

ボトムアップ・トップダウンの反復処理による画像解釈

Image interpretation by iterative bottom-up top-down processing

Shimon Ullman, Liav Assif, Alona Strugatski, Ben-Zion Vatashsky, Hila Levy, Aviv Netanyahu, Adam Yaari (2021)

情景理解には、物体やその部品、人物、場所などの情景構成要素を、それらの個々の特性だけでなく、それらの間の関係や相互作用とともに抽出し、表現する必要がある。ボトムアップ(BU)ネットワークとトップダウン(TD)ネットワークを組み合わせ、それらの間の対称的な双方通信(逆行流 counter stram 構造)を通じて相互作用する反復処理によって、意味ある情景構造を画像から抽出するモデルについて述べる。BU-TD モデルは、情景の構成要素を抽出し、その選択された特性と関係を認識し、それらを用いて画像を記述し理解する。情景表現は、3つの成分を繰り返し使用することで構築される。最初のモデル成分は、選択された情景要素、特性、および関係を抽出するボトムアップ流である。2つ目の成分(認知増強 cognitive augmentation)は、関連する非視覚的な保存表現に基づいて、抽出された視覚表現を補強する。また、第3の成分であるトップダウン流に、TD 命令の形で入力を提供し、次に実行すべきかだいをモデルに指示する。トップダウン流は、次のサイクルで選択された課題を実行するよう、BU 視覚流を誘導する。この過程で、画像から抽出された視覚表現は、関連する非視覚表現と組み合わせることができ、最終的な情景表現は、情景から抽出された視覚情報と、関連する世界の記憶された知識の両方に基づく。我々は、BU-TD モデルが、個々の TD 命令によって呼び出される一連の段階から、複雑な視覚課題をどのように構成するかを示す。特に、TD 命令の系列が、情景から興味のある構造を抽出するためにどのように利用されるかを、系列の次の TD 命令を自動的に選択するアルゴリズムも含めて説明する。TD 命令の選択は、一般に、目標、画像、および前段階で画像から既に抽出された情報に依存する。したがって、TD 命令系列は、最初に決まった固定的な系列ではなく、目標と画像に依存する進化するプログラム(または「視覚ルーチン」)である。この抽出処理過程は、組み合わせによる汎化という点で好ましい特性を持ち、新しい情景構造や、訓練中に見られなかった物体、特性、関係の新しい組み合わせに対してよく汎化することが示される。最後に、このモデルを人間の視覚の関連する側面と比較し、情景理解の処理において視覚的要素と認知的要素を統合するために BU-TD スキームを利用する方向性を提案する。

- 1. 一般的な背景と目標：視覚と認知の融合
- 2. BU-TD 逆行流ネットワーク
 - 2.1 ネットワーク構造
 - 2.2 逆行流モデルにおける学習
- 3. トップダウンの指示
 - 3.1 物体
 - 3.2 プロパティ
 - 3.3 関係
 - 3.3.1 人間と物体の相互作用
 - 3.3.2 物体の関係
 - 3.3.3 空間関係
 - 3.3.4 高次構成の抽出
 - 3.3.5 位置による汎化
 - 3.3.6 指示の参照
- 4. BU-TD 系列を用いた情景構造の抽出
 - 4.1 情景記述の抽出
 - 完全構造の抽出
 - ガイド付き構造抽出
 - 4.2 視覚ルーティンの構成：次の TD 命令を選択する
 - 非誘導情景解析
- 5. 容量と一般化
 - 5.1 マルチタスクと容量
 - 5.1.1 課題選択性
 - プログラム修正としての課題選択
 - 5.2 組み合わせによる汎化
 - 組み合わせ汎化: 右隣と左隣の関係
 - 汎化と複合命令の組み合わせ

- 5.3 命令の記号表現から埋め込み表現へ

- 6. 人間の視覚との関係

- 6.1 靈長類の視覚における反対流路

- 6.2 記号的表現と構成的／組込みの表現

- 6.3 TD ガイダンスの機能と利点

- 文献

- 補足資料

- 補足資料目次

- S1. 心理物理学的タイムライン研究

- S2. BU-TD 構造

- S3. The Persons data set

- S4. The Actors data set

- S5. 位置に基づく分類

- S6. 空間関係

- 非参照型の空間関係

- S7. 全情景構造の抽出

- S8. 情景構造のガイド付き抽出

- S9. 次のTD命令の選択：視覚ルーチンの構成

- S10. 組み合わせによる汎化

- 実験の詳細

- 組み合わせ汎化 - Persons データセット

- 組合せ汎化 - EMNIST データセット

1. 一般的な背景と目標：視覚と認知の融合

視覚は、我々の目に届く画像に基づいて、我々を取り巻く世界を理解することを可能にする。このような理解には、物体、人、場所などの様々な情景の構成要素を、それぞれの特性とともに、それらの間の関係や相互作用とともに抽出し、表現することが必要である。以下に、情景の構成要素をそれらの特性と関係とともに抽出・認識し、それらを用いて画像を記述・理解するモデルについて述べる。以下に述べる一般的なアプローチでは、ボトムアップ(略称BU)とトップダウン(略称TD)の流路を組み合わせた反復処理によって画像から意味のある情報を抽出することを提案する(図1 視覚-認知: 視覚と認知の融合)。このスキームでは、モデルの視覚部分(「視覚」とラベル付けされる)は、モデルの認知部分(「認知」とラベル付けされる)と呼ぶ、より高いレベルで情景表現を構築する。情景表現は、図1に模式的に示した3つの成分を繰り返し使用することで構築される。「視覚流路」とラベル付けされた最初の成分は視覚から認知へ、「認知増強」とラベル付けされた2番目の成分は認知部内で行われ、「トップダウン流」とラベル付けされた3番目の成分は認知部から視覚部に戻る。簡単に説明すると、視覚流路は、高レベルで有用な視覚表現を認知部分に届ける。「認知増強」成分は、関連する非視覚的な保存表現に基づいて、抽出された視覚表現を補強する。また、次に抽出する関連情報を選択するトップダウン流にも入力を提供する。トップダウン流は、次のサイクルで選択された情報を抽出するように視覚流を誘導する。この過程で、画像から抽出された視覚的表現が関連する認知的表現と組み合わされ、最終的な情景表現は、情景から抽出された視覚的情報と、世界に関する関連する記憶された知識の両方に基づいている。

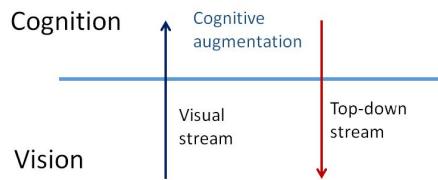


図1：視覚と認知。視覚から認知へ向かう「視覚流」、認知部内で行われる「認知増強」、認知部から視覚部へ戻る「トップダウン流」。

この3つの構成要素の役割とそれらの統合を、図2aに示す画像例(財布を盗む人)を用いて模式的に示す。関連する情景構成要素(例: 男、女、バッグ、椅子)、特性の一部(例: 赤いバッグ)、および関係(例: 男が財布をつかむ)を含む情景表現例を、図2bのグラフ(青い部分)で示す。この場面で重要な関係は、座っている女性が赤い財布の持ち主である可能性が高いということである。「所有権」は直接的な視覚的関係よりもむしろ意味的な関係であり、しばしば特定の知識に依存する(例えば、ほとんどの場合とは異なり、ポーターや警備員は自分が操作するバッグを所有していない)。上記のアプローチでは、モデルの提案された「認知増強」部分は、関連する非視覚認知表現に基づいて関係と属性を追加する。具体的には、現在の情景では、関連する記憶された知識を使用して、座っている女性とバッグの間の所有関係を追加することができる(図2bの赤破線)。続く段階では、トップダウン成分を使用し、次に女性の視線方向(赤の実線)を抽出するようモデルに指示する。これにより、バッグをつかんでいる人物が彼女の視野内にいるかどうかが判断され、その結果、**盗んでいる**という解釈が導き出される。

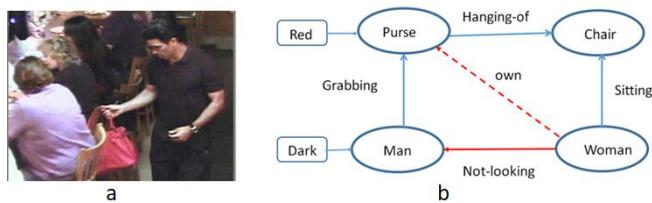


図2：盗み。座っている女性から財布を盗む男。a. 入力画像。b. 情景の構成要素、その特性、関係から見た情景記述。赤破線：認知補強による追加。赤実線：最終的なトップダウン命令によって抽出。

この画像(図2a)を他の情景とともにオンライン・人間研究に使用し、画像から構造化された情景情報(物体、人物、特性、関係)の抽出を時間(50ミリ秒～2秒)で追跡した。この心理物理学的タイムライン研究の結果は別項で述べる。詳細は補足S1(心理物理学的タイムライン研究)を参照。図2(盗み)画像について、(b)の記述は「盗み」イベントが報告されるまでに、大多数の被験者によって報告されたすべての情景構成要素、特性、関係をグラフにまとめたものである。複数の情景構成要素とその相互関係の抽出は、時間の経過とともに徐々に発展する過程を示した。例えば、100ミリ秒(msecs)という短い呈示時間では、男性と女性のみが報告され、バッグ(125msecs)、「座る」と「椅子」(200msecs)はその後に現れ、「盗む」は500msecs以上の呈示時間をかけて報告された。情景記述の経験的な時間軸は、所有関係の出現が遅く、視線方向の抽出がそれに続くなど、上述の時間的順序を支持している。(補足S1の図30(盗みタイムライン)参照)。

この一般的なアプローチと3つのモデル成分の中で、視覚流路は、経験的にも計算的にも、過去に広く研究してきた。認知増強の部分は主に非視覚的なものであり、ここでは説明しない。本稿の焦点は、いわゆるカウンターストリーム構造でボトムアップ流路とトップダウン流路を組み合わせた、視覚成分と認知成分間の双方向通信にある。次節では、BU-TD ネットワークの構造とその学習手順について述べる。

2. BU-TD 逆行流ネットワーク

2.1 ネットワーク構造

まず、反対流モデルの一般的な構造を説明し、次に、モデルの訓練に使用される学習手順と、視覚処理をガイドするためのTD命令の使用について説明する。このスキームは、視覚成分と認知成分間の双方向通信を提供する2つのネットワーク流路、すなわち、ボトムアップ(BU)流路とトップダウン(TD)流路に基づいている。この2つは構造は同じであるが、反対方向に向いているため、視覚野の階層に沿った皮質結合の側面を捉える「反対流」構造を作り出す(Ullman1995, Markov+2013, 2014)。BU流路は標準的なディープネット(ResNetなどHe+2016を使用)としてモデル化される。反対流構造を図3(反対流構造)に示す。

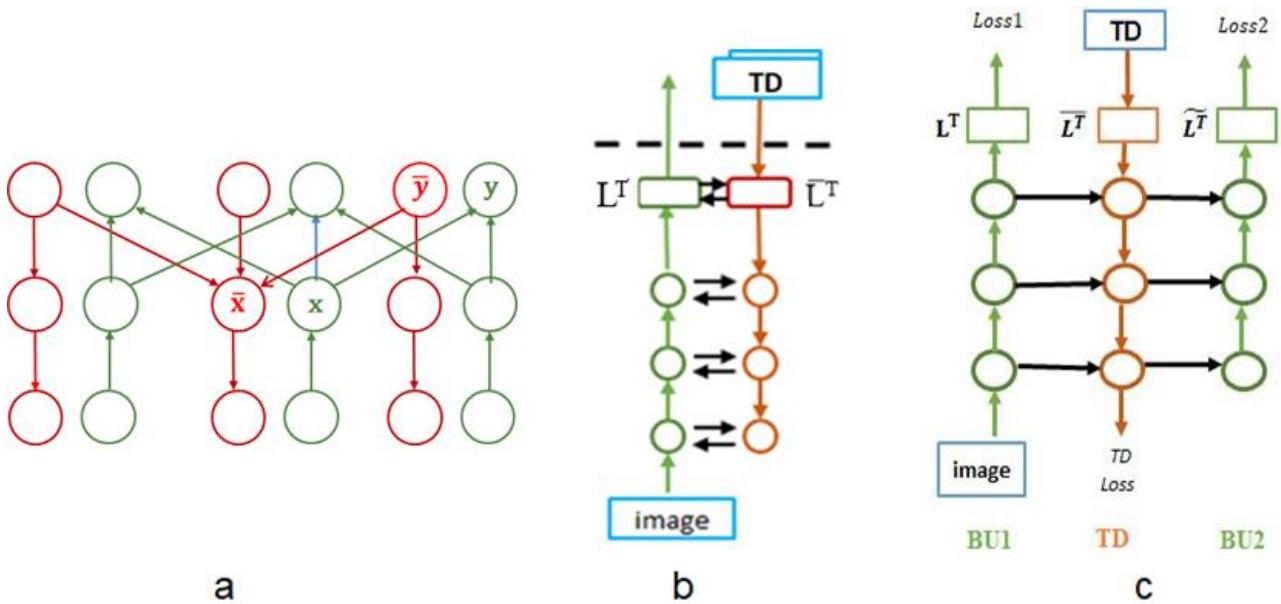


図3：逆行流構造

- a. ボトムアップ(緑色の「ニューロン」)流とトップダウン(赤色の「ニューロン」)流内のユニットの相互接続(交差流接続は図示せず)。
 - BU 流への入力は画像であり、TD 流への入力は認知部からのトップダウン命令である。緑と赤の円は、BU 流と TD 流に沿った層を表す。 L^T は BU 流の最上位層、 \bar{L}^T は TD 流の入力層である。
 - c. BU-TD ネットワークの時間的「展開」による学習。BU1, BU2 は同じ重みを持つ。最終出力は BU2 の一番上(L^T)で、学習誤差は Loss2 で測定される。BU1 の最後(Loss1)と TD 流路(TD loss)の 2 つのオプションの補助測定がある。詳細は図 5(全構造)を参照。

図3aは、BUネット(緑)と相補的なTDネット(赤)のごく一部を示している。図のように、ボトムアップ流路上の各ユニット x には、トップダウン流路上に対応するユニット x が存在し、 x がターゲットユニット y に接続されている場合、トップダウン流路上には y から x への反対方向の接続が存在する。反対側の流路上の対応する層は、交差流路接続(図示せず)によって両方向に接続される。構造の詳細は、以下の図5(完全構造)および補足S2(BU-TD構造)に記載されている。

図3bに示すように、ボトムアップ流路への入力は画像であり、トップダウン流路への入力は認知部からのTD命令である。この命令はいくつかのベクトル(通常2~3個)から構成され、次の処理サイクルのガイダンスとして使用される。TD流路のトップレベル(図中のLTはTop-Layerを意味する)で、この命令はBU流路(図中のLT)によって生成されたトップレベルの画像表現と結合され、TD部の全層を伝播する。TD流へのこの入力は、図5および補足S2(BU-TD構造)でより詳細に示され、説明されている。交差流接続を通じて、次のBUパスはこのTD表現の文脈で行われる。その結果、BUパスは、命令、画像、以前のサイクルによって制御され、導かれる。図3cは(b)の構造を「展開」したもので、次節で説明するネットワークの学習に使用される。この結合されたネットワーク、またはその展開されたバージョンをBU-TDネットワークと呼ぶこととする。

訓練処理の後、異なる視覚課題を実行するための指示をTDネットワークのトップダウン指示入力として与え、次の処理段階を選択・誘導することができる。後述するように、BU視覚流は、指示された課題を効率的な方法で選択的に実行し、関連する課題にうまく汎化する。さらに、4節(BU-TD系列を用いた情景構造の抽出)で説明するように、TD命令の系列は、構造化された視覚表現を抽出するためのプログラムとして適用することができる。

ネットワーク、訓練、機能を詳細に説明する前に、以下の図4(TD指示)は、ネットワークにトップダウンの指示を与え、適切な応答を得ることが何を意味するかを示す簡単な例を示している。入力画像(図4の例)には、我々のPersonsデータセットから、可変数の合成人物の顔(顔、首、肩)が含まれている。各人物は(モデルが認識するように訓練された番号、または名前で識別される)、いくつかの不变の顔の特徴と、異なる眼鏡、シャツ、髪型などの可変の特性との組み合わせによって定義される(補足S3(Personsデータセット)参照)。

