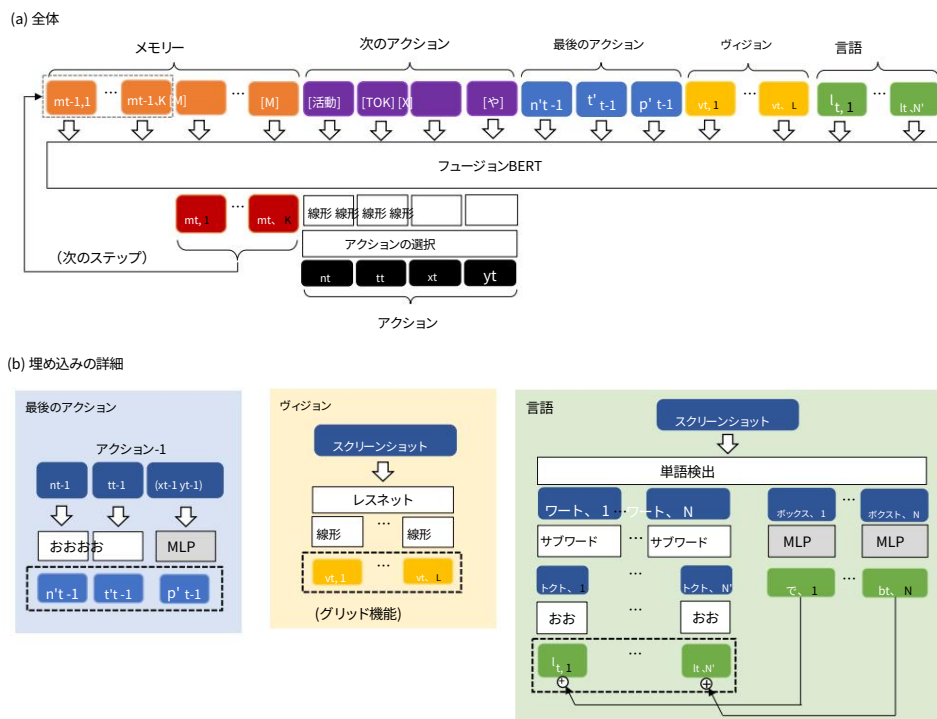


図 4: BUI-BERT の概要。この図は、モデルがステップ t のアクションを予測する方法を示しています。(a) モデルは、事前にトレーニングされた BERT (Fusion BERT) で構成されます。この BERT は、視覚、言語、記憶、およびいくつかの補助トークンを使用して、次のアクションを出力します。Fusion BERT は、元の BERT と同じ方法で位置とセグメントの埋め込みを追加します (ビューでは省略されています)。(b) モデルは、ディープ ニューラル ネットワークを使用して、アクション (3 つのトークン)、ビジョン (L トークン)、および言語埋め込み (N 単語から N トークン) を取得します。

トレーニング可能な MLP によってエンコードされました。単語内のサブ単語は、単語の埋め込み位置を共有します。

5.3 メモリメカニズム

私たちのモデルはメモリを使用して以前の入力を処理します。 K がメモリ トークン長のハイパーパラメータであるとして (このペーパーでは 64 を使用しました)。私たちの記憶メカニズムは $2 \times K$ 埋め込みを使用しました。前半の K は以前のメモリ出力からコピーされ、後半はトレーニング可能なワンホット ベクトルである $[M]$ 埋め込みで埋められました。入力を融合した後、次のステップのために後半に対応する K 個のエンコードされた埋め込みを保持しました。

トレーニング中、ステップ数が最大 (私たちの研究では 50) を超えないようにしながら、メモリ埋め込みを繰り返し入力しました。

5.4 補助入力

最後のアクション。最後のアクションは、アクション名、カーソル位置、⁸ およびサブワードの埋め込みで表されます。トレーニング可能なワンホットを使用しました

アクション名とサブワード embeddings のベクトル。を使用してカーソル位置をエンコードしました

⁸カーソル位置またはサブワードに関係のないアクションの場合、それらの埋め込みはゼロで埋められました。

幅と高さをゼロに設定することにより、単語位置と同じ MLP。

次のアクション。[ACT],[TOK],[X],および[Y] トークンにトレーニング可能なワンホット ベクトルを追加し、これらのトークンを入力して次のアクションを予測しました。

5.5 次のアクションの予測と損失

Fusion BERT が [ACT],[TOK],[X],および[Y] トークンをエンコードした埋め込みから、次のアクションを予測しました。エンコードされた埋め込み D_{ign} が e_{act} , e_{Tok} , e_x , および e_y であるとして。最初に e_{act} からアクション名を分類しました。次に、Fusion BERT のボキャブラリのトークン id を e_{Tok} から TOKEN に分類し、ピクセル座標 9 を e_x および e_y から MOVETO に分類しました。すべての埋め込みは、トレーニング可能な線形レイヤーを使用してクラス分布に射影されます。

トレーニング中、アクション名、トークン、 x , および y に Softmax クロスエントロピー損失を使用しました。これらはミニバッチで均等に追加されました。

$$L_{mb} = L_{name} + L_{token} + L_x + L_y,$$

ここで、 \cdot は非パッド ラベルの平均を示します。

損失関数はメモリ使用量を明示的に参照していませんが、結果は次のようになることに注意してください。

⁹ $x \in \{1, \dots, \text{画面の幅}\}$ および $y \in \{1, \dots, \text{画面の高さ}\}$ 。

モデル	カーソルボタン	エリアテキスト	0.90	0.56
BUI-BERTsmall 1.00	BUI-		0.87	0.62
BERTmedium 1.00	確率レベル		0.88	0.43

表 2: アクションの事前トレーニングでの BUI-BERT の正確な一致精度。モデルは 4 つのタスクで共同でトレーニングされました。ボタン数の逆数をボタンとエリアタスクのチャンスレベルとして表示します。モデルが応答を送信しなかった、つまりタイムアウトになったケースの比率は、すべてのタスクで 0.03 未満でした。

セクション6.1と6.3は、モデルが学習することを示しています。この損失でメモリを使用します。

6 実験10

まず、PTA タスクで BUI モデルをトレーニングして、モデルを事前トレーニングしました。次に、マルチタスク設定で BUI モデルをトレーニングしました。その後、BUI モデルをさまざまなタスク スタイルのモデルと比較しました。最後に、モデルを分析しました。

6.1 アクションの事前トレーニング

中小規模の BUI-BERT11 を 60,000 のトレーニング例を使用して PTA タスクでトレーニングしました。設定。両モデルのメモリ長は64でした。画面を640px×448pxに設定し、スクリーンショットを半分になりサイズしてから ResNet18 に入力しました。最大エポックは 50 でした。各エポックの終わりに検証損失を追跡し、検証損失が最小のモデルを評価に使用しました。

実際のブラウザで確認してください。評価中、ゴールド シーケンスのステップ数の1.5 倍以内にモデルが回答を提出しなかった場合、トライアルは中止され、不合格と見なされました。ADAM オプティマイザー(Kingma and Ba, 2014)を $5e-5$ の固定学習率で使用し、ミニバッチ サイズが 128 になるように勾配を累積しました。

結果。表2に PTA タスクの結果を示します。私たちのモデルは、カーソル、ボタン、およびテキストのタスクでうまく機能し、そのような単純なアクションで言語モデルを微調整できることを示しています。モデル スコアは、エリアのチャンス レベルを上回っています

このタスクでは、モデルが正しいボタンを選択するための指示を覚えている必要があります。これは、私たちの記憶メカニズムの有効性を示唆しています。BUT BERTmediumは、パラメータ サイズが大きいかかわらず、ボタンとエリアのタスクでBUI-BERTsmall よりもパフォーマンスが低下します。その理由はまだ明らかではありません。検討中

¹⁰結果は 1 回の実行に基づいていることに注意してください。

11 から事前に訓練された BERT で初期化

<https://github.com/google-research/bert>

どちらもテキスト タスクで同じレベルのスコアを獲得したため、可能な解釈は、メディアモデルがサブワードの位置の埋め込みを十分に学習せず、ビューでサブワードを誤って配置した可能性があることです。

6.2 主なタスク

CoLA (Warstadt et al., 2019)、STS-B (Cer et al., 2017)、MNLI-matched (Williams et al., 2018)、SQuADv2、VQAv2、および私たちのSA は共同でタスクを実行します。トレーニング例の数は、それぞれ 8.6 k、5.7 k、393 k、130 k、444 k、および 50 k でした。最初の3 つのタスクは、GLUE ベンチマークからのものです。