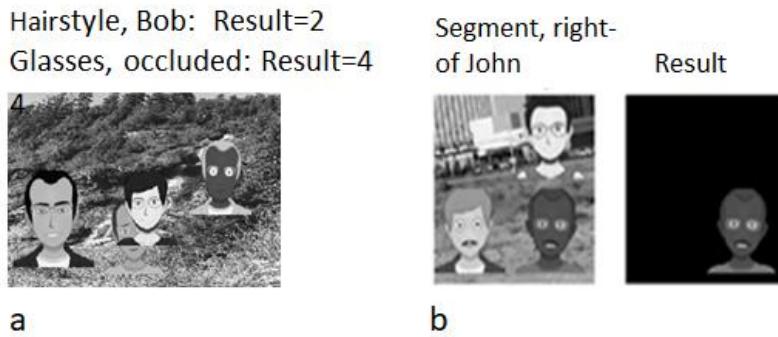


図4: TD命令TD命令とモデルの応答の例。

- a. 同じ入力画像に対する異なるTD命令: 特定の人物('Bob')のヘアスタイルを分類する命令ではHairstyle=2が生成され、隠れている人物のメガネのタイプを分類する命令ではGlasses-type=4が生成される。
- b. ジョンの右の人物をセグメンテーションするTD命令は、右のセグメンテーション画像を生成する。

ネットワークは、選択された人物の特定の特性を抽出するために、後述する手順で学習される。現在の目標は、モデルの視覚流路に、画像中の人物の一人であるBobの髪型を抽出するよう指示することであるとする。これは、トップダウン流路にTD命令を供給することで得られる。TD命令は、2つの学習済み命令ベクトルによって、(i)実行すべき課題(抽出すべき特性)、および(ii)引数、すなわち課題を何に適用するか(例えば上記のBob)を指定する。以下でさらに説明するように、TDネットは、TD命令によって指定された課題と引数を適用するようにTD-BUサイクルを導くために、命令ベクトルを使用することを学習する。人間の視覚(Fink+1997)に類似して、引数は、物体ベースの注意に類似した人物の同一性、または空間的注意に類似した位置、さらに関係(例えばright-of, occluded-by)によって、反対流路で指定することができる。同じ入力画像に対して、適切なTD命令を選択することで、異なる課題を適用することができる。「ボブの髪型」(図4a)という命令に対して、モデルは「タイプ2」という正解を出す。他の特性は、一度に1つずつ、または複数の特性を組み合わせて抽出することができ、異なる人物に適用することができる(3.2節特性を参照)。例えば<メガネのタイプ、メガネをかけた人>という命令では、メガネのタイプ4が識別される。図4bでは、TD命令はジョンの右の人物をセグメンテーションすることであった。出力はセグメンテーションマップであり、この場合はTD流路の下部に生成される。

2.2 逆行流モデルにおける学習

次に、どのように逆行流ネットに命令が与えられ、どのようにネットが命令に従って学習し、指示された出力を生成するかを説明する。BU-TDモデルの学習は、リカレント版のディープ・ネットワークの標準的な手法(Werbos1990, LeCun+2015)を用いて達成される。リカレントネット(RNN)を扱うために、誤差逆伝播ベースの訓練は、図3cに示すように、ネットワークの時間的な「展開」を使用して、リカレントネットを標準的なフィードフォワードモデルに変える「通時逆伝播」を使用して拡張される。これは図3bを展開したもので、3つの列があり、最後の列は最初の列と同じであるが、後の時点のものである。モデルには入力画像とTD命令の両方が同時に与えられ、学習は命令で

指定された出力を抽出するために使用される。図3cでBU1とマークされた左の列は、ネットを通る最初のBUパスであり、その後にTD(トップダウンパス)が続き、その後にBU2が続く。BU1もBU2も同じボトムアップ回路を使用しているため、展開ネットの重みは同じになるように制約されている。展開ネットでは、出力はBU2パスの先頭で生成される。BU1とBU2が同じネットワークを使用するリカレント・ネットでは、出力はボトムアップ部分の先頭、BU-TD-BUサイクルの終わりに読み出される。BU-TDモデルはリカレント・ネットワークであり、複数のサイクルに使用できる。図3cの展開されたバージョンは、単一のBU-TD-BUサイクルのためにネットワークを訓練する。しかし、補足S2(BU-TD構造、複数サイクルの訓練)で説明するように、ネットワークを複数サイクル用に訓練することもできる。訓練は、各々がTD命令と組み合わされた一組の訓練画像を用いて進められる。図4の画像(TD命令)では、特定の人物の選択された心像性を抽出する命令、選択された人物をセグメント化する命令、人物間の所定の関係を計算する命令などが例示される。学習は、トップダウン接続と交差流路接続の重みを含むネットワークの重みを学習するために用いられる。図5(全体構造)は、展開されたネットワークの全体図であり、トップダウンの指示がTD流路に与えられる部分(赤枠)を示している。図6(命令の提供)で拡大されたネットのこの部分については、次節で詳しく説明する。交差流路接続を含む全構造の詳細については、補足S2(BU-TD構造)に記載されている。図5のBU-TDネットワークはResNet-18(He+2016)に基づいているが、BU-TD構造のBU部分にはどのような標準的なディープネットワークでも使用することができ、完全なBU-TDネットワークは、補完的なTDネットワークと前述の図3に模式的に示された横方向の接続を追加することによって作成することができる。

1つは実行する課題(I_{task} と表記)を指定し、もう1つ(I_{arg} と表記)は引数、つまり課題を適用する物体を指定する。本節で示す例では、課題 I_{task} は「ワンホット」ベクトル表現を用いて学習中にモデルに指定される。例えば、 k 番目のエントリにある1つの「1」は(可能なタスクのリスト中の)課題 k を表す。ある場所の1は、メガネのタイプを分類する課題、別の場所はシャツのタイプを分類する課題、または物体の分類などを表す。後述するように(5.3節記号表現から埋め込み命令表現へ)、これは、課題や他の課題との関係に関する情報を提供せずに課題を指定するという意味で、「純粹な」記号表現形式と考えることができる。引数 I_{arg} も同様に、例えばモデルに提示された異なる人物に対して、ワンホットベクトルを用いて指定された。特に、引数が画像の位置である場合、入力画像と一緒に空間地図によって提供することもでき、後述する特定の関係を扱うようないくつかのケースでは、引数は1つではなく2つのベクトルを含む。

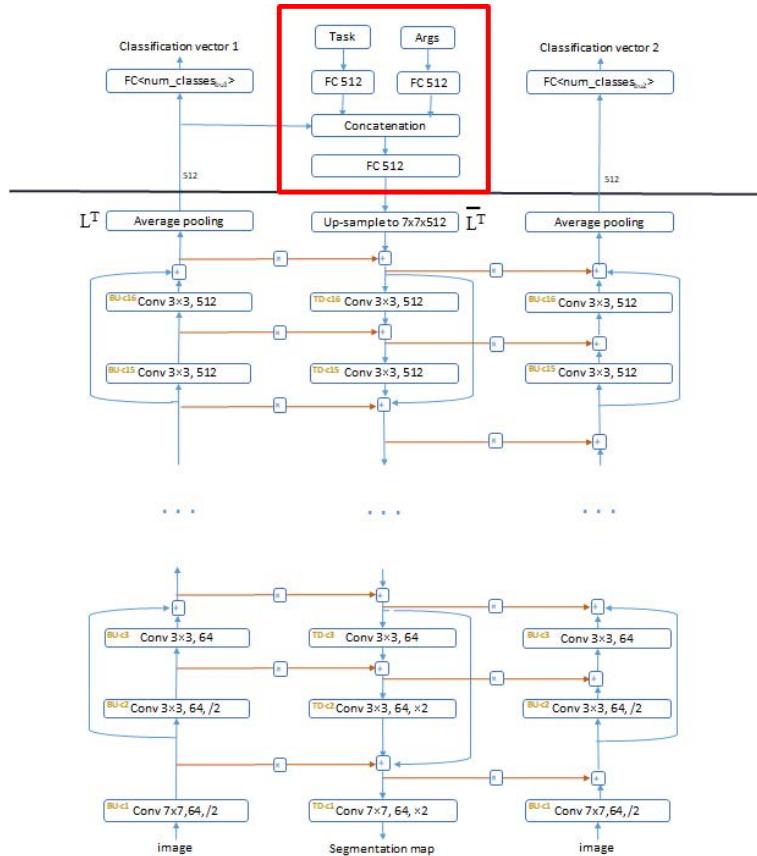


図5: 完全構造。BU-TDネットワークの完全な構造。この例のBUネットワークはResNet-18である。赤枠は、図6(命令の提供)で詳述したTD命令が提供される部分を囲んでいる。各層は、層識別子(BU-c3など)、実行する演算(畳み込みはconv、完全連結はFC)、次元を示し、+やxのついた小さなボックスは加算演算や乗算演算を示す。

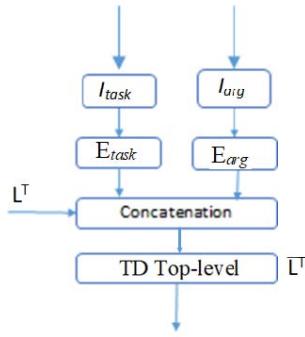


図 6: 指示の供給。TD 命令を BU-TD ネットワークに提供する。 I_{task}, I_{arg} は記号表現(例えばワンホットベクトル)で課題と引数を提供するベクトルである。 E_{task}, E_{arg} は課題と引数の埋め込み形式で、学習された重みによって前の表現から生成される(全結合)。 \bar{L}^T は TD 流の最上位層、 L^T はボトムアップ流の最上位層 BU1 から来る入力。

学習中 ワンホットベクトル I_{task}, I_{arg} は新しい学習済み表現(E_{task}, E_{arg} , E は埋め込み embedded を表す)に変換される。この表現を作成するネットワーク重み(最初はランダム)は、学習期間中に学習される。後述するように、最初の記号表現とは異なり、変換された「埋め込み」表現は、表現された物体に関する有用な情報を持っている。

これら 2 つの学習済み命令ベクトルは、BU 流のトップレベル表現(図 3 逆行流構造の L^T)と共に、すべて同じ次元で、TD 流のトップレベル表現(図 3 の \bar{L})への入力となる。図 4 の例では、2 つの損失関数を用いてネットワークを学習した(单一の損失を用いるか、あるいは補助的な損失を追加することも可能である、補足 S2 (BU-TD 構造) 参照)。最初の損失は最初の BU 段階(BU1)の最後に適用される。人物の画像を抽出するために、ネットワーク BU1 は画像内の全ての人物(最大 6 人)を識別するように学習された。次の BU 段階(BU2)で望まれる出力は、選択された人物の正しい特性であり、訓練で使用される損失関数は、正解とネットが生成したものとの間の距離(クロスエントロピーなど)である。訓練で使用される 3 つの損失関数は、TD 流(図 3)の最後にあり、セグメンテーション課題などで使用される。訓練中、BU-TD スキームは命令ベクトル E_{task}, E_{arg} を使用することを学習し、後続の BU パスが関連する引数の必要な特性(例えば Bob の眼鏡)を抽出するように導く。訓練処理の後、異なる視覚課題を実行する指示を TD ネットワークに与えることができ、視覚流は、選択された引数に適用される指示された課題を選択的に実行する。ネットワークの構造と訓練の詳細については、補足 S2 (BU-TD 構造) を参照。コードとデータは <https://github.com/liavassif/BU-TD> にある。

3. トップダウンの指示

はじめに述べたように、画像内の複雑な構造を扱うには、視覚処理で物体や人物のような情景の実体を、それらの個々の特性だけでなく、それらの間の関係や相互作用とともに識別する必要がある。以下に、選択された物体、特性、関係を抽出する例を示す。同じネットワークに異なる課題を実行させ、選択した心像性に適用させることもできる。次節以降では、個々の指示がどのように組み合わされて、複雑な構造を効率的かつ広範に汎化しながら抽出できる「視覚ルーチン」の指示系列とフォームを形成するかを示す。

物体、特性、関係のトップダウン指示。 物体、特性、関係の指示抽出について、位置による関係の汎化、参照指示の利用を含めて以下に述べる。

3.1 物体

画像内の物体を抽出し識別する際に、最も頻繁に行われる課題は、物体分類、位置の特定、背景からの物体の分割である。これら 3 つの側面は、ほとんどすべての視覚課題に必要とされる、本質的な物体特性である。人間の視覚では、物体を認識することは通常、それが何であるか(基本クラス)、その位置を特定すること、そして背景から分離することと自動的に結びついている(Rosch+1976、Grill-Spector&Kanwisher2005)。コンピュータビジョンでは、多物体画像に適用される最初の段階に、画像内的一部またはすべての物体の分類、位置特定、分割が含まれることが多い(He+2017)。

逆行流ネットワークは、分類、局在化、分割を行うよう指示することができる、それらをさまざまな方法で組み合わせることができる。例えば、図 7a では、課題は選択された人物を分割することである。最初のボトムアップパス(BU1)では、TD 命令の前に、モデルは画像内に存在するすべての人物の名前を生成するように学習される(ここでは、1 人の人物が画像内に複数回登場することはないと仮定する)。選択された人物を背景から分割するために、モデルは `<segment, person>` という命令を用いて学習され、次に実行する課題は物体の分割であることが指定される。実行する課題と人物の身元は、ワンホットベクトルで表現される。分割結果は、次の BU パスではなく、TD パスの一番下に作成される。図 7b では、課題は再び分割であるが、引数には、選択された人物ではなく、場所(右上など)が指定される。図 7c,d では、分割のための引数は、隠蔽された人物を分割するか(c の場合)、カメラに最も近い人物を分割するか(d の場合)というプロパティによって与えられる。(d) の画像は、コンピュータが生成した複数の人物や物体のある複雑な情景を作成することができる、我々の Actors データセットから取り出したものである。

このネットワークによって生成される分割は、「実体」分割であり、いわゆる「意味」分割ではない。意味分割では、画像内のすべての物体が同じクラスに属する。異なるクラスや異なる位置の分割は、同じ BU-TD ネットワークでも、異なる TD 命令を使用することで生成される。

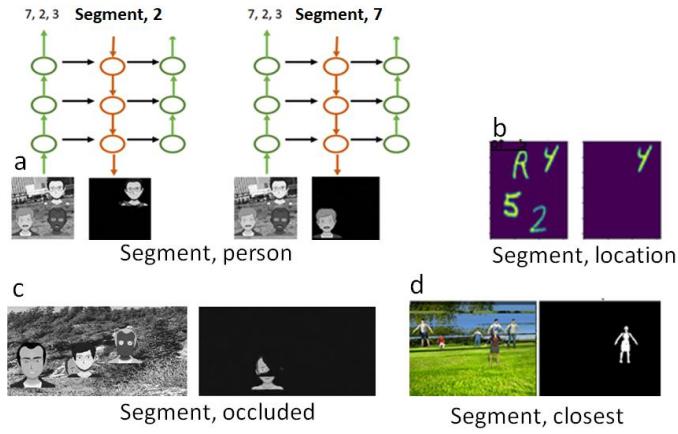


図 7: 選択的分割。TD 命令は各画像下に表示。

- a. 個々の人物ごとの分割 同一画像に、異なる命令を与えることで、異なる出力が得られる。個々の実体（人物 7 と人物 2）ごとの分割が、TD 流の最後に得られる。
- b. 場所による分割。MNIST 数字に対して、右上を分割せよとの命令による。
- c. 属性による分割。隠蔽された人物 (c) または、カメラに最も近い人物

以下の例では、視覚流は、選択された場所で物体の分類を行うように誘導される。このような選択 (Treisman&Gelade1980), つまり視覚過程を選択された場所に誘導する形式は、しばしば「注意のスポットライト」(Buschman&Miller2010) と呼ばれる。BU-TD モデルでは、この場合の誘導指示は <分類, 場所> という形式である。位置はさまざまな方法で指定することができ、特に、物体のバウンディングボックス、分割マスク、または中心位置によって指定することができる。

位置による分類の例を図 8a に示すが、これはいわゆる Multi-MNIST 画像 (Sabour+2017, Sener&Koltun2018) に適用したものである。2~9 桁の数字を用い、画像の固定された部分領域に部分的に重なるように配置し、TD 命令で示された位置にある数字を分類する課題とした。BU-TD ネットは LeNet (LeCun+1998) をベースとし、各位置で 60,000 例を用いた。ネットワーク、手順、結果の詳細は (Levi&Ullman2020) に記述されている。難易度の高い 9 桁のケースの分類精度は 88 %であり、これは 9ヶ所のうちの 1ヶ所でそれぞれ訓練された 9 個の個別の LeNet ネットワークによって得られた精度 (86.6 %) と同程度であった。

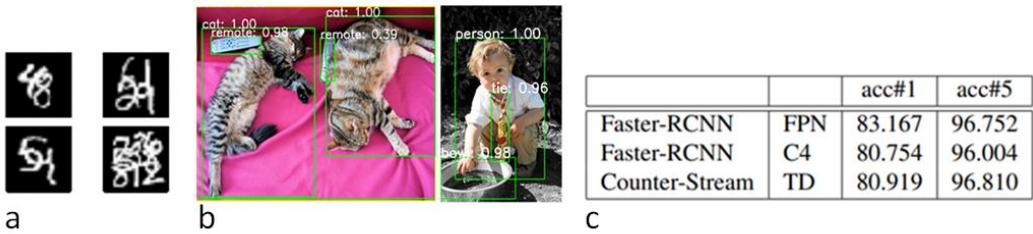


図 8: 分類-位置 Classify-location 位置による分類。

- a. 数字を最大 9 つの位置の 1 つに分類。
- b. 入力画像と一緒に与えられたバウンディングボックス内の各物体を分類し、その結果を信頼度とともにフレーム上部に示す。
- c. 最先端の手法と比較した分類精度。acc#1 と acc#5 は、それぞれトップ 1 分類とトップ 5 分類の精度である。FPN と C4 は Faster-RCNN モデル (He+2017) の 2 つの変種であり、詳細は補足 S5 (位置による分類) を参照。