比較したモデル。表3にまとめを示します。BERTsmall/+V:パフォーマンスの上限を推定するために、タスク固有のヘッドを使用して、SA を除く各タスクに対して個別に BERTsmallを微調整しました。BERTsmall-s2s+V:テキスト生成モデルとの比較のために、BERTsmallの重みに基づいてエンコーダとデコーダの重みを初期化したエンコーダ/デコーダ モデルを用意しました。SA を除くすべてのタスクでそのモデルを共同でトレーニングしました。入力シーケンスは、テンプレート(付録B.1)を使用して、タスクの説明、質問、およびクラス ラベルなど、タスクを解決するために必要な最も完全な情報を提供するように生成されました。サフィックスが +V のモデルは、VQA 用のイメージ入力を使用します。BUI-BERTと同様の方法で、ResNet18を使用してグリッド機能を取得しました。機能を頭に挿入しました

入力埋め込みの。付録B.2に詳細が記載されています。

設定。すべてのモデルを 10 エポックでトレーニングしました。各エポックの終わりに検証損失を追跡して、最適なモデルを選択しました。その他の条件は BUI-BERTはPTAのものと同じでした

トレーニング。比較モデルのハイパーパラメーターを最適化しました(付録B.3を参照)。

結果。表4に結果をまとめます。非アクションまたはシングルページ タスクの場合、BERTsmall (シングルタスク設定) と BERTsmall-s2s+V (マルチタスク設定) のどちらか高いパフォーマンスを採用することで、事前トレーニング済みの BERTsmallパフォーマンスの上限を推定します。ご覧のとおり、BUI-BERTsmallは、タスク フレームワークの CoLA を除いて、その見込みの約 76% ~ 90% を達成しています。これは、事前に訓練された言語モデルが、完全ではありませんが、行動シーケンスで言語知識を使用することを学習できることを示唆しています。マルチタスク設定で CoLA スコアが小さいのは、

モデル	ベース LM #params	エグゼクティブ	建築
BERT小/+V	31M	42M	タスク仕様。 head BERT (Devlin et al., 2019)
BERTsmall-s2s+V	BERTsmall 74M		PrからのEnc-dec. LM (Rothe et al., 2020) text gen.
BUI-BERT小	42M	BUI アクション	BUI-BERT BUI アクシヨ
BUI-BERT中	BERT中54M		ン
			バイパート

表 3: 比較するモデル。 +V を使用するモデルは、凍結された事前トレーニング済みの ResNet18 から取得した画像入力を使用します。 #params のうち、ResNet18 とその関連レイヤーは約 11M を占めています。

モデル	マルチタスク	マルチステップ	非アクションまたはシングルページ				マルチページ SA ACC。
			CoLA M	STS-B	MNLI	QQP	
BERTsmall/+V			31.3	75.6	71.8	71.3	56.8 47.0
BERTsmall-s2s+V BUI-	いい		0.0		42.9/51.4		43.6 49.4
BERTsmall BUI-	SAなしすべ		-1.0		51.4		-
BERTsmall BUI-	すべて		-1.8		48.0 48.8		75.5
BERTmed.							69.7

表 4: 検証分割の全体的なスコア。 M と P は、それぞれマッシュアップとピアソンの相関を表します。

比較的小さなトレーニング データ。モデルのパフォーマンスとサイズ (BUI BERTsmall/medium) の関係において、すべてのタスクに当てはまる傾向は見られませんでした。中間モデルは、言語モデルのサイズと一致する単一ページのタスクでより優れたパフォーマンスを発揮します。ただし、複数ページのタスクではパフォーマンスが逆になります。パフォーマンスがタスクのモデル サイズに比例するかどうかを確認するには、さらに調査する必要があります。

6.3 分析

アブレーション研究。PTA と記憶メカニズムを検証するために 2 つのモデルを追加しました。BUI BERTsmall w/o PTA の場合、重みを BERTsmall で初期化し、マルチタスク トレーニングで直接トレーニングしました。BUI-BERTsmall w/o mem の場合、メモリ シーケンスを省略しました。このモデルは、PTA で再初期化およびトレーニングされました。表 5 に示す

結果。アブレーションされたモデルは、ほとんどすべてのタスクで BUI BERTsmall よりも低かった。これは、PTA と記憶メカニズムが有効であることを示しています。特に、SA タスクで BUI BERTsmall w/o mem の大幅な低下は、メモリが重要な役割を果たしていることを示唆しています。

SA タスク。表 5 に SA タスクの結果の詳細を示します。ご覧のとおり、BUI-BERTsmall w/o mem を除いて、ほとんどの場合、すべてのモデルが回答の送信に成功しました。ただし、SA-A タスクのスコアは低いです。この低いスコアは、タスクの違いを反映しています。SA-QID、-Q、および -H タスクは、表示された結果をコピーすることで解決できますが、SA-A タスクでは、質問に対する答えを推測する必要があります。このギャップを埋めることは、将来の課題です。SA-QID、-Q、および -H のスコアも難易度を反映していることに注意してください。答え (ヒット カウント) が常に

最初の行に表示されます。SA-Q は、ヒット カウントが常に一定であるため、比較的簡単です。SA-QID は、ヒット カウントが複数あるため、より難しく、モデルはスクロールして答えを見つける必要がある場合があります。SA-Q および -QID タスクに関しては、まだ改善の余地があります。

目に見えないタスク。モデルが命令をどのように理解したかを確認するために、目に見えないタスクでモデルを評価しました。

WNLI (Levesque et al., 2012)、MRPC (Dolan and Brockett, 2005)、SST-2 (Socher et al., 2013) の 3 つの GLUE タスクを使用しました。これらは 2 択のタスクであり、学習したタスクとの類似性は異なっていました。WNLI と MNLI はテキストでした

含意タスク。MRPC と STS-B は、同等性と類似性のタスクでした。SST-2 は、モデルにとって新しい感情予測です。提出数 (#sub) と正しい提出数 (#cor) を表 6 に報告します。これらのモデルの提出率は、画面上のボタンが占める領域のパーセンテージ。提出された回答の正しい答えの割合に基づいて、モデルが自信がないときに回答を提出することを控えているようには見えません。この低いサブミッション率は、モデルが手順のアクションを指示の単語に合わせることなく、手順全体を記憶する可能性が高いことを示唆しています。

7 ディスカッション

NLP が扱うことができる世界を拡大することは、私たちの長期的な目標の 1 つです (Bisk et al., 2020)。GUI を操作するモデルは、動的相互作用に関する研究の統一された基礎を提供します。たとえば、動的グラウンディング (Chandu et al., 2021)。

	CoLA STS-B MNLI-m VQAv2 SQuADv2					検索と回答 (SA)							
	#steps	#cases	2.0	5.8	9.2	-QID-Q-H-A-all	23.0	30.8	13.1	19.4	24.0	1623	
	71.6	1040	1500	1500-1.0	マクロ f1	ACC _c	ACC _c	ACC _c	ACC _c	ACC _c	1634	182	
	1500	1500	1500	1500-1.0	ACC _c	ACC _c	ACC _c	ACC _c	ACC _c	ACC _c	1561	5000	
BUI-BERTsml.	1500	0.0	51.1	1500	78.3	9815	213236	9563	1592	1615	182	1514	4903
	スコア					70.4	48.0	43.6	88.8	96.0	100	37.2	75.5
PTAなし	#サブ					9813	213122	9011	1559	1612	182	1507	4860
	スコア					68.1	45.9	38.9	60.0	94.8	100	36.1	65.4
メモリなし	#サブ					9815	212334	7669	747	943	182	1122	2994
	スコア					69.0	46.9	27.9	25.3	23.7	100	9.7	22.7
BUI-BERTmed.	#sub	1040	9815	213599	10082	1591	1614	182	1591	1614	182	1506	4893
	スコア	-1.8	75.6	48.8	49.4	93.7	96.0	100	93.7	96.0	100	13.8	69.7

表 5: 検証分割による切除研究。#steps: ゴールド シーケンスの平均ステップ数。#cases: 評価されたケースの数。#sub: モデルがサブミッションを行ったケースの数。提出がない場合は不合格としてカウントした。M と P は、それぞれマッシュアップとピアソンの相関を表します。

(#cor / #sub)	WNLI MRPC SST-2 408 0 / 0 17 / 33
#cases	71 181 / 227 120 / 2812 /
BERTsmall-s2s+V 8 / 14	29 29 / 53 1 / 2
BUI-BERT小	
BUI-BERT中4 / 9	