図 8b は、自然画像における位置による分類の結果を示している。各物体について、正しい位置が分類指示とともにネットワークに供給された。(b) の位置は、物体のバウンディングボックス地図（緑色の輪郭で示す）の粗い地図によって、TD 方式で提供された。各フレームの上部には、信頼度付きの分類結果が示されている。物体の中心位置(x,y 座標)、バウンディングボックスの座標、空間地図としてのバウンディングボックス、物体の分割地図などである。これらのうち、バウンディングボックス地図が最も良い結果をもたらした。さらに、物体の位置を入力画像とともに BU 方式で提供する可能性を検証した。その結果、バウンディングボックス地図や分割地図を提供しても、同様の精度で利用できることがわかった。(c) に示す分類結果は、報告されている最良の結果に近い。比較における実装の詳細は、補足 S5 (位置による分類) に示す。

3.2 プロパティ

次例は、物体プロパティのガイド付き抽出を示している。図 9a は、画像内の異なる人物に関する選択された特性の抽出を示している。画像は Actors データセット (4.1 節 シーン記述の抽出、補足 S4 The actors data set) から取られた。同じ画像が与えられた場合、異なる TD 命令は、メガネを掛けている、髪型、その他(ここではプロパティとして「メガネを掛けている」を使用する)のような、同じ人物または異なる

人物の異なるプロパティに注目するようにモデルを導く。この場合、特性を抽出する人物は、モデルによって抽出されたセグメンテーション地図によって与えられ、BU2 の下部に補助入力として与えられる(詳細は4.1節および補足S7完全な情景構造の抽出で説明する)。その後、図のように、同じ人物の特性を追加で抽出し、さらに人物を選択し、その人物の特性を抽出するという処理を繰り返す。Actorsデータセットにおけるプロパティと関係の抽出に関する詳細と結果は、4節BU-TD系列を用いた情景構造の抽出でさらに説明する。

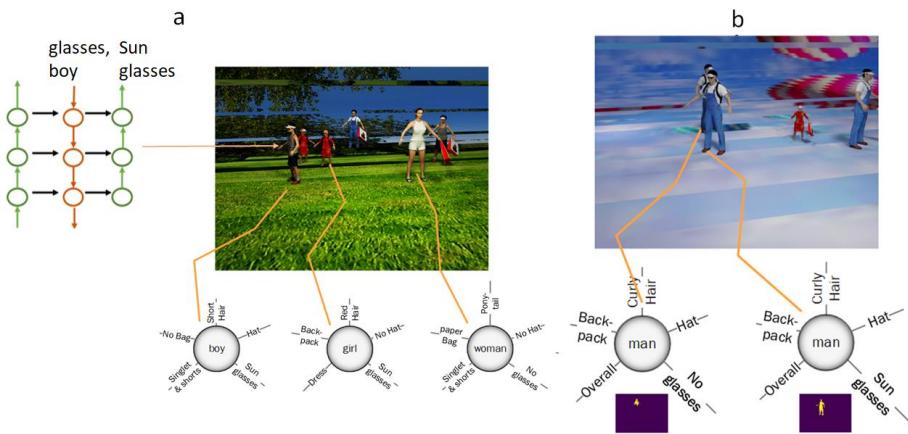


図9: プロパティの抽出選択された人物から選択されたプロパティを抽出。選択された人物の「メガネをかけている」プロパティを抽出するTD命令による、1つのプロパティの抽出。ネットワークへの命令は、少年を指す矢印で示されている。a(右)。同じ人物のすべてのプロパティを抽出するために、プロパティ抽出のプロセスが繰り返され、さらに人物が選択され、そのプロパティが抽出される。b. 重複度の高い2人の人物のプロパティを抽出する。下図は抽出された人物のセグメンテーション地図である。

プロパティとオブジェクトを結びつける：図に示すように、物体（または人物）とその特性を選択することで、画像内の物体とその特性を正しく関連付けることができる。この関連付けは知覚におけるいわゆる「バインディング」(Malsburg1995, Feldman2013)の一部であり、5.2節(組み合わせ汎化)でさらに詳しく説明する。図9Bは、非常に重なり合う2人の人物から特性を抽出する様子を示している。異なる物体からの特性のガイド付き抽出は、それらの物体間の空間的な分離を必要とせず、同じ空間領域内の異なる物体の特性は、対応する物体に正しく関連付けられる。

逐次抽出と並列抽出：上記の例では、モデルは一度に1つのプロパティを抽出するように学習され、その後、異なる物体の複数のプロパティ(例えば、図9aで選択された人物)を抽出するために、プロパティ抽出命令が繰り返された。抽出するプロパティの選択は、単一のプロパティに限定する必要はない。代わりに、物体に関連する複数のプロパティを、单一の命令を使用して同時に抽出することができる。選択された人物のすべてのプロパティを抽出する命令でネットワークを訓練することにより、人物とそのプロパティの図9と同じ表現も得られた。同じ物体の複数のプロパティを抽出する際、人物の「服装」や「ポーズ」のプロパティのように、関連するプロパティを群化し、1つの命令でそれぞれのセットを抽出することも可能である。

複合的な指示：最後に、物体の個々の性質を抽出する訓練は、追加訓練なしに、一度に1つ以上の性質を抽出することに自然に一般化することもわかった(Wagner2020)。

上述したように、特定のプロパティを抽出する命令は、プロパティベクトルによるワンホット符号化を使用した：選択されたプロパティは、表現においてオンになった1つの「1」によって表される。訓練後、2つのベクトル位置で‘1’をオンにした命令を使用すると、訓練では一緒に使用されることになったが、2つの選択されたプロパティを同時に抽出することにつながる。例えば、メガネタイプと髪型を抽出する命令を1つのベクトルにまとめると、選択された人物の両方の特性を正しく抽出することにつながる。複合命令は、より多くのプロパティに高精度で拡張することもできる。我々はPersonsデータセットで複合命令をテストし、個々の特性についてのみ訓練し、次に複合命令でテストした。各プロパティを個別に訓練し抽出した場合、平均精度は96%であった。追加訓練なしですべてのプロパティを同時に抽出すると、94%の精度が得られた。

我々は、このような複合命令は、同一人物の2つ以上の異なる特性を同時に抽出するためには使用できるが、2人以上の異なる人物の同じ特性を同時に抽出するためには使用できないことを発見した。この非対称性(特性は組み合わせられるが、人物は組み合わせられない)は驚くべきことではない。例えば、2人の人物のメガネ型を抽出する場合、人物とその正しい性質を組み合わせるという問題が追加されるが、同じ人物の2つの異なる性質を抽出する場合には生じない問題である。さらに、单一の特性の抽出と同様に、合命令による複数の特性の抽出は、新しい物体と特性の対に対してよく汎化することがわかった(詳細は5.2節組み合わせ汎化で述べる)。最後に、ネットワークのアブレーション研究を行うことで、複合TD命令への汎化は、交差流横方向の接続に依存することがわかった。横方向の結合がない状態で同じ実験(单一命令で訓練した後、複合命令を適用してすべての特性をまとめて読み出す)を繰り返すと、精度が96%から84%に低下する。また、2つの方向の横方向の接続のうち一方に限定すると、TD流からBU流への接続よりも、BU流からTD流への接続がより大きな役割を果たすことがわかった。

3.3 関係

情景で注目される関係には、人物間、人物と物体間、物体間の相互作用がある。図10aに示す人物の相互作用の例は、「対面」関係であり、Actorsデータセットから取り出した、通常近接した位置で向かい合う2人の間の関係である。我々は、<facing, person-1>という形

式の命令を用いて、`faceing` を抽出するように訓練した。Actors セットのネットワークは、person-1 と向かい合っている人物をセグメント化し、分類する（または、person-1 と向かい合っている人物がいないことを示す）ように学習された。上述したように（分類と分割について）、物体は、TD ベクトルとして、または補助入力画像を使用して、TD 命令で指定することができる。画像データセットでは、参照人物（上記の person-1）は、入力画像の追加チャンネルとして与えられた補助分割地図を使用してネットワークに供給された。出力は、TD パスの最後に生成された分割地図と、それに続く BU パスの最後に生成された分類であった（補足 S4 The Actors データセット）。図 10b は、画像内の人間間の「接触」関係を抽出するために同じネットワークを学習した例である。追加の関係と、それらが情景分析中にどのように使用されるかは、4.1 節 情景記述の抽出で詳しく説明する。

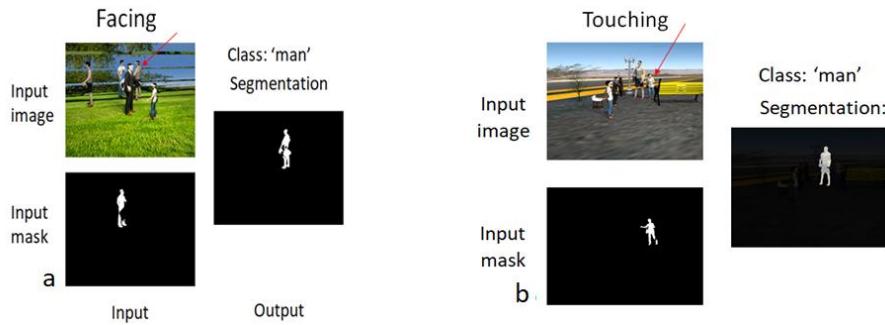


図 10：交流する人々

- a: <対面、人-1> という命令を用いて、「対面」という関係を抽出する。赤い矢印は奥にいる対面している人を指している。参照人物 (person-1) は入力画像とともに分割地図 (入力マスク) によって与えられ、出力は対象人物の分類と分割である。
- b: 「接触」関係。赤の矢印は相互作用している人物を指す。

これらの例や他の例における関係抽出は、入力を与え、関係を計算するという代替的な方法で訓練することもできる。相互作用の参加者の1人を提供し、2人目を見つける代わりに、代替の、より標準的なモードは、2人の人物、x と y と関係 R (2つの命令引数を使用) でモデルを訓練し、 $R(x,y)$ が成り立つかどうかを決定する。例を図 11 (Facing(x,y)) に示す。このネットワークは、Facing(person-1, person-2) という形式の命令で学習された。入力画像とともに提供された分割地図を用いて、2人の入力人物が提供された。出力は「対面」か「非対面」の2値選択である。別の例として、Occlude(x,y) という関係を補足 S6 空間関係に示す。TD 関係命令を使用する2つの形式の使用については、小節「参照命令」でさらに後述する。

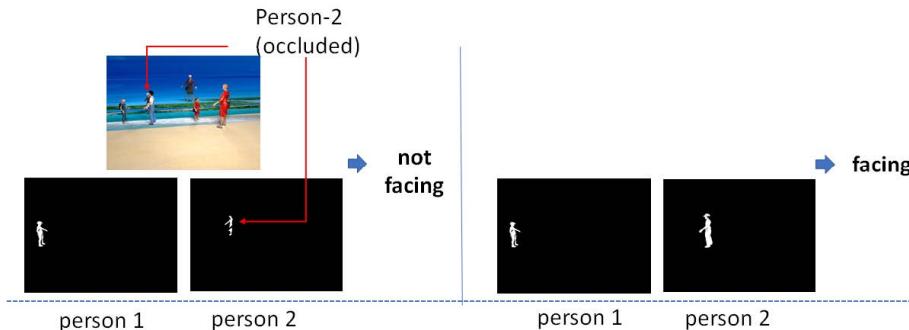


図 11: Facing(x,y) このモデルは、画像中の2人の人物の間の関係 Facing(person-1, person-2) を計算する。入力画像は、2人の人物の2つの分割地図とともに与えられる（下段）。左側の選択された人物の場合、出力は「向かい合っていない」となる。人物 2（赤い矢印で示す）は非常に隠蔽されていることに注意。右側の選択された人物については、「対面」という関係が成り立つ。

3.3.1 人間と物体の相互作用

人間と物体の相互作用の例を、関係 hold の図 12 (holding) に示す。このモデルは、保持された物体（ハンマー、手・持つ標識、バッグ、ラケットなど）を分類し、分割地図を生成する。



図 12: 保持 同じ情景から 2 つの「保持」関係を抽出。上段は、入力、情景画像、および関係の参照人物のガイドィング・マスクを示す。下段は、保持された物体の分割地図を示す。

3.3.2 物体の関係

図 13a は、物体とその部品の間の関係 part-of の抽出を示す。CUB200 データセット (Wah+2011) は、200 種の鳥の画像を、頭、目、冠、喉、脚など 15 の鳥の部位と属性とともに提供する、きめ細かな認識データセットである。このデータセットでは、部位は、見える部位ごとに中心点によって注釈されている。

この場合、BU-TD ネットは、部品の位置 (TD 流の下部にあるヒートマップ) と部品の色 (BU2 パスの出力) を生成するように学習された。異なる部品は異なる課題として扱われ、それぞれが TD 命令のワンホットベクトルによって指定された。BU-TD モデルによる正しい特性の予測精度は 80.9% であり、各課題に個別のネットワークを学習させた場合の 74.3% と比較した (詳細は Levi&Ullman2020 参照)。

関心のある部分を選択する能力は、複雑な情景を処理する上で重要である。情景処理に対するいくつかのアプローチでは、初期段階として、潜在的に関心のある、認識可能なすべての構成要素の識別 (分類とセグメンテーション) が行われる (Redmon+2016, He+2017, Yang+2018)。しかし、複雑な情景に含まれるすべての物体の部位や下位部位の数が多いため、このようなアプローチは実行不可能であり、したがって、複数のレベルで部位を選択する能力が不可欠になる。下の図 13b は、2 つの物体間の「オン」関係の抽出を示している。この情景には、別の「参照」物体の上に置かれた物体がある。TD 命令は、参照物体地図を持つ On 関係として与えられ、その結果は、トップ物体のクラスとセグメンテーション地図となる。他の物体関係には空間関係があり、これは人と物体の両方に適用される。

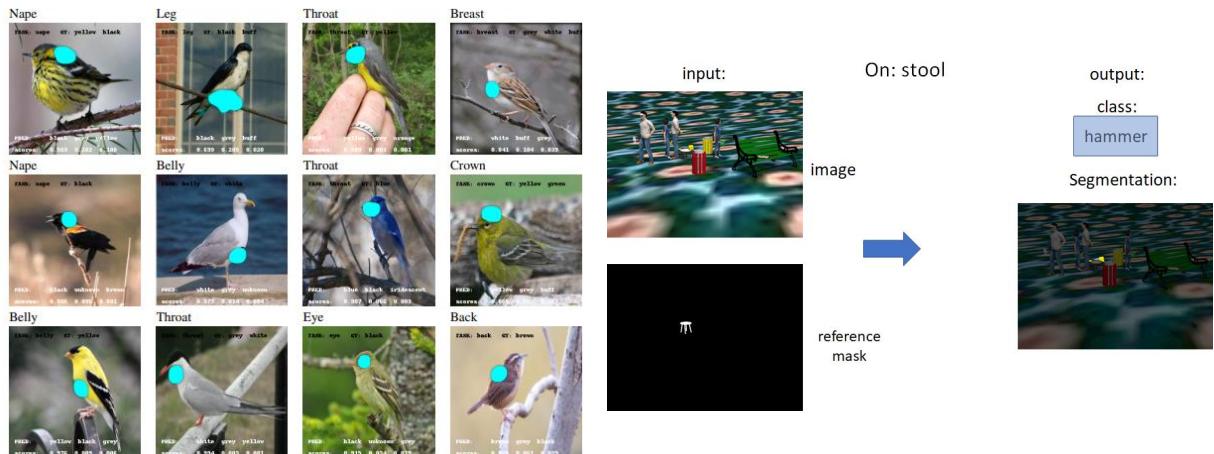


図 13 : 物体の関係

- a. 物体と部品の関係：部品の位置を特定し、その特性を抽出する。画像は CUB200 のもので、200 種の鳥と 15 個の部品が注釈されている。単一のネットワークが使用され、各部位は TD 命令を使用して検出される。部位の位置の予測は TD 流の下部で生成される。
- b. 物体間の関係 on。画像は、テーブル、ベンチ、スツール、椅子の上に置かれた道具のような、情景に含まれる物体間の Actors データセットから得られる。Actors データセットにおける関係抽出の詳細は、4.1 節 (情景記述の抽出) および補足 S4 Actors データセットに記述されている。

3.3.3 空間関係

図 14 (Spatial-2D) は、right-of, left-of, above, below といった空間関係の抽出を示している。これは、EMNIST データセット (Cohen+2017)、固体物体の CLEVR セット (Johnson+2017)、Actor セットからの画像からの例を示している。最初の BU パス (BU1) は、情景内の物体の分類 (つまり、どの物体クラスが存在するか) について学習され、TD 命令は、画像内の選択された物体を分類またはセグメント化することであつ

た。ほとんどの場合、命令の物体はそのクラスによって指定された(このクラスの物体が画像内に1つ存在すると仮定)が、参照物体を指定する他の方法も使用された。その関係を満たす近傍物体が情景内に存在しない場合、「no-neighbor」が返される。命令中の関係は記号的なワンホットの形で表現され、同様に参照物体(引数)のクラスはワンホットベクトルで表現された。異なるドメインのネットワークは別々であり、各ネットワークは <right-of, cylinder> の<above, w> のように、異なる関係と物体の複数課題で学習された。詳細は補足 S6(空間関係)にある。

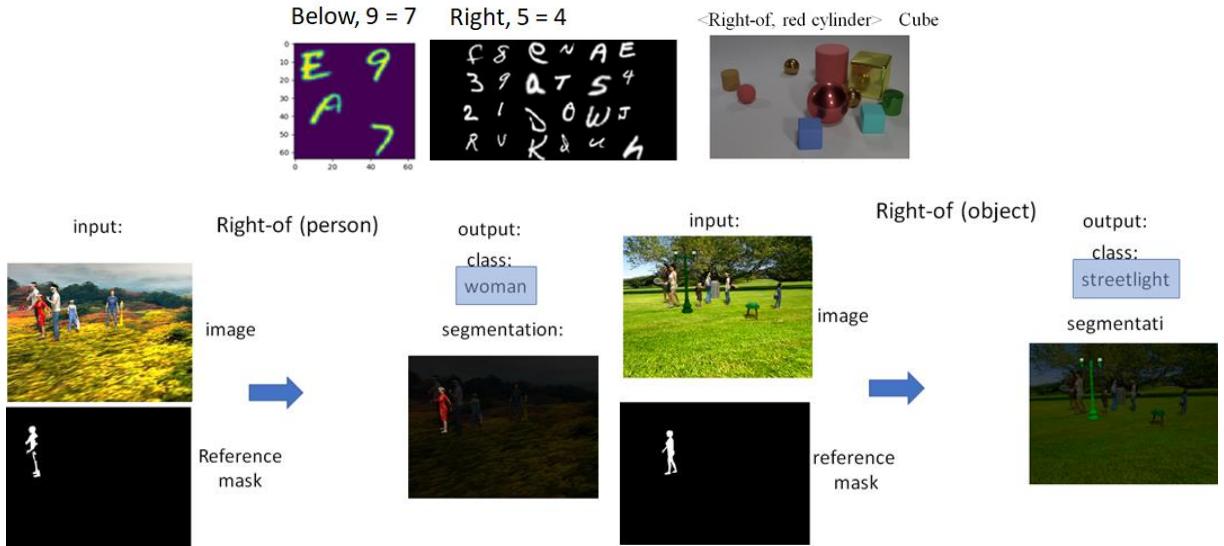


図 14: Spatial-2D 右、左、上、下の関係を学習する。EMNIST 画像(上、左、中央)、CLEVR(右上)、Actors データセット(下段)から得られた例。Actors の例では、対象が人物である場合(左)と、対象が物体である場合(右)の right-of を示している。

下の図 15(奥行き)は、奥行きの空間関係を含んでおり、物体が別の物体に覆い隠されたり、別の物体よりも遠くにあることがある。図 15a では、BU-TD モデルは、隠蔽または隠蔽された人物のいずれかをセグメント化し、分類するように学習された。画像内の隠蔽は最大でも 1 つであると仮定したが、この手順は複数の隠蔽にも拡張できる。図 15b では、モデルを Actors データセットで訓練し、情景内の人物を深度次数に基づいてセグメンテーションし分類した。この課題の入力は、情景と補助的なセグメンテーション地図である。画像内の参照人物 'person-1' のセグメンテーションされた領域が与えられると、命令 <behind, object> は、画像内の次の人物をカメラからの距離の観点から分類し、セグメンテーションする。その他の詳細については、補足 S6(空間関係)と補足 S4(Actors データセット)に記述されている。

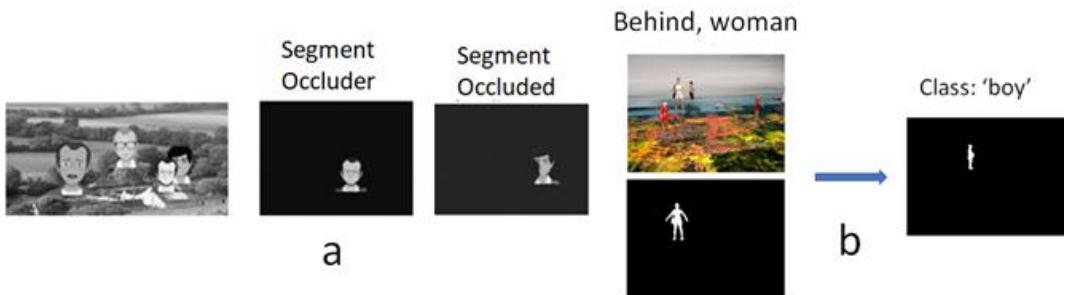


図 15: 空間-奥行き。

- a. 課題は人物のセグメンテーションである。
- b. 他の人物よりも遠くにいるという意味での、後ろという関係。そのセグメンテーション地図によって基準となる人物が与えられると、<behind, object> 命令は、人物-1 に対してカメラからの距離の点で次に位置する画像内の人物をセグメンテーションし、分類する。

3.3.4 高次構成の抽出

左、右、上、下の基本的な命令を用いて、BU-TD スキームは、基本的な命令の適切な系列を適用することにより、追加訓練なしで、高次の空間構成を抽出するために使用することができる。例えば、図 16(EMNIST-structure)を見ると、右側のターゲット構成(「ターゲット」とラベル付けされている)は、文字 W とその左のどこかにある 2、そしてその真上にある 5 から構成されている。この構成テストでは、複数の EMNIST 文字からなる心像性を用いた。課題は、左側の文字配列から、ターゲットとなる構造(真上に 5、左に 2 を持つ W)を見つけることである。

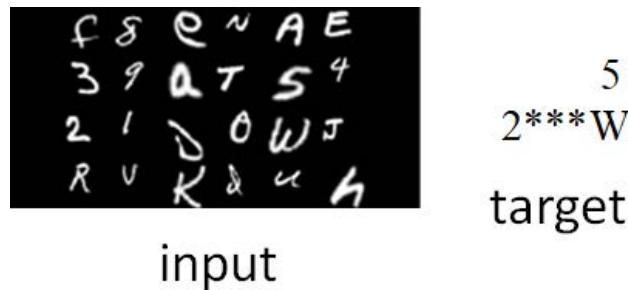


図 16: EMNIST 構造

EMNIST データセットでは、最初の BU パスは画像に存在するすべての文字を生成するように訓練されている。次に、例えば `<above, w>` から始まる一連の TD 命令を適用することができる。もし答えが `5` であれば、次に `Left-of` 命令系列を、`2` に達するか(つまりターゲット構造が見つかった)、行末に達するか(つまりその構造が画像に存在しないことを意味する)まで適用する。これは、4 節(BU-TD 系列を使った情景構造の抽出)でさらに説明する、TD 命令で構成されるプログラムを用いた構造記述の抽出に関する一般的な話題の単純な例である。

3.3.5 位置による汎化

空間的関係を抽出する学習は、ディープネットワーク学習の基本的な限界を明らかにする。例えば、EMNIST データセットの 0~9 の数字で空間関係を学習しても、A~Z の文字に汎化することはできない。対照的に、人間にとって `right-of` のような関係は抽象的な関係であり、特定の物体クラスに依存しない。我々は、单一の命令ではなく、適切な短い一連の TD 命令を使用することで、これらの関係を使用する際の広範な汎化につながることを発見した。この一般化の基本的な考え方とは、空間関係とは物体そのものではなく、物体の位置間の関係であるということである。直接的な命令の代わりに、例えば `right-of` という関係は、以下のように 3 つの命令系列を用いて学習された。画像中の数字 `n` について `right-of(n)` を抽出するには、まず、その数字の位置 `q1(x,y 座標)` を抽出する。第二に、`n` の右隣の予測位置 `q2=(x',y')` を求める。第三に、位置 `q2` にある文字を分類する。課題 1 と課題 3 の画像入力は数字と文字であったが、課題 2 では文字への汎化をテストするため、学習は数字画像に限定された。さらなるテストとして、EMNIST 文字に加えて、複数の追加图形(動物、記号、Omniglot 文字 Lake+2011)を認識する訓練を行ったが、空間関係の訓練中に追加物体が使われることはなかった。この訓練の根拠は、課題 1 と 3、つまり物体の位置を抽出したり、与えられた場所で物体を分類したりする課題は、空間関係とは無関係な一般的な課題であるため、より大きなセットで訓練されたからである。しかし、ステップ 2 の空間関係は、限られた画像セットで学習された。汎化をテストするために、テスト画像には文字や、拡張セットからの形状を含めることができた(図 17 関係汎化のように)。その結果、新規物体に対する精度は、訓練されたクラスの精度と同程度であった。新規の対象形状でテストされた課題 2 では、精度は高いままであった(97 %)。3 つの段階を順次使用した総合的な精度は、見慣れた形状と新規の形状の両方で 88 % であった(Netanyahu2018)。関連するアプローチとして、ニューロンモジュールネットワークを用いたものがあるが、一般性は低いので、(Bahdanau+2019) を参照されたい。

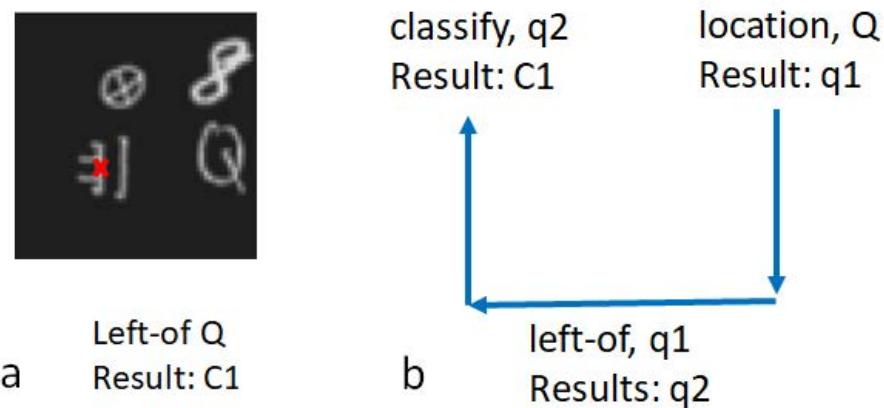


図 17：関係の汎化

- a. 課題は関係 `<left-of, Q>` を計算することであり、答えは `Q` の左にある形状である。
- b. この課題は 3 つの TD 命令によって達成される。最初の `<location, Q>` は、文字 `Q` の位置(画像座標 `q1=(x1,y1)`)を生成する。次に、`<left-of, q1>` は、`q1` の左側にある物体の予測位置 `q2=(x2,y2)` を生成する。最後に、`<classify-location, q2>` は指定された位置 `q2` にある物体を分類する。系列学習は `C1` のような新しい形状にも一般化する。空間関係 `left-of` は形状ではなく、位置間で学習される。

場所を介した同様の汎化は、単純な空間関係以外の他の関係にも適用できる。例えば、人間は、人と動物の間の「乗馬-動物」関係を、既知のクラスから新しいクラスへと汎化することができる(図 18 乗馬動物)。上記のアプローチを適用すると、対象となる動物の位置を特定

し、それを分類するという中間段階を経て、再び関係が成立することになり、乗馬関係では見たことのない動物が分類されることになる。予備的なテスト (Netanyahu2018, Levi2021, Krishna+2018 も参照) では、この汎化様式の実現可能性が支持されている。

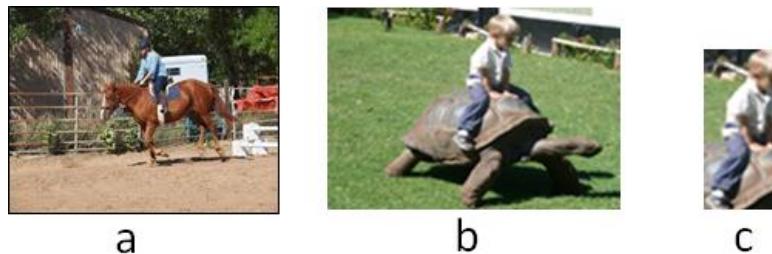


図 18: 乗馬する動物 人間は「乗馬」を (a) のような見慣れた例から (b) のような新しい例まで一般化することができます。関係そのものは、必ずしも対象物を認識しなくとも知覚できることが多い。

3.3.6 指示の参照

関係の抽出に関して注目すべき性質は、関係 R を抽出する指示には2つの形式があるということである。1つ目は、関係計算でよく使われる形式(例えば Lu+2016, Santoro+2017, Yang+2018)で、2つの物体 (x, y) を指定し、関係 $R(x, y)$ が成り立つかどうかを判断するモデルを学習させる。 $R(x, ?y)$ では、1つの物体 x のみが指定され、関係課題は、 x と必要な関係を満たす物体 y が画像中に存在する場合、それを特定することである。前節で、推論命令の例をいくつか取り上げた。例えば、 $\text{Right-of}(x, ?y)$ は、物体 x の右隣を見つける課題である。同様に、 $\text{Occlude}(p1, ?p2)$ は、 $p1$ に隠されている人物を見つけることであり、 $\text{Facing}(p1, ?p2)$ は、 $p1$ に面している人物 $p2$ を見つけることである。

参照表現における関係は、参照物体 x との関係 R を満たす画像内の特定の物体を識別するために使用される。この関係の使い方は、自然言語における「参照表現」の概念に似ている。この概念は、皮質計算における一般的な基本操作として (Valiant1994, 2005) を拡張して (Papadimitriou&Vempala2015) によって提案された「予測結合」(pjoin) の概念とも関連している。参照表現は、言語と視覚の組み合わせでも研究された。その目的は、自然言語の参照表現に基づいて画像内の物体を識別すること、あるいは画像内の与えられた物体を一意に識別する言語表現を生成することである (Mitchell+2013, Mao+2016, Luo&Shakhnarovich2017, Liu+2019)。(Krishna+2018) は、視覚における「参照関係」という用語を、特定の型の参照表現として使用した。そこでは、他の実体との所定の関係に基づいて、画像内の特定の物体を識別することが目標とされる。例えば「帽子をかぶっている少年」は、同じ画像内に存在する他の少年から特定の少年を区別することができる。

我々の使用方法も同様であるが、言語に依存しない。BU-TD ネットワークは「参照命令」と呼ばれるもので、TD 命令によって与えられる関係と参照物体に基づいて、ターゲット物体の位置を特定する。

さらに 4 節で述べるように、関係 R と物体 x を用いて、 $R(x, y)$ を満たす物体 y を探し出す参照指示は、BU-TD モデルによる情景の構造記述の抽出に有用である。簡単に説明すると、情景解釈の過程において、ある物体 x から、 x と特定の関係を満たす新しい物体 y に移動することが有用な場合が多いからである。 x に基づいて y の位置を特定する際、 y を特定することなく、物体 x に基づいて、物体 x が関係 $R(x, y)$ に参加しているかどうかを知ることができる場合が多い。例えば、上図 18c では、少年を中心とする部分画像は、「乗馬」という関係を示唆するのに十分である。同様に、人間の情景解釈(補足 S1 心理物理学的時間軸研究)においても、行為者のみに基づいて行為が示唆され、行為の対象が後から特定されることが多いことがわかった。例えば、図 19a では、「金髪の女性」と「飲酒」が最初に認識され、「グラス」が認識されるのは 300 ミリ秒後である。図 19b では、「女の子」、「持っている」、「注いでいる」がこの順番で認識され、「漏斗」と「ボトル」はかなり後になってから認識される(詳細は補足 S1 参照)。このような進行は、次に適用する TD 指示を示唆することで、情景の解釈に利用することができる。例えば、行動認識では、可能性の高い行動を最初に認識し、その後に行動対象を見つける TD 指示を行うことができる。



図 19: 参照-動作。人が物を使って動作をしている画像では、人と動作が物より先に認識されることが多い。

- a. 「金髪の女性」と「飲んでいる」は最初に認識され、「グラス」は 300 ミリ秒後に認識される。
- b. 「女の子」、「持っている」、「注いでいる」の順に認識され、「漏斗」と「瓶」はかなり遅れて認識される(補足 S1 心理物理学的タイムライン研究)。

In extracting structures of interest from the scene (Section 4 (Extracting scene structures using BU-TD sequences)), we found that referring relations play a major role. However, the model also uses non-referring relations discussed above, of the form $R(x,y)$, where the two objects x,y are given as arguments of the instruction, and the model is trained to determine whether or not $R(x,y)$ holds (see Supplementary S6 (Spatial Relations)).