表 6: 検証分割に関する目に見えないタスク評価。#sub (#cor): モデルが回答の提出に成功したケースの数 (正解)。#cases: 評価されたケースの数。

els は、GUI を介してさまざまなソフトウェアに適応できます。このホワイト ペーパーでは、既存のベンチマーク タスクを GUI タスクに変換しました。このような GUI タスクは、言語モデルのパフォーマンスが GUI モデルにどの程度うまく移行されているかを評価するのに役立ちます。この評価の視点は、MiniWoB (Shi et al., 2017) および MiniWoB++ (Liu et al., 2018) を補完します。

私たちの実験では、言語モデルが言語知識をある程度保持しながら GUI 操作を学習できるという有望な結果が示されましたが、私たちの研究には次の制限があります。• 計算コストが高いため、より大きな LM を使用したスケールの検証など、詳細な実験ができません。効率的な変圧器 (Tay et al., 2020) や Perceiver (Jaegle et al., 2021) などの軽量モデルを探する必要があります。• 目に見えないタスクのパフォーマンスが低い。データ拡張 (つまり、ページ レイアウトの変更) によってパフォーマンスが向上すると予想されますが、探索には強化学習 (Sutton and Barto, 2018) が不可欠になる可能性があります。• 私たちのモデルは、ブラウザに依存するエミュレートされた OCR を使用します。モデルをブラウザ GUI 以外の GUI に適用するには、その機能を変更することが今後の課題です。実際の OCR を使用することは簡単な解決策ですが、その処理時間は私たちの目的にとって無視できません。したがって、ピクセルから直接読み取る LM を作成する場合があります。

8 結論

この作業では、ブラウザー UI を使用するために複数のアクションを必要とするタスク フレームワークに BERT を適用できることを示しました。マルチタスク トレーニングでは、メモリ メカニズムを備えた BERT 拡張機能が、タスク ページによって提供されるハイパーリンクを含む UI に従って 6 つのタスクを解決することを学習しました。同時に、目に見えないタスクを解決する能力の低さも観察されました。現在のモデルの計算コストは詳細な実験を制限するため、次のステップとして既存の効率的な方法で軽量モデルを作成する予定です。

謝辞

この作品は JSPS 科研費の支援を受けました 助成金番号 21H03502 および JST SPRING 助成金番号 JPMJSP2104。

参考文献

ピーター・アンダーソン、チャー・ウー、ダミアン・テニー、ジェイク・ブルース、マーク・ジョンソン、ニコ・サンダーハウフ、イアン・リード、ステューブ・グールド、アントン・ヴァン・デン・ヘンゲル。2018。視覚と言語のナビゲーション: 実環境における視覚的に接地されたナビゲーション指示の解釈。IEEE Conference on Computer Vision and Pattern Recognition の議事録、3674 ~ 3683 ページ。

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh. 2015. VQA: ビジュアル クエスチョン 回答。コンピュータ ビジョンに関する IEEE 国際会議の議事録、2425 ~ 2433 ページ。

Chongyang Bai, Xiaoxue Zang, Ying Xu, Srinivas Sunkara, Abhinav Rastogi, Jindong Chen, Blaise Agüera y Arcas. 2021. UIbert: UI アンダースタンディングの汎用マルチモーダル表現の学習。第 30 回インターナショナル会議

- 人工知能に関する国際合同会議、1705 ~ 1712 ページ。
- IEEE Conference on Computer Vision and Pattern Recognition, 1 ~ 10 ページ。
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, Joseph Turian. 2020年。
[グラウンド言語を体験してください](#)。自然言語処理における経験的手法に関する 2020 年会議の議事録, 8718 ~ 8735 ページ。
- ジェイコブ・デブリン、ミンウェイ・チャン、ケントン・リー、クリスティーナ・トウタノフ。 2019. [BERT: 言語理解のためのディープ双方向トランスフォーマーの事前トレーニング](#)。計算言語学協会の北米支部の 2019 年会議の議事録: 人間の言語技術、第 1 巻、4171 ~ 4186 ページ。
- トム・B・ブラウン、ベンジャミン・マン、ニック・ライダー、メラニー・サブピア、ジャレッド・カプラン、プラフラ・ダリワル、アービンド・ニラカント、プラナフ・シャム、ギリッシュ・サストリー、アマンダ・アスケル、サンディニ・アガルワル、アリエル・ハーバート=ヴォス、グレッチェン・クルーガー、TJ・ヘニガン、リウオン・チャイルド、アディティア Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam Mc Candlish, Alec Radford, Ilya Sutskever, Dario Amodei. 2020. [言語モデルは少数の学習者です](#)。arXiv プレプリント arXiv:2005.14165.
- ウィリアム・B・ドーランとクリス・ブロケット。 2005. [センテンスパラフレーズのコーパスの自動構築](#)。言い換えに関する第 3 回国際ワークショップの議事録。
- Izzeddin Gur, Natasha Jaques, Yingjie Miao, Jongwook Choi, Manoj Tiwari, Honglak Lee, Alexandra Faust. 2021. [ゼロショット構成強化学習のための環境生成](#)。神経情報処理システムの進歩。
- Izzeddin Gur, Ulrich Rueckert, Aleksandra Faust, Dilek Hakkani-Tur. 2018. [Webの操作方法を学ぶ](#)。学習代表者の憤りに関する国際会議。
- リッチ・カルアナ。 1997. [マルチタスク学習](#)。マシン学習, 28: 41 ~ 75 ページ。
- ダニエル・セル、モナ・ディアブ、エネコ・アギレ、イニゴ・ロペス・ガスピオ、ルシア・スベシア。 2017. [SemEval-2017タスク 1: セマンティック テキストの類似性 多言語およびクロスリンガルに焦点を当てた評価](#)。意味評価に関する第 11 回国際ワークショップの議事録, 1 ~ 14 ページ。
- 開明何、シャンコ・チャン、シャオチン・レン、ジャン・サン。 2016. [画像認識のための深層残差学習](#)。コンピューター ビジョンとパターン認識に関する IEEE 会議の議事録, 770 ~ 778 ページ。
- キャティ・ラガヴィ・チャンドウ、ヨナタン・ビスク、アラン・W・ブラック。 2021. [グラウンディング NLP の「グラウンディング」](#)。計算言語学協会の調査結果: ACL-IJCNLP 2021, 4283-4305 ページ。
- Zecheng He, Srinivas Sunkara, Xiaoxue Zang, Ying Xu, Lijuan Liu, Nevan Wichers, Gabriel Schubiner, Ruby Lee, Jindong Chen. 2021. [ActionBert: ユーザー インターフェイスのセマンティックな理解のためのユーザー アクションの活用](#)。人工知能に関する AAAI 会議の議事録, 第 35 巻, 5931 ~ 5938 ページ。
- Xingyu Chen, Zhao Zhao, Lu Chen, JiaBao Ji, Danyang Zhang, Ao Luo, Yuxuan Xiong, Kai Yu. 2021. [WebSRC: Web ベースの構造的読解のためのデータセット](#)。自然言語処理における経験的方法に関する 2021 年会議の議事録, 4173 ~ 4185 ページ。
- Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez Opazo, Stephen Gould. 2021. [VLN BERT: ナビゲーションのための反復的な視覚と言語の BERT](#)。IEEE/CVF Conference on Computer Vision and Pattern Recognition の議事録, 1643 ~ 1653 ページ。
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, および Jingjing Liu. 2020. [UNITER: ユニバーサルイメージ テキスト表現の学習](#)。The 2020 European Conference on Computer Vision, 104 ~ 120 ページ。
- Ronghang Hu と Amanpreet Singh. 2021. [UnitT: 統合トランスフォーマーを使用したマルチモーダル マルチタスク学習](#)。IEEE/CVF International Conference on Computer Vision (ICCV) の議事録, 1439 ~ 1449 ページ。
- Jaemin Cho, Jie Lei, Hao Tan, Mohit Bansal. 2021. [テキスト生成による視覚と言語のタスクの統合](#)。第 38 回国際機械学習会議の議事録, 第 139 巻, 1931 ~ 1942 ページ。
- Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, Jianlong Fu. 2020. [Pixel-BERT: Deep Multi-Modal Transformers による画像ピクセルとテキストの位置合わせ](#)。arXiv プレプリント arXiv:2004.00849.
- Lei Cui, Yiheng Xu, Tengchao Lv, Furu Wei. 2021年。
[Document AI: ベンチマーク、モデル、およびアプリケーション](#)。arXiv プレプリント arXiv:2111.08609.
- ピーター・C・ハンフリーズ、デビッド・ラボソ、トビアス・ポーレン、グレゴリー・ソートン、ラチータ・チャバリア、アリスター・ムルダル、ジョシュ・エイブラムソン、ペトコ・ゲオルギエフ、アレックス・ゴールディン、アダム・サントロ、ティモシー・P・リリクラップ。 2022. [コンピューターの制御を学習するためのデータ駆動型アプローチ](#)。arXiv プレプリント arXiv:2202.08137.
- アビシェク・ダス、サムヤク・ダッタ、ジョージア・ギキオサリ、ステファン・リー、デヴィ・パトリク、ドゥルヴ・バトラ。 2018. [Embodied Question Answering](#)。議事録において