4. BU-TD 系列を用いた情景構造の抽出

BU-TD モデルは、TD 命令によって呼び出されるステップの系列から、複雑な視覚課題を構成することを可能にする。すなわち、(i) 利用可能な TD 命令から幅広い新規課題を構成し、(ii) 与えられた課題を幅広い画像セットに適用する必要がある。各ステップで適用する TD 命令は、一般に、目標、画像、および前のステップで画像から既に抽出された情報に依存する。例えば、与えられた画像内の特定の物体を視覚的に検索する場合、実行すべきステップは、画像内の物体の数とその配置に依存する。その結果、マルチサイクル処理によって複合視覚課題を適用する処理は、画像に視覚プログラム（または「視覚ルーチン」Ullman1984）を適用すると考えることができる。冒頭（背景と目標）で簡単に述べたように、次に適用する TD 命令の選択も非視覚的知識に依存する可能性があるが、この側面は今回の議論の範囲外のままである。画像に視覚ルーチンを適用する主な目的は、シーンに関連し、かつ有益な記述を抽出することである。次節では、TD 命令で構成されるルーチンを用いた、関心のある構造的記述の抽出について説明する。

4.1 情景記述の抽出

完全構造の抽出

図 20（完全な構造）は、対話する人々の情景の完全な構造を抽出する例を示す。「完全な構造」とは、すべての構成要素、それらの特性、および関係を抽出することを意味する。この情景は Actors データセットから生成されたので、画像に見えるすべての構成要素、特性、および関係の集合は完全に既知であり、したがって抽出された構造記述は既知のグランドトゥルースと比較することができる。

完全な構造を抽出する処理は、補足 S7（完全な構造の抽出）に記載されている。簡単に説明すると、次のような段階を経る。第一段階は、画像内の全ての構成要素（人物と情景オブジェクト）を特定する。（情景オブジェクトには、人物が持っている物体や情景オブジェクトの上に置かれているオブジェクトは含まれない。）これは、TD 命令を用いて、それらを順次検出し、セグメント化することによって得られる。この命令は、カメラからの距離の順に、画像内の次の人物または物体をセグメント化し、分類するために使用される。この命令の入力には、累積セグメンテーション地図（入力画像とともに提供される）が含まれ、これにはこれまでに抽出されたすべての成分が含まれる。累積地図は最初は空であるため、この命令の最初の適用では、カメラに最も近い人物または物体が識別される。識別された各成分について、すべてのプロパティが抽出され、最終的な構造表現に追加される（下記参照）。次に、すべての人物や物体とそのプロパティが抽出されるまで、次の成分が識別される。次の段階では、すべての人物-関係人物-物体および物体の関係に対して、画像内のすべての参照関係を抽出する。最後に、非参照関係 $R(x,y)$ （3.3 節「関係」）を、抽出された物体の関連するすべての対に適用することで、関係を追加することができる。完全な構造の抽出の詳細については、補足 S7（完全な構造の抽出）を参照のこと。

BU-TD 処理による構造の抽出は、特定の構造を用いた訓練に依存しない。データセットによって生成された任意の構造を抽出するためには、我々のモデルは、個々の命令、すなわち、モデルによって使用されるすべての物体クラス、特性、および関係の視覚的認識のための、それぞれの命令を用いた訓練を必要とした。また、現在の構造を拡張するために、新たな画像構成要素を識別する（例えば、次の人物を識別する）など、いくつかの追加命令も必要となる。自然画像では、視覚分類子の集合は非常に大きくなるが、この大きな集合を扱うことは、あらゆる構造認識スキームに固有の要件である。この命令セットにより、スキームは訓練された構造に限定されず、あらゆる構造を復元することができる。単純な構造の完全な構造の抽出は、テストや評価の目的には有用であるが、実世界の自然な構造の場合、完全な構造の抽出は、次節で議論するように、しばしば実行不可能であり、通常は不要である。上記の処理によって抽出された構造を記述する構造は、構造グラフ（Santoro+2017, Xu+2017, Yang+2018）と同様のグラフとして考えることができる。グラフ構造は（Battaglia+2018）で説明されているような一般的な形式を持つことができ、物体プロパティや関係だけでなく、大域的な構造プロパティや、「温かい抱擁」のような属性を関連付けることができる関係も含む。計算モデルでグラフ構造を表現する一般的な方法は、構造のノードとエッジを記述する局所的記述のコレクション、通常はトリプレットである。例えば、プロパティのトリプレットは $(item, property, value)$ という形式を持つ。ここで $item-1$ または $item-2$ は、抽出処理中に生成される識別番号（i.d.）である。

構造に加えて、構造表現には、画像内の抽出された構成要素（人物と物体）の「グラウンドィング」が含まれる。このグラウンドィングとは、構造記述の構成要素から、その i.d. によって識別される、当該構成要素を含む画像領域へ参照し返す能力を意味する。このグラウンドィングは、構造記述において $(component-id, segmentation-map)$ という形式の対によって提供され、構造解釈処理において画像から抽出されたセグメンテーション地図と各成分を関連付ける。計算スキームと同様に、人間の知覚系もまた、外部構造の何らかの構造表現を作成するという問題に直面している。興味深く、まだ未解決の問題は、一般的に人間の脳におけるそのような構造の形態と神経表現、特に構造記述である（Smolensky1990, Thomas_McCoy+2019）。

D:\docs\Downloads\pp_full_structure_example_no_tool_size.png

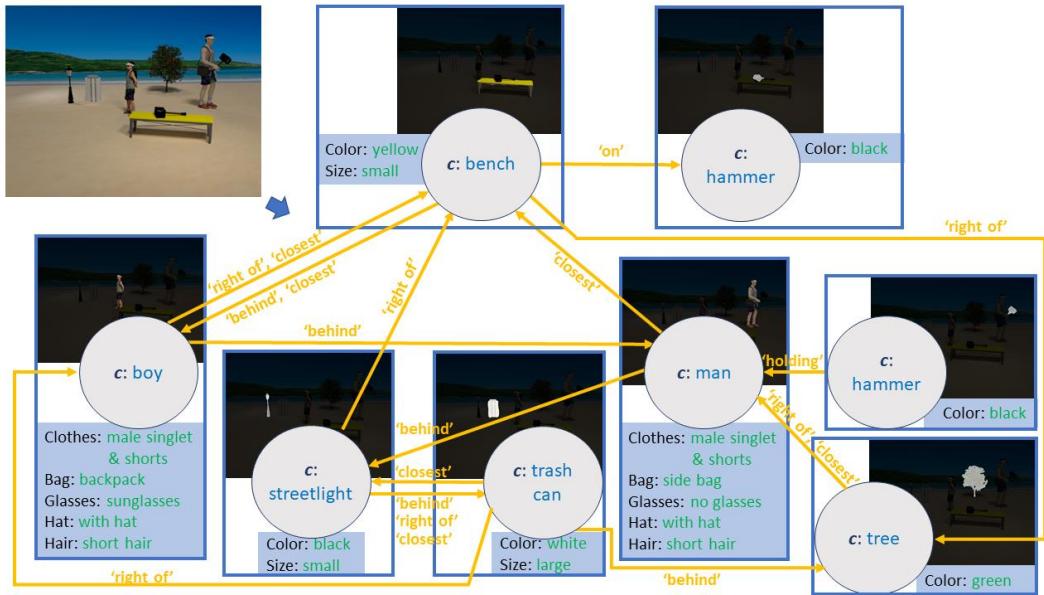


図 20: 完全な情景構造。入力画像(左上)には人物、情景物体、保持物体が含まれる。抽出された構造には、情景構成要素、それらのセグメンテーション地図、特性、情景構成要素間の関係が含まれる。この構造は、モデルによって自動的に選択された TD 命令を順次適用することによって抽出された。

ガイド付き構造抽出

自然な情景において、潜在的に関心のある全ての情景構成要素を、その全ての特性と関係とともに含む完全な「シーニングラフ」を抽出することは、一般的に不可能である。例えば、複数の人物と多くの物体がある部屋の情景において、我々は特定の側面、例えは特定の人物が持っているグラスが空なのか満杯なのかに興味があるかもしれない。もちろん、これは多くの可能性のひとつに過ぎず、例えは、画像から完全な情景グラフを抽出したり(Yang+2018)、画像内のすべての物体を、そのパートやサブパートとともに複数のレベルで分類・セグメント化したり(Redmon+2016, He+2017)するような、ガイドのない方法でそれらすべてに対処することは考えにくい。その代わりに、情景の構造の関連部分を素早く抽出することが目標となる。抽出すべき関連部分は知覚者の目標に依存するため、情景に適用される視覚プログラムは目標に依存すべきである。BU-TD 系列を目標指向的に情景に適用した例を図 21(ガイド付き情景解釈 Guided scene interpretation)に示す。この例の目標は、少女と向かい合っている女性が手にしている物体を特定することである。下の図 21a は入力画像である。処理を導く目標は、(b) に示す、画像から見つけるべき標的構造によって表される: それは、少女と向かい合っている女性が何を持っているかを特定することであり、物体とその心像性を特定する。(c) から (e) の行は、4.2 節(視覚ルーチンの構成: 次の TD 命令の選択)で説明したアルゴリズムによって自動的に生成された、ゴールに到達するために使用される TD 命令の系列を示す。(d) の行は、命令を用いて画像から抽出された人物の系列を示している。この命令は、カメラからの距離順に抽出された画像内の次の人物(または情景の物体)をセグメント化し、分類するために使用される。この命令の入力には、入力画像と一緒に提供された累積セグメンテーション地図が含まれる(c 行目)。

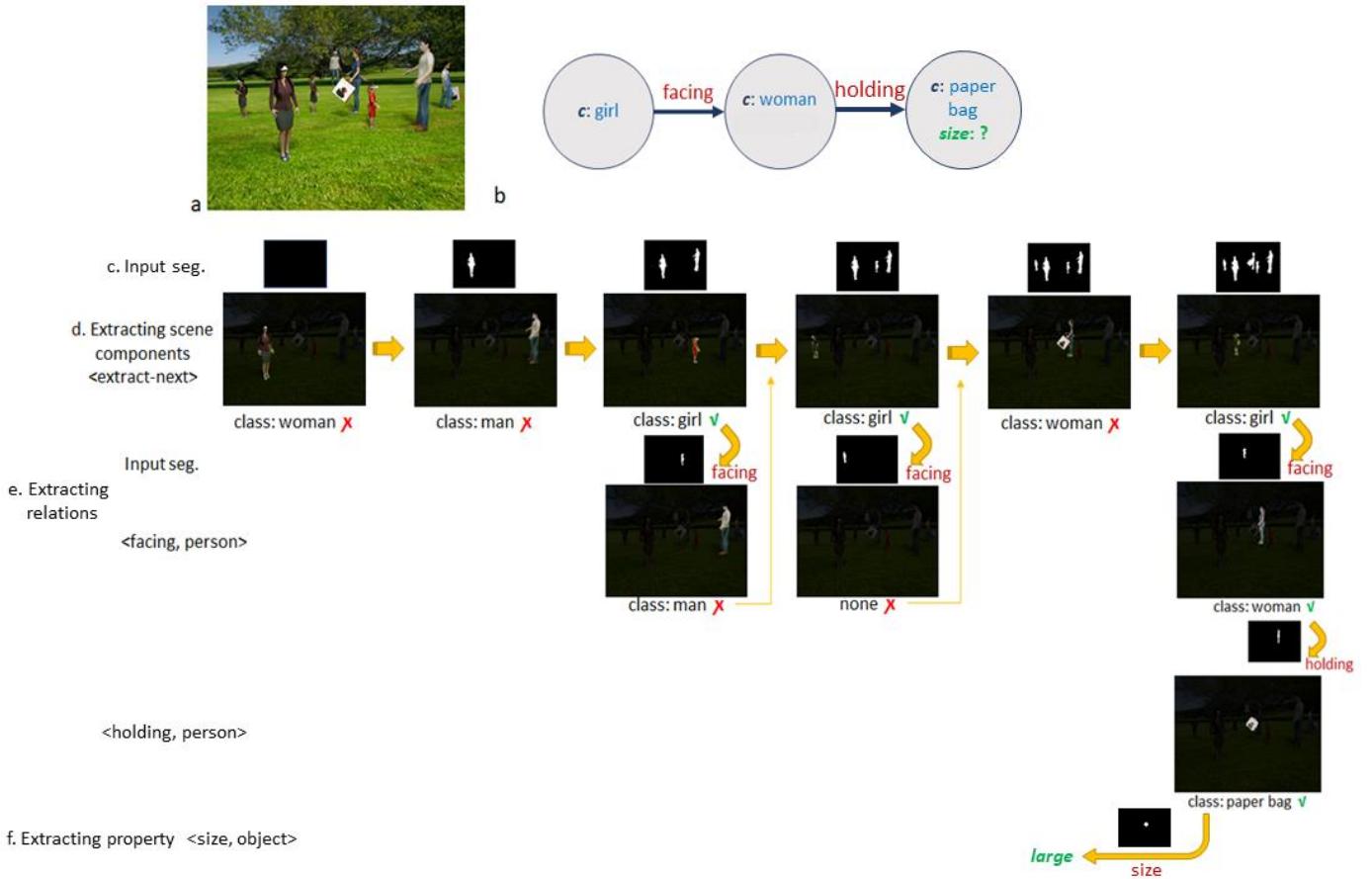


図 21: ガイド付き情景解釈。目標は、女の子と向かい合って紙袋を持っている女性を見つけ、その袋の大きさを特定すること。

- a: 入力画像。
- b: ターゲット構造。(c)から(f)の行は、アルゴリズムによって自動的に生成される、ゴールに到達するために使用される TD 命令系列を示す。
- c,d. TD 命令を用いた、情景内の人物の逐次抽出。これは、カメラからの距離の観点から、その次の情景構成要素を分類し、セグメンテーションする；c は、これまでにセグメンテーションされたすべての人物の累積セグメンテーション地図を示す。
- e: 参照人物の入力セグメンテーション地図(上、小画像)と出力(下、大画像)を示す。(d)による連続的な抽出で最初の女の子(d の 3 番目の画像)が特定されると、次の TD 命令で <facing, person> がテストされ、man が生成される。d,e の繰り返しは、女の子と対面している女性が見つかるまで続けられる(d, e の最後の列)。該当する女性を見つけた後、次の命令 <holding, person> は「紙袋」を取り出し、必要なターゲット構造を完成させる。(f)では、<size, bag> という命令によって、バッグの大きさが検索される。
- g. 解釈処理過程で見つかった関連する情景構成要素(少女、女性、バッグ)を示す最終結果を示す。最終的な出力には、画像から抽出された構成の構造化された記述(構成要素、特性、関係、およびセグメンテーションマップを使用した画像内の各構成要素の接地)が含まれる。

画像から最初に抽出された人物は女性と識別される。(b)の構造に対するガイド付き探索は女の子を探しているので、命令は女の子が見つかるまで繰り返し使われる。ターゲットグラフに統いて、ガイドアルゴリズムは次に <facing, person> 命令を生成し、先ほど発見された少女が他の人と向かい合っているかどうかをテストする。facing 命令の引数 ‘person’ は、女の子のセグメンテーション地図によって与えられ、入力画像と一緒にネットワークに供給される。‘next’ と ‘facing’ の一連の命令によって、最終的に少女と向かい合っている女性が特定される。<holding, person> 命令は、女性が紙袋を持っていることを確認し、紙袋を取り出す。(f)では <size, object> 命令を与え、バッグの大きさを取り出すことでゴールに到達する。

画像内で特定された構造は(g)に示されており、画像内の構成要素(少女、女性、紙袋)の位置を示している。解釈処理の出力には、画像から抽出された構成要素、関係、特性の構造記述も含まれ、それらの関連情報(特性、関係、セグメンテーション地図)を含む構成要素の配列として配置される。

TD 命令の系列を使用することで、この処理は画像の部分的な分析を行い、関心のある目標を得ることができた。目標を達成するためには、適切な命令系列(通常は一意ではない)を適用しなければならない。したがって、次節で議論する基本的な疑問は、各段階で適切な TD 命令がどのように選択されるかということである。

4.2 視覚ルーティンの構成：次の TD 命令を選択する

本節では、ガイド付き情景解釈を行うための、次の TD 命令を選択するアルゴリズムについて説明する。情景解析は特定の目標なしに進めることもできる。例えば、特定の目標を意識せずに新しい画像を見て、それでも有用な情報を抽出することができる。このガイドなしの抽出については後述する(4節の最後)。情景解析のガイドに使われるゴールは、画像から特定すべき情景構造として与えられ、必要な構造を見つけ抽出するための TD 命令からなる系列を自動的に生成すると仮定する。自然な状況では、目標構造は明示的に与えられないことがあり、構築する必要がある。例えば、ゴールは自然言語で記述された質問によって外部から与えられることがあり、目的の構造を構築するために最初の構文解析段階が必要となる。後述するアルゴリズムでは、画像に目的の構造が含まれていないと結論付けることもできるし、目標が1つ以上の構造を見つける必要がある場合には、目的の構造の複数の(またはすべての)実体を特定することもできる。ゴール指向の情景解析の例として、視覚的質問応答(VQA)の課題がある。この課題では、画像と特定の質問が与えられ、画像に基づいて適切な回答を生成することがゴールとなる(Kafle&Kanan2017)。VQA課題では、質問(クエリ)は外部から投げかけられるが、知覚者の現在の目標に応じて、クエリを知覚者が内部的に生成することもできる。