- アンドリュー・イーグル、フェリックス・ギメノ、アンディ・ブロック、オリオール・ベニャルス、アンドリュー・ジサーマン、ジョアン・カレイラ。
2021. 知覚者: 反復的な注意を伴う一般的な知覚。第 38 回機械学習に関する国際会議の議事録, Proceedings of Machine Learning Research の第 139 巻, 4651 ~ 4664 ページ。
- Sheng Jia, Jamie Ryan Kiros, Jimmy Ba. 2019年。
DOM-q-NET: 構造化言語に基づいた強化学習。
学習表現に関する国際会議。
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, Tamara Berg. 2014. Referitgame: 自然の風景の写真の中のオブジェクトを参照する。自然言語処理における経験的方法に関する 2014 年の会議の議事録 (EMNLP), 787 ~ 798 ページ。
- ディーデリック・P・キングマとジミー・バ。 2014. Adam: 確率的最適化の方法。 arXiv プレプリント arXiv:1412.6980.
- ヘクター・レベスク、アーネスト・デイビス、レオラ・モーゲン・スターン。 2012. ウィノグラード スキーマ チャレンジ。知識表現と推論の原理に関する第 13 回国際会議。
- Junlong Li, Yiheng Xu, Lei Cui, Furu Wei. 2021a.
MarkupLM: 視覚的に豊かなドキュメント理解のためのテキストとマークアップ言語の事前トレーニング。 arXiv プレプリント arXiv:2110.08518.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, Kai-Wei Chang. 2019a. VisualBERT: 視覚と言語のためのシンプルで高性能なベースライン。 arXiv プレプリント arXiv:1908.03557.
- Xiujun Li, Chunyuan Li, Qiaolin Xia, Yonatan Bisk, Asli Celikyilmaz, Jianfeng Gao, Noah A Smith, Yejin Choi. 2019b. 言語の事前トレーニングと確率的サンプリングによる堅牢なナビゲーション。自然言語処理における経験的手法に関する 2019 年会議の議事録および自然言語処理に関する第 9 回国際合同会議 (EMNLP-IJCNLP), 1494 ~ 1499 ページ。
- ヤン・リー、ガン・リー、シン・チョウ、モスタファ・デガーニ、アレクセイ・グリツェンコ。 2021b. VUT: マルチモーダル マルチタスク ユーザー インターフェイスモデリング用の汎用 UI トランスフォーマー。 arXiv プレプリント arXiv:2112.05692.
- Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tian lin Shi, Percy Liang. 2018. ワークフローに基づく探索を使用した Web インターフェイスでの強化学習。学習表現に関する国際会議。
- Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee. 2019.
ViLBERT: 視覚および言語タスクのためのタスクに依存しない視覚言語表現の事前トレーニング。神経情報処理システムの進歩, 第 32 巻。
- Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, Stefan Lee. 2020. 12-in-1: マルチタスクビジョンと言語表現の学習。コンピューター ビジョンとパターン認識に関する IEEE/CVF 会議の議事録, 10437 ~ 10446 ページ。
- Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, Dhruv Batra. 2020年。
Web からの画像とテキストのペアによる視覚と言語のナビゲーションの改善。欧州コンピュータ ビジョン会議, 259 ~ 274 ページ。
- Stephen Merity, Caiming Xiong, James Bradbury, Richard Socher. 2016. ポインター センチネル混合モデル。 arXiv プレプリント arXiv:1609.07843.
- Yuankai Qi, Zizheng Pan, Yicong Hong, Ming Hsuan Yang, Anton van den Hengel, および Qi Wu. 2021. 場所を知る道: 屋内視覚言語ナビゲーションのためのオブジェクトと部屋に情報を提供するシーケンシャル BERT。 IEEE/CVF International Conference on Computer Vision (ICCV) の議事録, 1655 ~ 1664 ページ。
- アレック・ラドフォード、カーシク・ナラシンハン、ティム・サリマンズ、イリヤ・サツケバー。 2018. ジェネレーティブプレトレーニングによる言語理解の向上。 OpenAI ブログ。
- アレック・ラドフォード、ジェフ・ウー、リウォン・チャイルド、デビッド・ルアン、ダリオ・アモデイ、イリヤ・サツケバー。 2019. 言語モデルは教師なしマルチタスク学習者です。 nAI ブログ, 1(8) を開きます。
- コリン・ラッフェル、ノーム・シェイザー、アダム・ロバーツ、キャサリン・リー、シャラン・ナラン、マイケル・マテナ、ヤンキ・チョウ、ウェイ・リー、ピーター・J・リュウ。 2020. 統合されたテキストからテキストへのトランスフォーマーで転移学習の限界を探る。 Journal of Machine Learning Research, 21(140): 1 ~ 67 ページ。
- Pranav Rajpurkar, Robin Jia, Percy Liang. 2018年。
あなたが知らないことを知ってください: SQuAD に対する答えのない質問。第 56 回計算言語学協会年次総会の議事録第 2 巻, 784 ~ 789 ページ。
- プラナヴ・ラージプルカール、ジャン・ジャン、コンスタンチン・ロピレフ、パーシー・リャン。 2016. SQuAD: テキストの機械理解に関する 100,000 以上の質問。自然言語処理における経験的手法に関する 2016 年会議の議事録, 2383 ~ 2392 ページ。
- サシャ・ローテ、シャシ・ナラヤン、アリアクセイ・セヴェリン。 2020. シーケンス生成タスクのための事前トレーニング済みチェックポイントの活用。計算言語学協会のトランザクション, 8: 264 ~ 280 ページ。
- セバスチャン・ルーダー。 2017. ディープ ニューラル ネットワークにおけるマルチタスク学習の概要。 arXiv プレプリント arXiv:1706.05098.
- Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, および Percy Liang. 2017. World of Bits: アン

- Web ベースのエージェント向けのオープン ドメイン プラットフォーム。機械学習に関する国際会議、3135 ~ 3144 ページ。
- Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, Dieter Fox. 2020. ALFRED: 日常業務の根拠のある指示を解釈するためのベンチマーク。コンピューター ビジョンとパターン認識に関する IEEE/CVF 会議の議事録、10740 ~ 10749 ページ。
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, および Christopher Potts. 2013. センチメント ツリーバンク上のセマンティック コンポジションの再帰的ディープ モデル。自然言語処理における経験的手法に関する 2013 年会議の議事録、1631 ~ 1642 ページ。
- Shashank Srivastava, Oleksandr Polozov, Nebojsa Jojic, Christopher Meek. 2020. コンテキスト内の説明とデモンストラーションを推論することにより、Web ベースの手順を学習します。計算言語学協会の第 58 回年次総会の議事録、7652 ~ 7662 ページ、オンライン。計算言語学の学会として。
- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, Jifeng Dai. 2020. VL-BERT: Generic Visual-Linguistic Representations の事前トレーニング。学習代表者の憤りに関する国際会議。
- リチャード・S・サットンとアンドリュー・G・バルト。2018. 強化学習: はじめに。MITプレス。
- ハオ・タンとモヒト・バンサル。2019. LXMERT : トランスフォーマーからのクロスモダリティ エンコーダー表現の学習。自然言語処理における経験的方法に関する 2019 年会議の議事録および自然言語処理に関する第 9 回国際合同会議、ページ 5100-5111。
- 田中良太、西田恭介、吉田千。
2021. VisualMRC: 文書画像の機械読解。人工知能に関する AAAI 会議の議事録、第 35 巻、13878 ~ 13888 ページ。
- イ・テイ、モスタファ・デガニ、ダラ・バーリ、ドナルド・メッツラー。
2020. 効率的な変圧器: 調査。 arXiv プレプリント arXiv:2009.06732.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin. 2017年。必要なのは注意だけです。神経情報処理システムの進歩、5998 ~ 6008 ページ。
- アレックス・ワン、アマンブリート・シン、ジュリアン・マイケル、フェリックス・ヒル、オマー・レヴィ、サミュエル・R・ボウマン。2019年。
GLUE: 自然言語理解のためのマルチタスク ベンチマークおよび分析プラットフォーム。学習表現に関する国際会議。
- Alex Warstadt, Amanpreet Singh, Samuel R. Bowman. 2019. ニューラル ネットワークの受容性判断。計算言語学協会のトランザクション、7: 625-641 ページ。
- アディナ・ウィリアムズ、ニキータ・ナンギア、サミュエル・ボウマン。
2018. 推論による文理解のための広域チャレンジ コーパス。計算言語学協会の北米支部の 2018 年会議の議事録: 人間の言語技術、第 1 巻、1112-1122 ページ。
- Te-Lin Wu, Cheng Li, Mingyang Zhang, Tao Chen, Spurth Amba Hombaiyah, Michael Bender sky. 2021. LAMPRET: ドキュメント理解のためのレイアウト対応マルチモーダル事前トレーニング。 arXiv プレプリント arXiv:2104.08405.

環境の詳細

ブラウザ。タスク ページのレンダリングと実行には、次の環境を使用しました。

- OS: Ubuntu 18.04.20.04
- ブラウザ: Firefox バージョン 87.0
- ブラウザドライバ: geckodriver 0.29.0
- セレン: バージョン 3.141.0
- デフォルトのフォント (本文): DejaVu Serif, 16px • デフォルトのフォント (テキスト、ボタン): 13px

パッケージとライブラリ。 Python 3.6 と PyTorch (1.10) を使用して BUI モデルを実装しました。シーケンスからシーケンスへのモデルには、 Transformers ライブラリ (4.12) を使用しました。 SQuAD と VQA を評価するために、公開スクリプトを使用しました¹²。

トレーニング。 32 GB VRAMを搭載した NVIDIA V100 GPUまたは NVIDIA RTX 3090 GPU を使用しました。

各トレーニング実行で 24 GM VRAM。

A.1 所要時間

実験に必要な時間を表7 に示します。

B 比較に関する追加の詳細 モデル

B.1 入力テンプレート

テンプレートを使用して、 seq2seq モデルのテキスト入力を作成しました。表8は、テンプレートと、各データセットのテンプレート内の文を埋める規則を示しています。 task_type、instruction、コンテンツ名、およびコンテンツ値を提供するようにテンプレートを作成しました。さらに、 VQA のシーケンスの先頭に画像の埋め込みが追加されます。データセットの内容を参考にTASK_TYPEとINSTRUCTIONを作成しました。

B.2 画像入力

まず、与えられた画像を 320px × 224px にリサイズします。縦横比が一致しない場合は、画像を中央に配置し、不足しているピクセルを黒いピクセルで埋めます。次に、凍結された事前トレーニング済みのResNet18モデルに画像を入力して、最後の特徴マップ (C4; 10x7) を取得します。特徴マップを1次元で平坦化し、各特徴を1つの完全に接続された線形レイヤー (トレーニング可能) に入力して、次元を LM の隠れた次元に合わせます。最後に、私たちは

位置およびセグメントタイプの埋め込みを追加する前に、これらの機能と言語の埋め込みを連結します。

B.3 ハイパーパラメータ

学習率 (LR) のスケジューリングなしで ADAM オプティマイザーを使用し、自動混合精度 (AMP) を有効にしました。すべてのトレーニングは 10 エポックでした。予備実験では、事前トレーニング済みの ResNet18 の最適化がVQA のパフォーマンスにほとんど影響しないことがわかったため、上記の固定設定のみを使用しました。

BERT小/ +V. GLUE タスクと VQA 300、および SQuAD 512の最大トークン長を修正しました。6つのハイパーパラメータの組み合わせを試しました: {64, 128, 256} からのミニバッチ サイズ、{1e-4, 5e-4} からの LR -5}。検証損失が最小となるハイパーパラメータセットを採用した。

BERTsmall-s2s+V.ベースのミニバッチサイズを 32 に、テキストのみのタスクの最大トークン長を512 に、テキストと画像のタスクの最大トークン長を 432 に修正しました (+70 の画像埋め込み)。 {1, 4} からの勾配累積と{1e-4, 5e-5, 1e-5} からの LRという 6 つのハイパーパラメータの組み合わせを試しました。検証損失が最小となるハイパーパラメータセットを採用した。 BERTsmall-s2s+Vの最適なハイパーパラメータは (4, 5e-5) でした

B.4 Seq2Seqモデルによる分類

分類タスクでは、生成されたテキストが指示で指定されたどのクラス ラベルとも完全に一致しなかった場合、モデルが回答を送信できなかったと見なされました。

C BUI セットアップのタスク

C.1 回答フォームの説明

ここでは、指示と回答フォームをイメージで示します。図5は、SQuAD および VQAページを示しています。図6は、GLUE タスクのタスク ページを示しています。図7と図8は、SA タスクのタスク ページを示しています。命令は、表8に示されている seq2seq モデルのカウンター パーツと基本的に同じですが、画面に収まるように単語の選択が変更されています。

C.2 アクション語彙の事前トレーニング用テンプレート

Wikitext103 (Merity et al., 2016)コーパスのトレーニング分割から語彙を作成しました。からなる言葉を守りました。


¹²分隊 : <https://rajpurkar.github.io/SQuAD-explorer/> および VQA : <https://github.com/GT-Vision-Lab/VQA>。

例のレコード ゴールド シーケンスを処理します。他の ~4d ~1.3M の例のゴールド シーケンスを記録します。

50 ep の PTA で小規模/中規模のトレーニングを行います。~2d / ~4d 60k の例。GPUで。		
10 ep のマルチタスク トレーニングで小規模/中規模のトレーニングを行います。~6d / ~12d ~1.0M の例。GPUで。		
val のモデルで予測します。val 上のモデルを使用し	20分~2	2k の例。GPUで。~230k の
た PTA Predict の分割。他の分割	日	例。GPUで。

表 7: 所要時間。異なる構成の複数のサーバーを使用したため、これらの値は概算です。モデルの画面サイズ、トークン化、およびアクションが同一である場合、ゴールド シーケンスは再利用可能です。1つの例のスクリーンショットとアクションを1つの json ファイルに保存し、使用するたびにディスクから読み取りました。トレーニングには float32 を使用しました。また、自動混合精度で float16 も試しました。トレーニング時間は約 30% 短縮されましたが (削減されたメモリ スペースを使用してバッチサイズを 2 倍にしました)、NaN が発生してトレーニングが停止することがありました。そのため、今回は使用しませんでした。

Visual Question Answering Task
See the picture below and answer the following question.



Question: How many sets of doors are open?

Answer:

Reading Comprehension Task
Read the next paragraph and answer the following question. Enter your answer in the textbox or check the unanswerable box when you think the question is unanswerable.

Paragraph: The Norman dynasty had a major political, cultural and military impact on medieval Europe and even the Near East. The Normans were famed for their martial spirit and eventually for their Christian piety, becoming exponents of the Catholic orthodoxy into which they assimilated. They adopted the Gallo-Romance language of the Frankish land they settled, their dialect becoming known as Norman, Normand or Norman French, an important literary language. The Duchy of Normandy, which they formed by treaty with the French crown, was a great fief of medieval France, and under Richard I of Normandy was forged into a cohesive and formidable principality in feudal tenure. The Normans are noted both for their culture, such as their unique Romanesque architecture and musical traditions, and for their significant military accomplishments and innovations. Norman adventurers founded the Kingdom of Sicily under Roger II after conquering southern Italy on the Saracens and Byzantines, and an expedition on behalf of their duke, William the Conqueror, led to the **Norman conquest of England at the Battle of Hastings in 1066.** Norman cultural and military influence spread from these new European centres to the Crusader states of the Near East, where their prince Bohemond I founded the Principality of Antioch in the Levant, to Scotland and Wales in Great Britain, to Ireland, and to the coasts of north Africa and the Canary Islands.

Question: Who ruled the duchy of Normandy?