次に、ターゲット構造と画像を入力とし、ターゲットに導かれた TD 命令からなるプログラムを生成する一般的なアルゴリズムについて述べる。この処理については(Vatashsky&Ullman2020)に詳述されている。VQA アプリケーションでは、アルゴリズムはまず、自然言語で与えられた質問をグラフ表現に変換し、画像から発見すべき情景構造を指定する。この段階は、自然言語で投げかけられた質問を処理することであり、ここでは関係ない。ターゲット構造が与えられた場合、我々は、ターゲットに導かれた適切な TD 命令列(一意でない場合もある)を生成する問題に焦点を当てる。ゴールに到達するための TD 命令系列を生成するための基本的なアプローチは、図 22(ターゲット-グラフ)の単純な例を用いて模式的に示される。左のグラフは、黄色いバスの右側に赤い車を識別するという目標を表している。ゴールに到達し、質問に答えるために、考えられる一連の手順は、バスを見つけ、その黄色の色を確認し、その右側にある車を見つけ、その色を識別することである。例えば、車が赤でなかったり、バスが黄色でなかったりした場合は、情景解析処理を拡張し、別の車や別のバスを探す必要があることに注意されたい。したがって、アルゴリズムによって生成される TD 命令の系列は、ターゲット構造によって決定される固定された系列ではなく、ゴールと画像に依存する進化するプログラムである。図 22 の右側のグラフ(Target-graph)は、クエリされた構造の抽象的な形を示す。これは、特定の特性や関係に依存しないという意味で抽象的である。これは、クラス c1 のオブジェクトと特性 p1 を持ち、クラス c2 の第 2 の物体との関係 R にあり、我々が識別する必要のある特性 p2 を持つ。

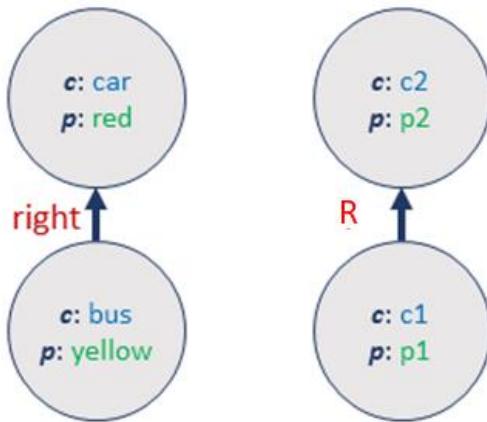


図 22: ターゲットグラフ

- 左: 画像内で識別すべきターゲット構造。
- 右: ターゲット構造の抽象形。

つまり、クラス c1 と特性 p1 の物体を見つけ、関係 R に従ってクラス c2 の2つの物体を特定し、要求された特性 p2 を特定する。大雑把に言えば、適切な TD 命令を選択する手順は、要求された情景構造の「グラフをたどる」ことによって進行し、物体を分類し、そのプロパティを抽出し、関係をたどるなどの命令を選択する。また、論理接続詞や量化詞を含む目標を扱うために、我々が使用する手順は拡張され、例えば、与えられた構造のすべての実体を見つけるために、例えば「テーブルの周りにいるすべての人がスピーカーを見ているか」という目標に対処する。

TD 命令を生成するために使用される手順は、一度に1つのノードを処理する再帰的処理である(部分グラフのルートノードがある場合はそのノードから開始し、そうでない場合は任意のノードから開始する)。この処理は、2つの BU-TD ネットワーク-「拡張ネットワーク」と「精緻化ネットワーク」-の組み合わせによって実行される。拡張ネットワークは、抽出された情景構造に新たな情景構成要素(人物または物体)を追加することで、表現グラフを拡張する。精緻化ネットワークは、TD 命令によって指定された追加プロパティと特定の関係を抽出することによって、既存のグラフを精緻化するために使用される。この2つのネットワークについて以下に簡単に説明するが、より詳細な説明は補足 S7(完全なシーン構造の抽出)に記載する。拡張ネットワークと精緻化ネットワークから構成されるモデルは、以下に説明するように、Actor データセットの情景構造の抽出に使用された。Persons データセットと EMNIST データセットの様々な課題に使用されたモデル(上記の節と以下の5節)は、これらの成分を使用せずに独自のネットワークを使用した。

拡張ネットワーク: 参照物体との関係に基づいて人物や物体を見つけ、それをセグメンテーションし、そのクラスを認識するように学習される。例えば、展開ネットワークへの TD 命令は <facing, person-1> とすることができ、ネットワークは、person-1 向いている画像の人物を見つけ、その分類とセグメンテーションを生成する。引数(person-1)は、入力画像と一緒に供給されるセグメンテーション地図によってネットに与えられる。ネットワークは、命令を使って、カメラからの距離に基づいて画像から新しい物体や人物を抽出することができる。命令の引数には、画像からこれまでに抽出された人物や物体を含むセグメンテーション地図が与えられる。精緻化ネットワーク: TD 命令に従って、特定の構成要素の選択された特性を認識するように学習される。同じネットワークは、R(x,y) 形式の関係を認識するためにも学習される。R(x,y) は、入力が 2 つの物体と関係で構成され、出力が関係の値(多くの場合、単に真/偽の値である)である。入力の引数は、関係に参加する 2 つの項目を表す 2 つのセグメンテーション地図で与えられる。前述したように、物体間の参照関係(3.3 節 関係で説明)は、進化する表現を拡張するために使用されることが多いため、拡張ネットワークによって抽出される。

次に、ターゲット構造を抽出するためにモデルを誘導する手順を概説する。完全なアルゴリズムのより詳細な説明は、補足 S9(次の TD 命令の選択)に記載されている。この手順では、展開ネットワークを用いて人物と物体を抽出する。人物と情景物体(ベンチや街灯など)は、命令を使用して 1 つずつ抽出される。アルゴリズムはターゲット構造のノードに沿って進み、ターゲットグラフで記述された構造に対応するようなプロパティと関係を持つ人物や物体を画像から抽出するようモデルを導く。抽出された人物のクラスが、グラフの現在のノードが必要とするクラス(例えば「少女」)に対応する場合、処理は次に、そのノードのプロパティと関係の観点から、そのノードの要件をチェックする。関連する特性は、適切な TD 特性命令を使用して抽出および検証され、関連する関係は、グラフ内の関係によって必要とされる人物および物体を検索する展開ネットワークを使用して抽出および検証される(たとえば facing)。空間関係命令('right-of' や 'behind' など)は、人物や情景物体を抽出し、「holding」や 'on' などの関係は、人物によって保持されたり、他の物体の上に置かれたりする道具や物体を抽出するために使用される。精緻化ネットワークは、2 つの引数によって定義される関係を担当する(3.3 節 関係)。ネットワークによるプロパティの抽出は、ターゲットグラフ内のプロパティを検証する(例えば、赤い物体を探す)か、プロパティを検索する(例えば、テーブル上の物体の色を見つける)どちらかに使用される。検索された情報は、各エントリーが 1 つの物体／人物のデータを表す配列に保存され、更新される。要件を満たす検索された物体は、ターゲット・グラフ内の対応する物体と対になり、後続のテストが正しい物体に適用されるようになる。すべてのノードのテストが完了すると、深さ優先(DFS)探索によって決定された次のノードに対して、この処理が繰り返される。再帰的な処理は、ターゲットグラフが画像に完全に接地された時点で終了するか、あるいは、チェックすべき代替案がなくなった時点で終了する。ガイドされた処理の詳細は、補足 S9(次の TD 命令の選択)に記載されている。

ガイディング・アルゴリズムは本質的に一般的であり、適切な TD 命令系列を自動的に適用することで、新規の情景において関心のある新規の構造を抽出することができる。この視覚的ルーチンを構成し適用する能力は、エンド・ツー・エンドで完全な構造を学習しようとする現在のほとんどのアプローチ(例えば VQA で使用されている)とは異なる。

非誘導情景解析

前述の説明では、完全構造の抽出と、目標構造のガイド付き抽出について述べた。さらに、情景理解には「デフォルト」モードがあり、そこでは特定のゴールもターゲット構造もなく、単に情景を見て、興味のある一般的な情報を抽出する。非ガイドモードでは、特定のゴールによるガイドの代わりに、情景分析は少なくとも部分的には一般的な優先順位によってガイドされ、興味のありそうな情報をできるだけ早く抽出する確率を高めることを提案する。特に、情景分析処理の早い段階で、情景内の人物の存在を抽出し、人物間の相互作用、および人物と物体間の相互作用を発見することは、一般的に有用である。例えば、人物の優先順位を得るために、次の情景成分を抽出するための TD 命令の代わりに、情景物体を無視して、シーン内の次の人物を抽出する命令を追加することができる。現在、モデルをガイドするために、次の TD 命令を選択する処理に一般的な優先順位を組み込む方法を模索している。このような優先順位に加え、ボトムアップの顕著性も画像からの情報抽出の順序に影響を与える可能性がある。

5. 容量と一般化

前節では、BU-TD スキームがどのように興味のある構造を抽出するために利用できるかを示した。本節では、多数の課題を学習し、情景構造を横断して汎化する際に、この形式の課題選択の利点を示す。

5.1 マルチタスクと容量

視覚処理の成功したネットワークモデルは、当初、物体の認識やセグメンテーションなど、单一または少数の課題に焦点を当てていた。最近の研究では、画像内の複数の物体の認識やセグメンテーション、同じ物体の複数の特性、物体間の関係など、同じモデルが多数の異なる課題を実行するように学習された場合に生じる問題を探し始めた。複数の課題に対して学習を行う際に生じる重要な問題は、課題間の相互作用と容量の制限であり、これにより学習が困難になり、結果の精度が低下する可能性がある。この影響は、(Sener&Koltun2018, Zhao+2018, Strezoski+2019) のように、課題数が少ない場合に既に顕著に現れることがあり、大規模な人間類似視覚モデルで扱う課題数が増えるにつれて、ますます深刻になる可能性が高い。そのため、マルチタスクと容量の問題は、その成功にとって極めて重要になる。

マルチタスクの課題に対する最近の一般的なアプローチは、共通のバックボーンの上に各課題専用のブランチを配置した多分岐アーキテクチャを使用して、複数の課題を処理することである(図 23a)。多分岐ネットの訓練には、重み付き損失関数の使用(Chen+2018, Sener&Koltun2018)や、いわゆる変調モジュールの使用(Zhao+2018, Strezoski+2019)など、いくつかの方法が提案されている。分岐アプローチの限界の一つは、いつでも漸進的に追加できる新規課題に対応するのではなく、固定数の課題に対応するように設計されていることである。対照的に、BU-TD モデルは、異なる TD 命令で訓練された单一の共通モデルを使用して、異なる課題に対して訓練される。分岐ネットワークや類似のモデルとは異なり、BU-TD モデルは、可能なすべての課題をまとめて実行するのではなく、各サイクルで選択された課題、

または少数の課題のみを実行する。後述するように、このアプローチの大きな利点は、その埋め込み表現を学習することで、既存のモデルに新しい課題を追加できることであり、既存の課題を損なうことなく新しい課題を追加することが可能になる。

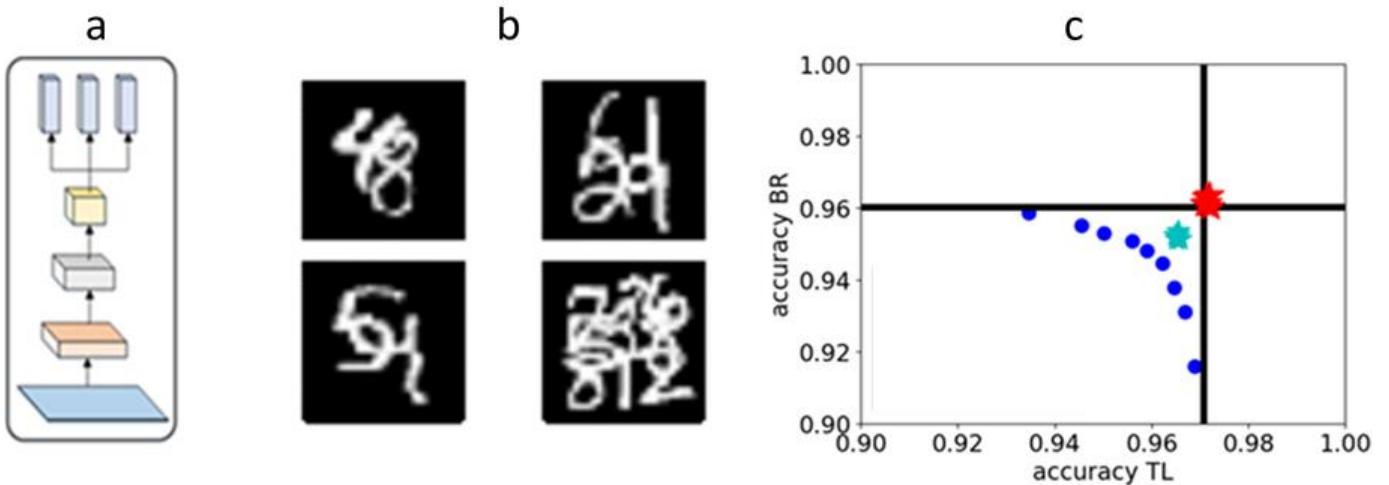


図 23: Multi-MNIST Multi-MNIST に適用されるマルチタスク。

- a: 共通のバックボーンに続いて複数の課題固有の分岐が続く、概略的な分岐アーキテクチャ。
- b: 複数の数字が部分的に重複する Multi-MNIST の例。
- c: 2 桁の数字認識課題の性能。x 軸は左上の数字の認識率、y 軸は右下の数字の認識率である。横棒と縦棒は各課題単体の性能、青点は重み付き損失を用いた分岐ネット、赤星は BU-TD ネット、シアン星はチャネル変調の性能を示す。

3.1節(物体)で簡単に説明した Multi-MNIST データセット (Sabour+2017, Sener&Koltun2018) を用いた図 23 (Multi-MNIST) に簡単な例を示す。(Sener&Koltun2018) で用いた 2 クラス実験では、画像の左上(TL)と右下(BR)の数字を分類する課題であった。同様に 4 クラスと 9 クラスの実験を追加し、各位置の数字を独立に選んだ(図 23)。

テストと結果について簡単に説明する。詳細は (Levi&Ullman2020) に記載されている。テストでは、各場所について 60,000 例の MNIST データセットを使用した。(Sener&Koltun2018) と同様に、テストしたモデルの共通のバックボーンとして LeNet ネットワーク(LeCun+1998) を用いた。我々は、BU-TD モデルをいくつかの代替案と比較した：各課題が独自のネットワークによって実行される単一課題ベースライン、多重目的アプローチ(異なる課題の個々の損失の加重和を最小化する)を持つ分岐ネットワーク、および「チャネル変調」と呼ばれる方法(Zhao+2018)。BU-TD モデルと同様に、チャネル変調は一度に 1 つの課題を実行し、チャネル単位のベクトル変調アーキテクチャを使用して、課題に応じて異なるチャネルの重みを変調する。図 23 (Multi-MNIST) には、2 つの課題の実験における性能がプロットされている。青点は、2 つの課題を異なる損失重みで組み合わせたときの性能を示している。このプロットは、一方の課題の精度を上げると、もう一方の課題の精度が下がるという能力問題を示している。対照的に、赤い星印で示した BU-TD スキームの結果(5 回の実験の平均)は、各課題に完全なネットワークを割り当てた単一課題の場合と比較して、精度の低下は見られず、わずかに優れている。チャネル変調アプローチ(水色の星印)は中間的な結果を示しており、2 つの課題の精度が若干低下している。表1:マルチ性能は、2 課題、4 課題、9 課題を使用した場合の Multi-MNIST 実験の結果をまとめたものである。各行について、5 回の実験に基づく全課題の平均精度を示している。2 列目は標準的な LeNet アーキテクチャのパラメータ数の倍数としてパラメータ数を示している。表中のチャネル変調(拡張)モデルは、パラメータ数を増やしたチャネル変調モデルである。BU-TD 課題選択の使用は、9 つのネットワークを使用する単一課題のベースラインを含む他のアプローチと比較して、有意に良好な結果を達成している(おそらく相関課題のより良い使用による)。課題数のスケーリングは精度のギャップを増大させる。桁数を増やすと、課題数が増えるだけでなく、桁間の重複が増えるため、問題が難しくなることに注意すべきである。