Your answer: unanswerable

図 5: VQAv2 の BUI バージョン (左) と SQuADv2 の BUI バージョン (右) の画面例。これらは、BUI モデルが受け取るスクリーンショットです。青い破線の長方形は、モデルの最初の可視領域を示しています。指示と回答フォームは、すべての例で共通です。

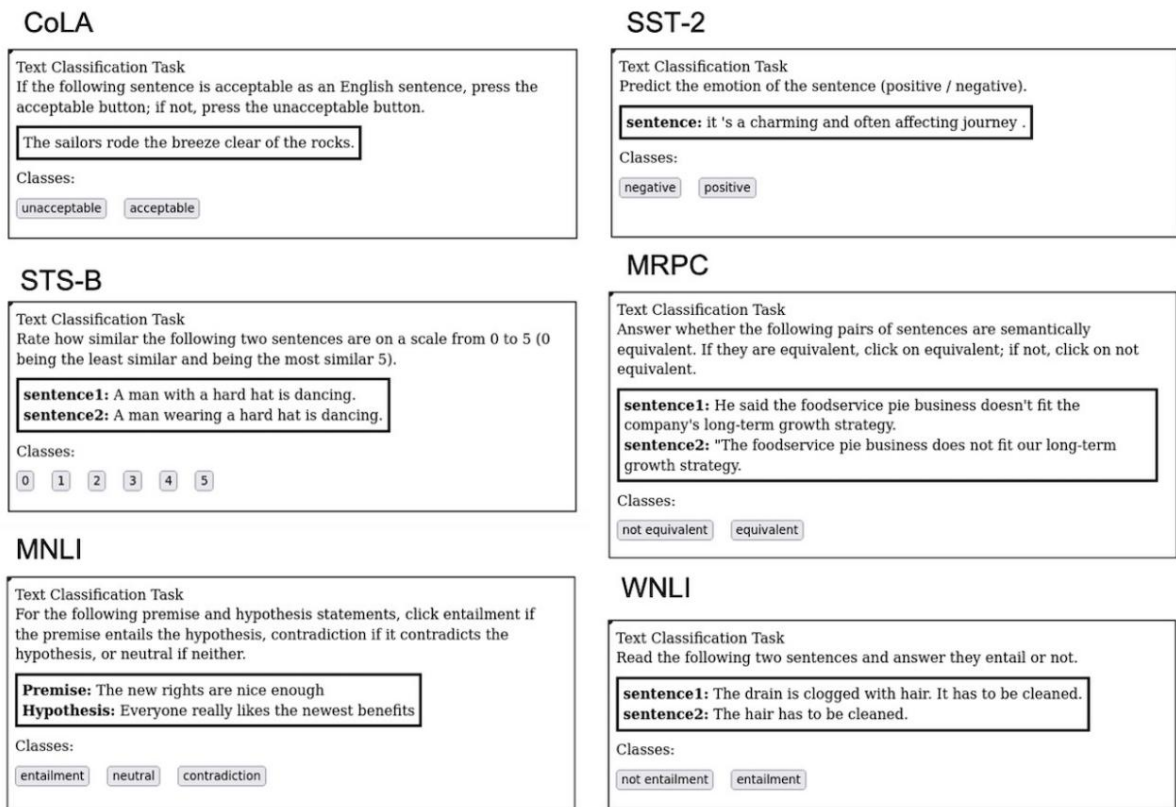


図 6: GLUE ベンチマークの BUI バージョンの画面例。下マージンは省略されます。太字の塗りつぶされたボックス内の内容は例によって異なりますが、命令とラベル ボタンは共通です。使用しなかったタスク (QNLI、QQP、および RTE) は示されていないことに注意してください。

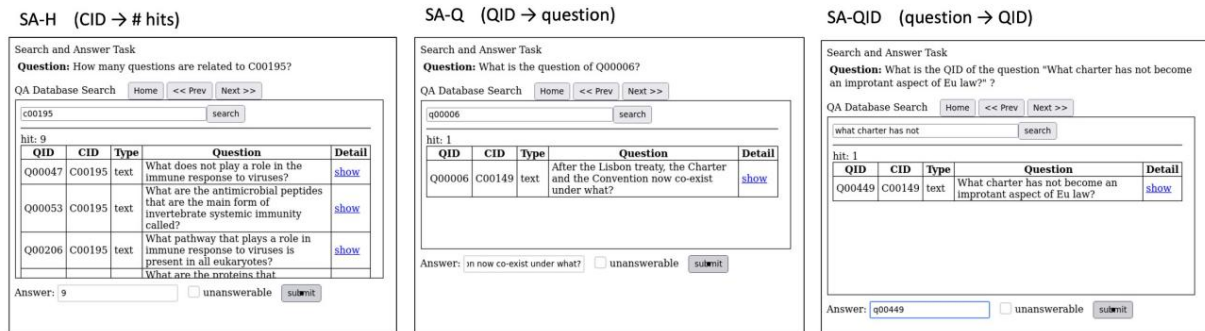


図 7: SA-QID、-Q、および -H の画面例。スクリーンショットは、タスクの最後のステップを示しています。これらのタスクは、(1) 与えられた命令からキー フレーズを抽出する、(2) キー フレーズをクエリする、(3) 回答セグメントを見つける、(4) セグメントに入る、によって解決されることが期待されます。

SA-A (QID → question → answer)

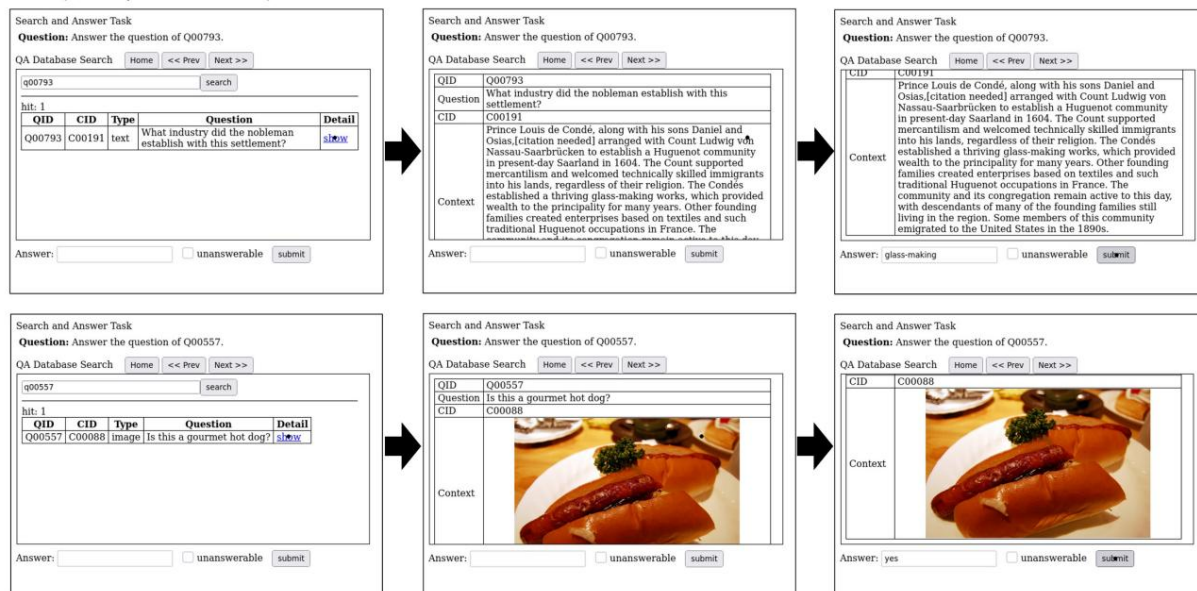


図 8: SA-A の画面例。これらのタスクは、(1) 与えられた命令からキー フレーズを抽出する、(2) キー フレーズをクエリする、(3) 詳細を表示する、(4) 問題を読む、(5) 中から答えを見つける、によって解決されることが期待されます。コンテキスト、および (6) 回答を入力します。

アルファベットと数字のみ。単語を小文字にしました。

バリエーションを作成するために、命令内の単語のセットが展開されます。タスク インスタンスの命令から1 つを一樣にサンプリングしました。

C.2.1 カーソル

手順：

- ボックス内でカーソルを移動します。
- カーソルでボックスをポイントします。

ボックスの座標は、ウィンドウから一樣にサンプリングされました。

C.2.2 ボタン

手順：

- {Click, Push, Press, Choose, Select} WORD というラベルの付いたボタン。
- {クリック、プッシュ、プレス、選択、選択} ワードボタン。

WORDは語彙からサンプリングされました。

C.2.3 エリア

手順：

- ボタンが表示されるまで下にスクロールし、WORD というラベルの付いたボタンをクリックします。
- ボタンが表示されるまで下にスクロールし、WORD ボタンをクリックします。

WORDは語彙からサンプリングされました。

C.2.4 テキスト

手順：

- {Type, Enter, Input}各テキスト ボックスの左側の文字列。で送信ボタンをクリックします
最後。

各ストリングは、語彙からサンプリングされた 2 つの単語をスペースで結合して作成されました。

C.3 検索および回答タスクの詳細

検索と回答のタスクでは、SQuADと VQA のそれぞれから 100 のコンテキスト (段落または画像) をサンプリングしてデータベースを作成しました。各コンテキストには約 10 個の質問があるため、データベースには約 2,000 個の質問が含まれます。このデータベース サイズを選択したのは、データ全体をモデルに入力するのが難しくするためです。データベース、CID、および QID 内の各コンテキストと質問に一意的ラベルを割り当て、4 つのタスクを作成しました。データベースは、200 個の SA-Hタスクと、約 2k の SA-QID、-Q、-A タスクを生成します。最後に、これらの生成されたタスクから 500 個のタスクをサンプリングしました。