ALG	2 digits Av. Acc	4 digits Av. Acc	9 digits by loc Av. Acc	9 digits by ref Av. Acc
Single task	96.46	94.15	86.62	50.33
task-routing	95.12	92.09	80.52	40.00
channel-modulation	95.87	91.38	76.56	32.69
channel-modulation (extended)	96.30	92.96	79.81	38.57
BU-TD	96.67	94.64	88.07	72.25

表 1: 2, 4, 9 課題の Multi-MNIST 課題におけるマルチ性能精度(5 回繰り返した平均値)。課題数が増えるにつれて、BU-TD と他のモデルとの性能差が大きくなっている。桁数が増えると、桁間の重複が増えるため、各課題の単体の難易度も上がる。

5.1.1 課題選択性

BU-TD スキームでは、TD 命令が特定の課題を選択し、TD 流が次の BU パスを選択された課題の実行に導く。課題間の競合や干渉を避けるために、選択された課題の処理を、競合する非選択課題に比べて選択的にすることが望ましい目標である。BU2 流によって生成される表示の選択性をテストするために、BU2 流の先頭から、選択された課題だけでなく、すべての位置の数字を予測することを試みた。この目的

のため、訓練後、4数字テストにおいて、BU2の最上層に4つの読み出し分岐(各2層)を取り付け、それぞれが1つの位置の数字を予測するように訓練した。4つの課題の予測精度を図24(課題選択性、左パネル)にまとめた。この結果は、選択された課題の場所ブランチでは90%以上の精度を示し、選択されていないすべてのブランチでは偶然レベルに近い精度を示している。BU1パスの先頭から同様の読み出しが行うと、すべての場所で30~40%の範囲の精度が得られる。課題選択性は、選択課題と非選択課題の(平均精度-偶然レベル)の比として定義される選択性指数で測定した。選択性指数は、4課題テストでは26.5、9課題テストでは37.2であった。その他の結果と代替案との比較は、(Levi&Ullman2020)に記載されている。

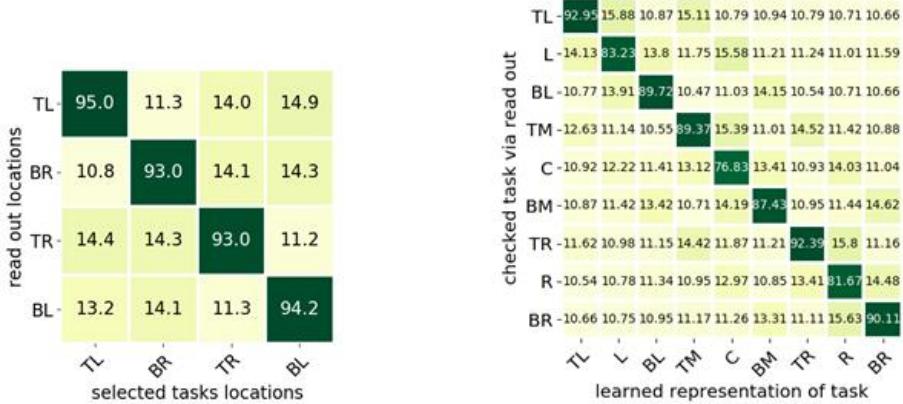


図24: 課題選択性。各プロットの列は選択された課題の位置で、行は読み出し位置(左上から左下/右下までの範囲)である。選択性が高ければ、対角線外の認識は偶然のレベルになる。選択性は4課題(左)と9課題(右)のMulti-MNIST課題について示されている。

BU-TD モデルは、指示された課題に対して高い選択性を示す。例えば、(図4 TD 指示のように)選択された人物の選択された性質を抽出する場合、選択された性質を読み上げる精度は高い(98%)が、選択されていない性質のいずれかを読み上げる精度は偶然のレベルに近い(これは選択された人物の選択されていない性質、選択されていない人物の性質の両方について当てはまる)。EMNIST 文字(3.3 節空間関係)については、選択された文字の左隣または右隣を報告するという追加課題もテストした(6 文字の列)。選択された文字の選択された隣人を読み上げる精度は 95% であったが、非指示の隣人または非指示の隣人のいずれかを読み上げる精度は偶然レベルに近かった。

プログラム修正としての課題選択

課題間の切り替えにおける TD 流の機能は、分岐ネットワークのより一般的な形と考えることができる。分岐ネットワークは、課題が特定の方法で関連しているという事実を利用することで、課題の集合を計算する。課題は、すべての課題に有効な共通部分を使用するネットワークによって実行でき、その後に各課題のための特定の分岐が続く。これは、独立したものではなく、共通部分 T と、それに続く各課題固有の追加部分 δ_i から構成される、異なる課題のためのプログラム T_i の集合に類似している。個々のプログラム (T_i) の代わりに、プログラムの集合 (T, δ_i) は、関連する課題の集合を扱う効率的な方法である。BU-TD モデルにおける TD 命令の使用は、関連する課題の集合を扱う、より一般的な形式である。BU 流は共通部分 T と見なすことができ、トップダウン部分は次に、課題固有の方法で T を修正する。これはプログラムの集合 $\delta_i(T)$ に類似しており、 δ_i は T を入力とし、課題 T_i のために修正されたプログラムを生成するプログラムである。分岐網と同様に、課題が比較的単純な修正によって関連している場合、これは効率的な表現となる。一般的なケースでは、分岐網によって実行される単純な拡張に限定されない。課題選択性に関する上記の結果は、TD 命令が BU 流による処理を効率的に変換し、選択された課題に選択性に集中できることを示している。分岐ネットワークと BU-TD ネットワークの両方において、課題の追加はネットワークに新しいパラメータセット(シナプス重み)を追加する。分岐ネットでは、新しい課題はネットに新しい分岐を追加する。BU-TD ネットでは、追加課題は TD 命令ベクトルの埋め込み形式を作成する重みを学習する必要がある。

分岐ネットワークの使用と BU-TD 処理は、自然な形で組み合わせることもできる。TD 反対流は分岐ネットワーク、さらには木構造にも適用できる。分岐木は「反対木」を持ち、TD 命令は木に沿って異なる分岐のトップに入力を提供する。木の各分岐の BU 部分は、関連する下位課題のセットを群化し、分岐の TD 部分によって制御される。このような BU-TD 反対木構造が、課題間の相関の程度が異なる多数の課題の学習に対する一般的な解決策を提供できるか、また、以前に学習した課題への影響を最小限に抑えながら、時間の経過とともに新しい課題を追加できるかを、今後の研究で検証することは興味深い。

5.2 組み合わせによる汎化

画像から構造記述を抽出する際、我々は物体の新しい構成、それらの特性、相互関係を扱うという問題に直面する。前節で使用した Actors データセットのような限定された領域であっても、様々な人物、その特性、相互関係を選択することで、膨大な数の情景を作成することができる。自然画像の場合、構成の数はもちろん多く、この大きな可能構成空間は、可能情景集合のごく一部に対する経験に基づいて、新規情景の構造表現を抽出できる必要があることを意味する(Tokmakov+2019)。組み合わせ汎化、または構成汎化と呼ばれる、新しい構造を汎化する能力は、情景理解や他の認知課題において重要な役割を果たす。組み合わせ汎化に関する最近の議論(Battaglia+2018)は、「...組み合わせ汎化は、AI が人間のような能力を達成するための最優先事項でなければならず、構造化表現と計算がこの目的を実現する鍵である」と結論づけている。情景解釈の領域において、組み合わせ汎化の達成とは、異なる抽象構造を持つ情景を扱う能力(4.2 節視覚ルーチン

の構成: 次の TD 命令の選択 参照), および同じ構造について, 物体とその特性の新しい組み合わせ, および新しい物体に適用される関係に汎化する能力を意味する。

組合せ汎化は, 図 25a の例のような単純な構成すでに生じる。我々はこのような画像を, 髮型, 眼鏡の種類, シャツなど(3.1 節 物体と同様)の属性を持つ 2 人の人物を各画像に含むテストに使用した。課題は, 画像から 2 人の人物の同一性と, それぞれに関連する特性のセットを復元することであった。組み合わせ汎化の側面については, 以下の方法でテストした。多くの人物と特性の組み合わせが訓練セットから除外された。例えば, 7 番の人物は訓練中にメガネタイプ 3 で提示されることはなかった。人物 7 は他のメガネ型と一緒に提示され, メガネ 3 は異なる人物と一緒に提示されたが, 特定の組み合わせ(人物 7, メガネ 3)は訓練中に除外された。テストは汎化テストと非汎化テストの 2 つに分けられた。より標準的な非一般化テストでは, テスト画像に写っている人物の少なくとも 1 人について, 新しい構成, つまり, 訓練では見られなかった人物, メガネ, シャツ, 髮型などの構成が示されたが, 除外された組み合わせは使われなかった。対照的に, 汎化テストでは, 各テスト画像は, 訓練で除外された人物・特性の少なくとも 1 つを使用した。このような新規の人物・物性の組に汎化する能力は, 組み合わせ汎化の単純な例である。このような「除外された対」に対する汎化の限界は, 視覚的・非視覚的課題の両方におけるネットワークモデルのモデリングにおいて指摘されているが(Kriete+2013, Lake&Baroni2018, Yaari2018, Bahdanau+2019, Vankov&Bowers2020), 以下のような課題も含め, この問題はほとんど未解決のままであった。

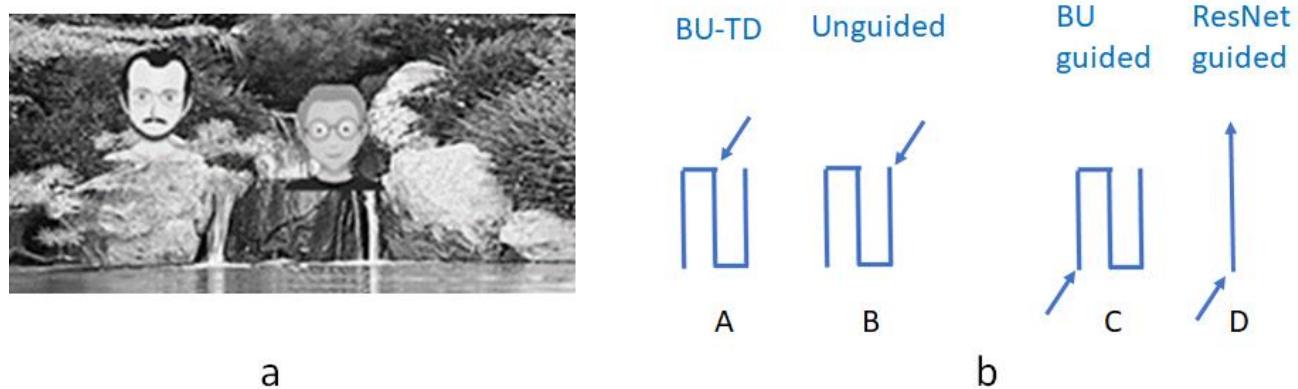


図 25: 組み合わせ汎化

- a. 画像の例。課題は, 人物の身元とそのすべての特性(眼鏡の種類, 髮型など)を抽出することである。
- b. 課題に使われる 4 つのネットワークモデルの概略図。図は各モデルのバックボーンを示しており, 矢印はネットに指示を与える場所を指している。主な比較対象は, BU-TD のガイド付きモデルである A と, 同様のネットワークだがガイドなしで, 代わりにすべての人物のすべての特性を同時に生成する B である。C,D はガイド付きモデルのバージョンであるが, 命令は TD 方式ではなく, 入力と一緒に与えられる。

図 25 (Combinatorial generalization) に模式的に示した 4 種類のネットワークモデルで汎化を比較した(詳細は補足 S10 組み合わせ汎化で説明)。最初のモデル(図 25b の BU-TD) は BU-TD モデルである。図は, モデルの BU-TD-BU バックボーンを模式的に示しており(図 3c と同様), 矢印は, TD 流の先頭で, ネットに命令を提供する場所を指している。この命令を使用して, ネットワークは選択された課題を適用し, 必要な情報を順次抽出する。主な選択肢は, 課題選択を使わず, 必要な情報を 1 回のパスでまとめて抽出することであり, これは(B)のモデルで実現されている。モデル間の比較ができるだけ近くするため, (B) のモデルでは, 命令ベクトルの位置を除き, BU-TD モデルと同じネットワークを使用しており, モデルの最終層(読み出し選択)に接続されている。BU-TD モデルでも読み出し選択モデルでも, 命令ベクトルは例えば<メガネ型, 人-3>を示すことがあり, この場合, 正しい出力は画像中の 人-3 メガネ型となる。この 2 つのモデルはほぼ同じ構造を共有し, 同じ学習手順に従っている。しかし, BU-TD モデルとは異なり, 選択された命令に対して正しい出力を提供するために, 読み出し選択モデルは, 1 回のパスで画像内的人物のすべての特性を同時に抽出しなければならない。読み出し選択は, どのプロパティを抽出するかを決定するのではなく, 最終層からの読み出しを選択するだけである。従って, 指示された課題のみを計算するようにガイドされるガイド付き BU-TD モデルとは対照的に, 読み出し選択モデルを「ガイドなし」と呼ぶ。選択的読み出しは, (非一般化テストでは)すべての出力が順次読み出されるのではなく, まとめて読み出されるモデルと同等の性能である。選択的読み出しを比較に用いたのは, 同じ構造で同じ学習手順を持つネットワークを比較することができ, 同時に一方は誘導され(指示された出力のみを計算), 他方は誘導されない(すべての潜在的出力を同時に計算)ためである。同時出力を生成するモデルとのさらなる比較については, 補足 S10 を参照のこと。

ガイド付きモデルとガイドなしモデルの比較に加え, ガイド付きモデルの 2 つの追加変種もテストした。「BU 誘導」モデル(図 25c) は, TD 流に対してではなく, (追加入力チャネルとして) 入力画像と一緒に指示が与えられることを除けば, BU-TD と同じである(補足 S10)。最後に, ResNet とラベル付けされたモデルは, 一般的に使用されている「残差ネットワーク」深層ネットワークアーキテクチャである(He+2017)。BU-TD モデルと同様の構造, ユニット数, 接続を持つが, BU 流と TD 流間の横方向の接続はない。標準的な ResNet モデルとは異なり, BU-TD ネットワークと同様に課題命令を使用する。

次に, 異なるモデルの組み合わせ一般化結果を考察する。中心的な比較は最初の 2 つのモデル間であり, BU-TD モデルと, すべての特性と一緒に生成する非ガイド・モデルとの比較である。すべてのモデルは, まず訓練画像のデータセットで訓練され, 除外された組み合わせを含まないテストセット画像でテストされ, 上述の非一般化テストが行われた。次にモデルを除外した組でテストし, その結果を汎化テストで報告する。各課題で, スモールサイズとラージサイズの 2 つの訓練セットを使用した。スモールサイズは, BU-TD が除外されたセット

(以下、「十分セット」と表記)において少なくとも85%の精度基準を達成した最小の訓練例セットであり、ラージサイズはその数倍の大きさであった(拡張セットと表記、図参照)。組合せ汎化テストと非汎化テストの結果を図26に示す(人物特性汎化)。図26の一番上の行は、1600枚の画像からなる訓練セットで、十分なセットに対する抽出された特性の精度を示している。各モデルの訓練は、事前に設定した収束基準(50エポックにわたって2%未満の改善)に達するまで、複数エポックにわたって継続した(補足S10訓練詳細を参照)。下段は、拡張セットに対する同じモデルの結果である。主な比較対象は、ガイド付きBU-TDモデル(図中A、緑)とガイドなしモデル(B、青)である。見てわかるように、BU-TDモデルは十分なデータですでに高い精度に達しており、非汎化テストと般化テストの両方で精度が高い。対照的に、ガイドなしモデル(B)は、最終的に非般化テスト(拡張セット、下段)で高精度に達するが、組み合わせ汎化での精度は低いままで、偶然のレベルに近い(23%)。代替のガイド付きモデル(C,D)は、より長い訓練を必要とするが、最終的にはすべて高い汎化精度に達する。