合計で、トレーニング分割用に 100 個のデータベース (50000 タスク)、および検証分割用に 10 個のデータベース (5000 タスク) を作成しました。コンテキストはデータベース間で重複しません。

検索 UI はエンタリで部分一致を使用します

C.4 ゴールド シーケンス長の分布

図9は、ゴールド アクション シーケンスの長さの分布を示しています。ほとんどすべての例が、トレーニング中に設定した 50 ステップの上限内に収まっています。テキストボックスに単語を入力する必要があるタスクは、手順の数が長くなる傾向があります。

D BUI-BERT に関する追加情報

D.1 OCRエミュレーション

この作業では、実際の OCR の代わりに、スパン タグを使用して HTML ソースの各単語を囲む OCR エミュレーションを使用しました。図10に例を示します。技術的な理由により、エミュレーションはテキスト ボックス内のテキストをキャプチャしません。単語は、上から順にソートされ、左から順にソートされます。ソートは基本的に自然な順序を保ちますが、図のように順序が崩れることがあります。

D.2 ミニバッチ戦略

図11は、トレーニングに使用したミニバッチ処理を示しています。充填率を上げるために、一連のミニバッチに複数の軌道を詰め込みました。記憶と最後のアクションを軌道に対して繰り返し入力し、軌道の各先頭でそれらをリセットします。

D.3 BUI-BERT の学習曲線

図12は、BUI モデルの学習曲線を示しています。PTA 研修では、3 つのモデルが大まかに収束しました。マルチタスクトレーニングでは、BUI-BERTsmall w/o PTAを除くすべてのモデルが 10 エポックでほぼ収束しました。ただし、PTAなしの BUI-BERTsmallの損失は、5k 更新前後で大幅に減少し始め、10 エポック後には小さくなる可能性があります。これは、PTAが損失の収束を少なくとも加速することを示していますが、より長い時間の後に達成される最終的なパフォーマンスには影響しない可能性があります。

D.4 タスク実行のケース

理解を助けるために、図13に BUI BERTsmallを使用したタスク実行のケースを示します。

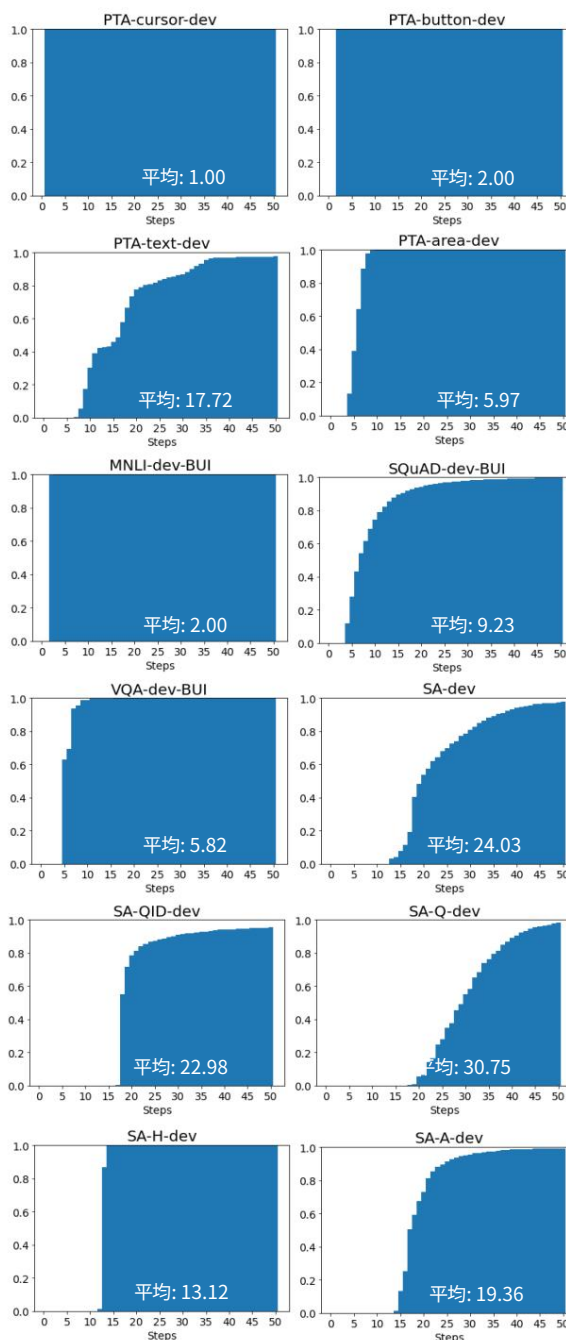


図 9: dev 分割におけるゴールド アクション シーケンスの長さの分布。累積値を表示します。文書分類タスクのアクション数は基本的に 2 つなので、代表的な例として MNL を示した。

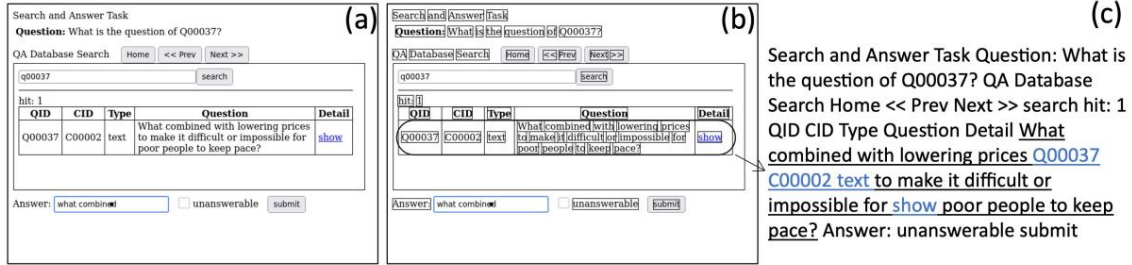


図 10: OCR エミュレーションの例。(a) 画面例。(b) 検出された単語。検出された単語は、塗りつぶされたボックスで囲まれています。(c) 得られたテキストシーケンス。順序が崩れている部分には下線を引いています。

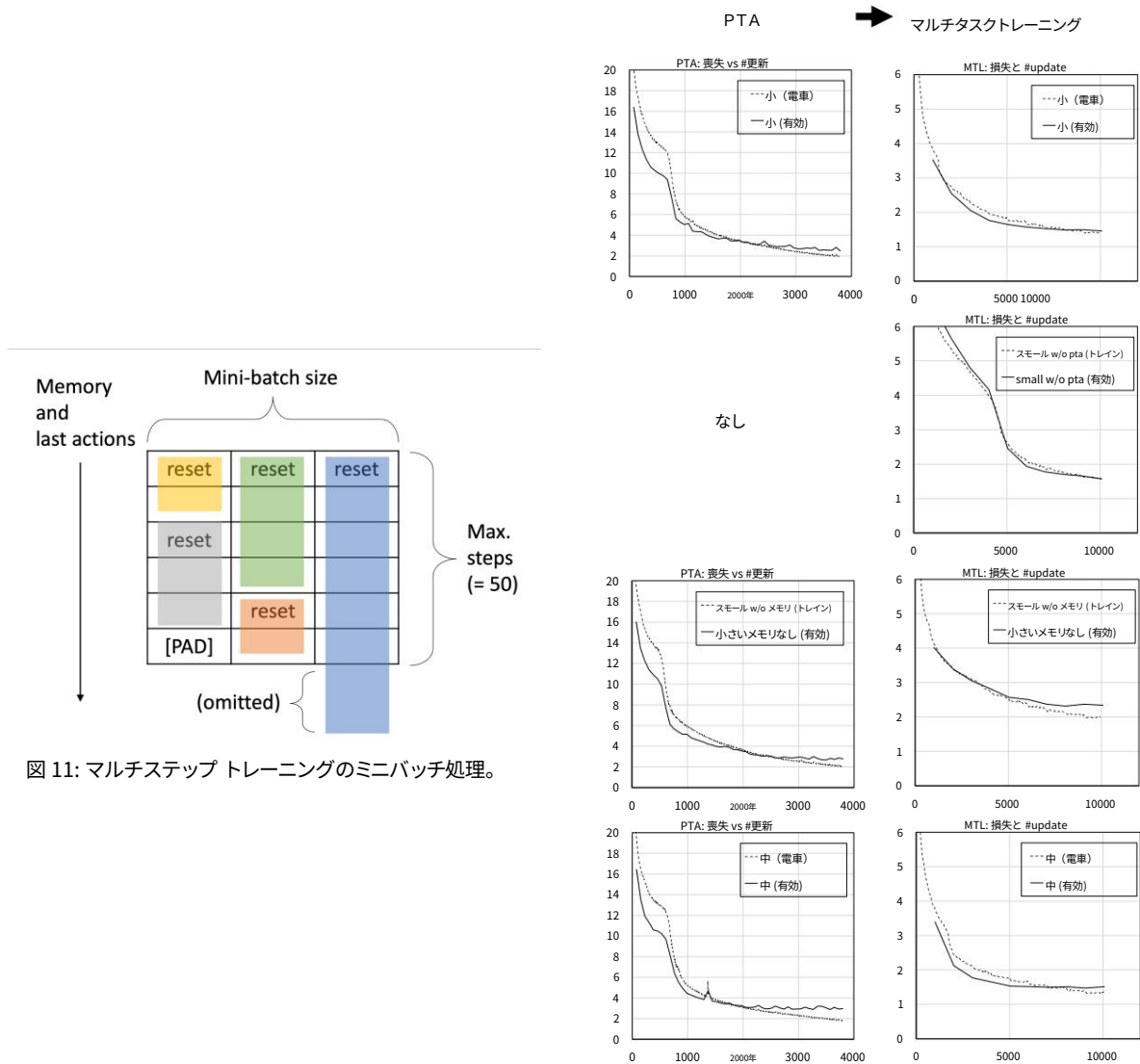


図 11: マルチステップトレーニングのミニバッチ処理。

図 12: BUI モデルの学習曲線。

(a) 失敗 (タイムアウト)。 (CoLA val.554)

Text Classification Task
If the following sentence is acceptable as an English sentence, press the acceptable button; if not, press the unacceptable button.

The newspaper has reported that they are about to appoint someone, but I can't remember who the newspaper has reported that they are about to appoint.

Classes:
 unacceptable acceptable

Text Classification Task
If the following sentence is acceptable as an English sentence, press the acceptable button; if not, press the unacceptable button.

The newspaper has reported that they are about to appoint someone, but I can't remember who the newspaper has reported that they are about to appoint.

Classes:
 unacceptable acceptable

Text Classification Task
If the following sentence is acceptable as an English sentence, press the acceptable button; if not, press the unacceptable button.

The newspaper has reported that they are about to appoint someone, but I can't remember who the newspaper has reported that they are about to appoint.

Classes:
 unacceptable acceptable

(b) 成功。 (SA 値 34)

Search and Answer Task
Question: What is the QID of the question "What type of revolution did Maududi advocate?" ?

QA Database Search Home << Prev Next >>

what type of search

hit: 36

QID	CID	Type	Question	Detail
Q00048	C00001	text	Constitutional impasse is different from civil disobedience because does not include what type of person?	show
Q00060	C00058	text	What type of numbers are always multiples of 2?	show
Q00078	C00064	image	What type of meat is on the plate?	show
Q00133	C00196	image	What type of vehicle is this, in the	show

Answer: unanswerable submit

Search and Answer Task
Question: What is the QID of the question "What type of revolution did Maududi advocate?" ?

QA Database Search Home << Prev Next >>

Q00233	C00001	text	What type of person can not be attributed civil disobedience?	show
Q00277	C00141	text	To what type of organisms is oxygen toxic?	show
Q00296	C00058	text	What type of numbers are always multiples of distinct divisors?	show
Q00319	C00039	image	What type of condiment is on the top shelf second from the right?	show
Q00416	C00008	image	What type of sink is in the bathroom?	show
Q00464	C00144	text	What type of revolution did Maududi advocate?	show
Q00494	C00168	image	What type of clouds are those on	show

Answer: unanswerable submit

Search and Answer Task
Question: What is the QID of the question "What type of revolution did Maududi advocate?" ?

QA Database Search Home << Prev Next >>

Q00233	C00001	text	What type of person can not be attributed civil disobedience?	show
Q00277	C00141	text	To what type of organisms is oxygen toxic?	show
Q00296	C00058	text	What type of numbers are always multiples of distinct divisors?	show
Q00319	C00039	image	What type of condiment is on the top shelf second from the right?	show
Q00416	C00008	image	What type of sink is in the bathroom?	show
Q00464	C00144	text	What type of revolution did Maududi advocate?	show
Q00494	C00168	image	What type of clouds are those on	show

Answer: q00464 unanswerable submit

(c) 失敗。ゴールド回答 : 第 30 条、モデル : 回答不可 (SA val. 42)

Search and Answer Task
Question: Answer the question of Q00773.

QA Database Search Home << Prev Next >>

q00773 search

hit: 1

QID	CID	Type	Question	Detail
Q00773	C00020	text	Which TEFU article states that no quantitative restrictions can be placed on trade?	show

Answer: unanswerable submit

Search and Answer Task
Question: Answer the question of Q00773.

QA Database Search Home << Prev Next >>

QID	Q00773
Question	Which TEFU article states that no quantitative restrictions can be placed on trade?
CID	C00020
Context	Although it is generally accepted that EU law has primacy, not all EU laws give citizens standing to bring claims: what is, not all EU laws have "direct effect". In Van Gend en Loos v Nederlandse Administratie der Belastingen it was held that the provisions of the Treaties (and EU Regulations) are directly effective, if they are (1) clear and unambiguous (2) unconditional, and (3) did not require EU or national authorities to take further action to implement them. Van Gend en Loos, a postal company, claimed that

Answer: unanswerable submit

Search and Answer Task
Question: Answer the question of Q00773.

QA Database Search Home << Prev Next >>

Context

regulations in their own law, in order to prevent confusion. For instance, in Commission v Italy the Court of Justice held that Italy had breached a duty under the Treaties, both by failing to operate a scheme to pay farmers a premium to slaughter cows (to reduce dairy overproduction), and by reproducing the rules in a decree with various additions. "Regulations," held the Court of Justice, "come into force solely by virtue of their publication" and implementation could have the effect of "jeopardizing their simultaneous and uniform application in the whole of the Union." On the other hand, some Regulations may themselves expressly require implementing measures, in which case those specific rules should be followed.

Answer: unanswerable submit

(d) 失敗。金の答え : グレー、モデル : 青。 (SA 値 46)

Search and Answer Task
Question: Answer the question of Q00249.

QA Database Search Home << Prev Next >>

q00249 search

hit: 1

QID	CID	Type	Question	Detail
Q00249	C00050	image	What colors are the planes?	show

Answer: unanswerable submit

Search and Answer Task
Question: Answer the question of Q00249.

QA Database Search Home << Prev Next >>

QID	Q00249
Question	What colors are the planes?
CID	C00050
Context	

Answer: unanswerable submit

Search and Answer Task
Question: Answer the question of Q00249.

QA Database Search Home << Prev Next >>

Context



Answer: blue unanswerable submit

(e) 成功。金の答え : 3 番目、モデル : 3 番目に豊富な元素。 (SA 値 67)

Search and Answer Task
Question: Answer the question of Q01114.

QA Database Search Home << Prev Next >>

q01114 search

hit: 1

QID	CID	Type	Question	Detail
Q01114	C00075	text	Compared to other elements, how abundant does oxygen rank?	show

Answer: unanswerable submit

Search and Answer Task
Question: Answer the question of Q01114.

QA Database Search Home << Prev Next >>

QID	Q01114
Question	Compared to other elements, how abundant does oxygen rank?
CID	C00075
Context	Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O ₂ . Diatomic

Answer: unanswerable submit

Search and Answer Task
Question: Answer the question of Q01114.

QA Database Search Home << Prev Next >>

Context

number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O₂. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

Answer: hird-most abundant element unanswerable submit

図 13: ケーススタディ。(a) モデルは、move_to (172, 200)、click.token (「受け入れられない」)、move_to (172, 178)、click、token (「受け入れられない」)、move_to (172, 200)、を繰り返しました。.. (b) モデルは、ゴールド シーケンスと同じ戦略である最初の 3 つの単語をクエリし、リストを取得しました。質問が表示されるまで下にスクロールし、QID を正常に抽出しました。(c) モデルは詳細に進み、すべてのコンテキストを読み取ります。ただし、回答可能な質問には回答不可のチェック ボックスをオンにしました。(d) モデルは写真を見るために詳細に行きました。回答タイプは正解でしたが、回答がゴールド回答とは異なりました。(e) モデルは詳細に進み、すべてのコンテキストを読んで正しく答えました。