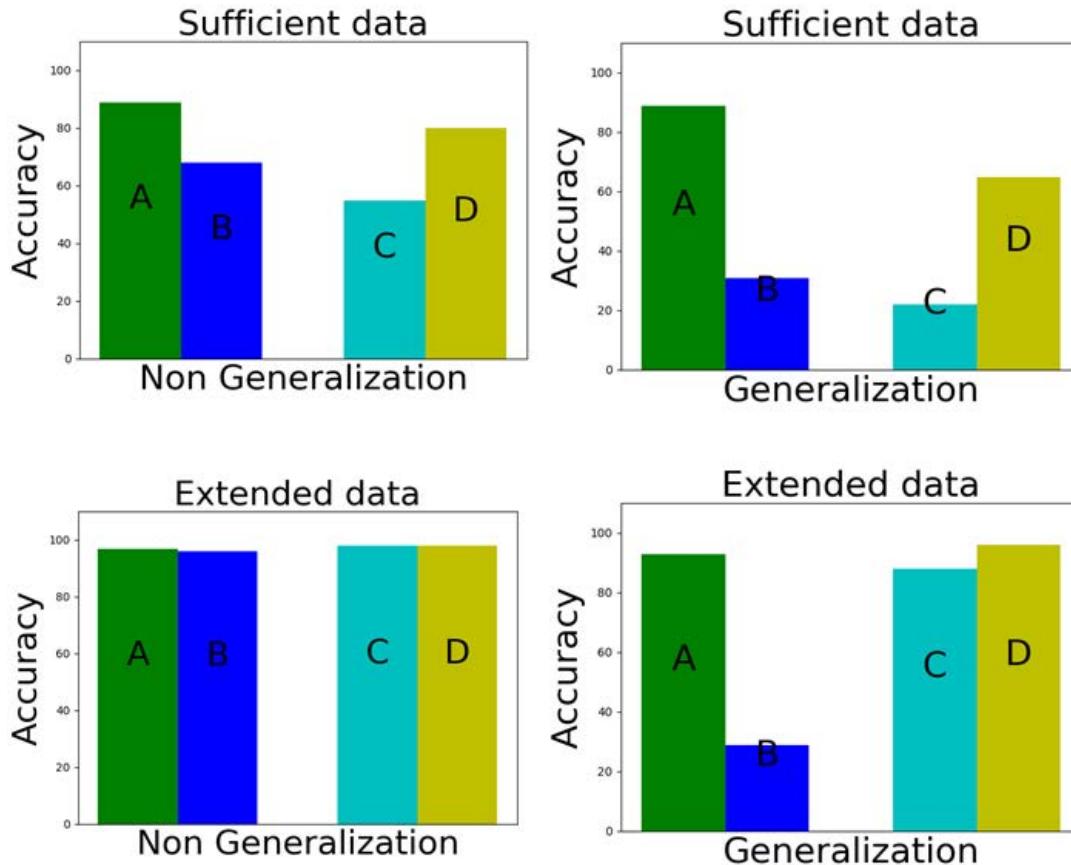


図26: 人物プロパティの汎化。組み合わせ汎化、人物特性。

- 上段: 1600例(十分なデータ)で収束まで学習。
- 下段: 収束するまで4800例(拡張セット)で訓練。

結果(平均精度)は図25の4つのモデルのものである。主な比較は、ガイド付きBU-TDモデル(Aと表示)と、似ているがガイドなしのバージョン(B)である。C,Dのモデルは代替のガイド付きモデルである。組み合わせ汎化テストでは、ガイドなしモデルは低いままであり、変化レベルに近い(23%)。拡張データ(下)では、すべてのガイド付きモデルが組み合わせ汎化を示すが、ガイドなしモデルは低いままである。

組み合わせ汎化: 右隣と左隣の関係

図27(関係汎化)に示すように、left-ofとright-ofの空間関係を汎化する別の課題で、組合せ汎化の同様のテストを行った。このテスト用の画像には、全29文字から選ばれた24文字のEMNIST文字が含まれていた。課題は各文字の右隣と左隣を生成することであった。BU-TDネットワークはこの課題のために2つの損失で訓練された。ひとつは、最初のBUパスの最後に、画像に存在する24文字すべてを識別する精度を測定する(二値交差エントロピー損失)。2つ目は、<direction-of, x>という形式の命令を受け、2回目のBUパス終了時にxの右隣または左隣を生成することである(交差エントロピー損失)。比較に使用したネットワークは、前の例(図25組み合わせ汎化)と同じである。



図27: 関係汎化EMNISTのright-of, left-of課題の組み合わせ汎化。

- a. 24文字の入力画像の例。
- b. 100,000文字対の例(十分なデータ)を用いた訓練結果(平均精度)。

- c. 400,000 文字対の例からなる拡張学習セットでの結果。テストされたモデルは図 26(人物特性の汎化)と同じである。主な比較は、ガイド付き BU-TD モデル (A) と、類似しているがガイドなしバージョン (B) である。C,D のモデルは代替のガイド付きモデルである。例数が少ない場合、BU-TD モデルだけが非汎化テストと組み合わせ汎化テストの両方で高い精度を達成する。十分なデータがあれば、すべてのガイド付きモデルは組み合わせ汎化を得るが、非ガイド付きモデルは汎化に失敗する。

訓練は 2 つのデータセットで前回と同様に行われた。小さい(十分な)セットは 10,000 枚の画像で学習され、各画像からランダムに 5 文字を選択し、学習用と両方向で合計 100,000 個の学習実体(文字対)を使用した。より大きな(拡張)セットでは、各画像と両方向から 20 対すべてを使用し、合計 400,000 の文字対を使用した。このセットの連続する文字の対は、どの訓練画像にも表示されていない。前の例と同様に、BU-TD モデルが課題(十分なデータセット)を学習するまでに、代替モデルは汎化テストと非汎化テストの両方で低い精度を出す(図 27b)。訓練例数が増えると(400,000 まで)、指示されたモデルは追いつき、良い汎化を達成するが、指示のない BU モデルは再びほとんど汎化を示さない(図 27c)。除外される対の割合を増やす(全対の 59% まで)この実験を繰り返したが、BU-TD モデルによって得られた組み合わせ汎化は上記と変わらなかった。このテストの詳細は補足 S10 組み合わせ汎化に示す。

上記の結果では、非誘導モデル(課題選択を行わない他のモデルも同様、補足 S10 参照)は、訓練で見られなかった組み合わせに対する汎化を示さなかったか、あるいはわずかであった。さらに、課題を十分に単純化することで、非ガイド付きモデルで汎化が増加するかどうかを検証した。図 28 (EMNIST-6) に示すように、EMNIST 課題を 6 文字に削減し、多くの訓練例を使用することで、汎化が可能になることが示された。図 28 は、テスト画像 2,000 枚(10,000 組)の十分なセットで学習した後の精度結果である。前回と同様に、BU-TD は高い精度に達しているが、ガイドなしモデルは低いままである。大きな訓練画像セット(50,000 対で 10,000 枚)では、非ガイドモデルの組み合わせ汎化精度は向上し、非汎化の場合の精度に近づくが、BU-TD モデルより一貫して低いままである(77% 対 94%)。学習には時間がかかり、ガイド付きモデルに比べて約 10 倍の例提示(より多くのエポックで例を繰り返す)が必要であった。最後に、このテストを繰り返したが、学習中に除外する桁関係対の割合を増やすことで、組み合わせ汎化をより難しくした。図 28d に見られるように、非ガイド付きモデルの精度は急激に低下し、除外された対の割合が 63% になると 3% に低下する。

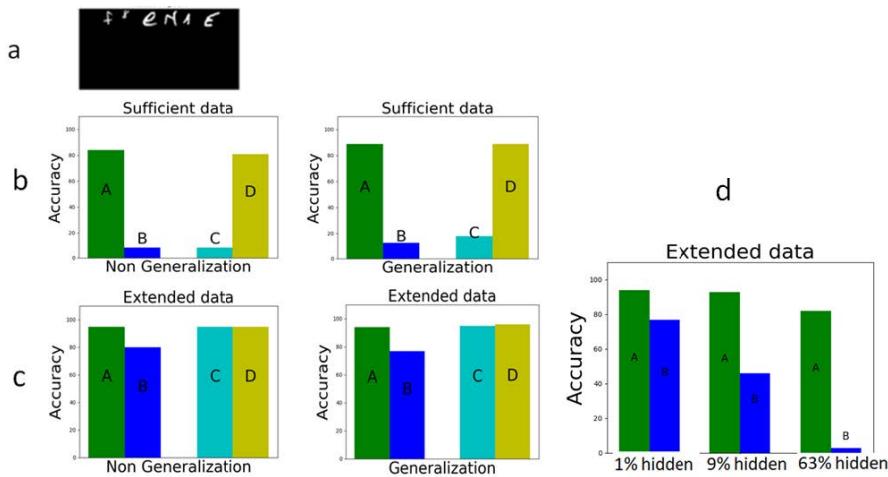


図 28: EMNIST-6 汎化 6 文字の EMNIST の right-of 課題と left-of 課題の組み合わせ汎化。

- a. 入力画像の例。
- b. 10,000 文字対(十分なセット)での学習後の結果。BU-TD モデルは高い精度に達するが、非ガイドモデルは低い精度のままである。学習画像セットを拡張した場合(50,000 文字対)、非ガイドモデルは有意な組み合わせ汎化を達成するが、非ガイドモデルよりも低いレベルで飽和する(77% 対 94%)。
- c. 学習中に除外する文字対の割合を増やす。除外された対の割合が 63% になると、ガイドなしモデルの精度は 3% に低下する。

これらの結果は、課題が十分に単純化され、訓練例の数が増えれば、無誘導モデルでもある程度の組み合わせ汎化を達成できることを示している。対照的に、複数の課題を同時に学習する一般的なモデル(上述した多分岐ネットワークなど)では、モデルで使用される特定のアーキテクチャによる制限のため、組み合わせ汎化が起こらないことは注目に値する(補足 S10)。

まとめると、汎化テストでは、組み合わせ汎化について主に 2 つの結論が得られた。第一に、指示されたモデルは、すべての特性や関係を同時に抽出した非指示モデルに比べて、はるかに優れた汎化をもたらした。第二に、代替の指示モデルの中で、BU-TD による指示の使用は、汎化の性能、例数、学習に必要な全体的な提示回数の点で、代替モデルと比較して優位性を示した。

汎化と複合命令の組み合わせ

最後に、3.1 節(物体、下位節「プロパティ」、複合命令)で説明した、複合命令の使用と汎化の組み合わせもテストした。複合命令では、BU-TD モデルは複数の課題で学習され、各課題は単独で学習される。訓練後、複数の特性に関する命令は 1 つの命令にまとめられ、順次ではなく同時に抽出される。つまり、別々に学習した課題と一緒に実行した場合、前節のように別々に学習し実行した課題と比較して、同じような組み合わせによる汎化を示すかどうかである。上記と同様に、今回のテストでは 2 人の人物を示す画像を用い、それぞれの特性を抽

とする課題を行った。BU-TD モデルは、画像中の人物の 1 人の特性を各サイクルで抽出する TD 命令を用いて学習された。テストでは、同じ人物の 2 つの特性を抽出する TD 命令を 1 つの命令に統合し、それにより 2 つの特性を順次ではなく 1 サイクル内で抽出した。組み合わせによる汎化をテストするため、いくつかの人物と特性の組み合わせ(例: 人物-2, メガネ型3)は、訓練中に再び除外された。次に、2 つの課題を 1 つのインストラクションに統合し、2 つのうちの 1 つが除外された組み合わせを使用するようにした。上記のテストと同様に、BU-TD モデルは汎化テストでも非汎化テストでも同じ精度を得た。複合命令に対する汎化の結果は、逐次学習から生じる利点と並列実行の利点を組み合わせる可能性を提起している。異なる特性を独立に、異なるタイミングで学習することができるが、追加学習を必要とせず、組み合わせ汎化を損なうことなく、後で同時に回復することができる。

5.3 命令の記号表現から埋め込み表現へ

TD 指示の一般的な側面は、系の高レベルな部分では、最初は記号的な形式で表現できるが、学習によって、視覚流の後続処理を導くのに便利な埋め込み形式に変化することである。

記号系では、1 つの領域の要素間に写像が存在し、2 つ目の、いわゆる「意味的」領域を表現するのに使われる。ここで使われる方法では、「記号的」(または「純粹に」記号的)表現は、表現される物体に関する情報を持たない。埋め込み表現では、記号はベクトル空間に埋め込まれ、物体のベクトル表現が表現された物体に関する情報をを持つようになる。

本節では、BU-TD モデル(2.2 節(対向流モデルにおける学習)および図 6(命令の提供))を導く TD 命令において、純粹に記号的な形式から埋め込み形式への移行の側面について述べる。我々が実施したほとんどの実験では、ネットワークに提供された TD 命令は、実行すべき課題と、選択された人物や物体の特性を抽出するような、課題を適用する引数の 2 つの構成要素にすでに分割されていた。以下の実験では、代わりに完全な命令が 1 つのワンホットベクトル(課題と引数を組み合わせたもの)で提供され、学習された埋め込み表現において、その構成形式が独自に獲得されることがわかった。さらに、モデルが組み合わせ汎化に到達する能力は、構成的埋め込み表現の出現に依存していた。

例として、図 4(TD 指示)、2.1 節(ネットワーク構造)のように、人物から選択された特性を抽出する。我々は、6 人の人物から 5 つの異なる特性を持つ人物を選択し、1 つの画像につき 2 人の人物を持つ画像のデータセットを使用した。従って、 $6 \times 5 = 30$ 個の課題が存在し、課題は例えば人物 4 のシャツの種類を特定することである。各課題はワンホットベクトル、すなわち適切な位置に 1 つの「1」を持つ 30 個の長さのベクトルで表現された。この命令ベクトルは、2.2 節(反対流モデルにおける学習)と図 6(命令の提供)と同様に埋め込み形式を作成するが、元の命令が課題と引数の部分に分かれていなければ、このような形式は存在しない。つまり、課題と引数に Itask, Iarg を使用する代わりに、1 つのワンホット命令ベクトル I を使用した。前述と同様に、ベクトル I は学習された埋め込み形式 E に変換され、TD 流に入力を提供する。

その後、モデルをさまざまな課題で訓練した。上記と同様に、ある課題の一部を訓練から除外することで、組み合わせ汎化をテストした。第一に、単一記号命令形式と、TD 命令が物体とプロパティに分割された構成形式を用いたモデルの性能を比較する。第二に、最初に非構成形式で課題が与えられた場合、組み表現を学習する際に、構成形式が自ずと出現するのだろうか? 第二の疑問を検証するため、学習後に埋め込み表現からの読み出しを行った。読み出し(ReLU を用いた 2 つの線形層)は、指示中の選択された人物と性質を回復するように訓練され、訓練中に指示の一部を使用し、残りでテストした。例えば、ワンホットの命令ベクトルの 1 つが、(glasses-type, person-2) という組み合わせを指定する場合、読み出しの訓練中に命令を見ずに、埋め込まれた命令から人物とプロパティ glasses-type を復元することは可能だろうか? このような読み出しが記号表現からは不可能であるが、学習中に埋め込み形式で生じる可能性があり、これは構成表現の学習を示す。

モデルの性能をテストするために、我々は学習課題の単純バージョンから始めた。この課題では、可能な対のごく一部(7.5%)(メガネ型-4, 人-2 など)を訓練から除外した。非構成モデルは課題をよく学習し、91% の組み合わせ汎化精度を達成した(Itask Iarg)。埋め込みフォームから人物と物件を読み上げる学習では、テスト精度は 100% であった。この場合、モデルは課題をよく学習し、構成表現への遷移を自ら生成した。

次に、人物と特性の対のうち、低い割合(7.5%), 中程度(40%), 高い割合(75%)を除外することで、学習は次第に難しくなった。例えば(person-2, glasses-type-k)の指示はすべての k の値で訓練から除外され、person-2 は glass-type プロパティで訓練されることなく、課題全体(glasses-type, person-2)は訓練で使用されなかった。30 個の課題のうち、合計 9 個が訓練中に除外された。

Exclusio n	Readout accurac y	Predicti on accurac y	Compos itional instructi on
Low (7.5%)	100	91	93
Medium (40%)	68	62	91
High (75%)	9	32	80



表2: 記号的埋め込み TD 命令の記号的表現から構成的表現へ。課題は、選択された人物から選択された性質を抽出することである(右の画像例)。「読み出し精度」欄は、埋め込まれた命令表現から、指示された人物と性質を読み出す精度(%)である;高い値は、構成構造の読み出しが成功したことを示す。「予測精度」は組み合わせ汎化の精度である。「構成命令」は、TD 命令が標準的な構成形式(2つの命令ベクトルを使用)で与えられた場合の予測精度を示す。この結果は、組合せ的汎化(予測精度)が、構成的表現(読み出し精度)の出現とともに生じることを示している。

2つ目の疑問(構成表現の出現)に答えるため、3つの条件それぞれについて、学習後に埋め込み表現からの読み上げを行った。すべての条件において、読み上げ課題を訓練する際、除外は完全な性質であった: 例えば <glasses-of, person-2> という命令を読み上げる場合、読み上げ訓練では、どのようなタイプのメガネをかけた person-2 の例もすべて除外する。従って、glasses-of という特性を持つ人物-2 の組み合わせは、読み上げテスト時には新規のものであった。

結果は表2: Symbolic-embedded にまとめられている。この表は、異なる条件(訓練において、人物と特性の組み合わせの割合が低いもの、中程度のもの、高いものを除く)について、3つの行がある。最初の2列は、各条件に対する2つの指標を示している。すなわち、埋め込み表現から指示された人物とプロパティの読み出し精度と、組み合わせ汎化のテストにおけるモデルの精度である。比較のために、3列目の構成指示は、TD 指示の構成形式を使用した場合の精度を示している。表が示すように、TD 命令が非構成形式(1つのワンホットベクトル)で与えられた場合、2つの指標は連動して低下する: 予測精度は低下し、同時に埋め込み表現からの人物と特性の読み出しも低下する。加えて、記号的指示はまた、著しく長い訓練を必要とした(例えば、40% 除外の場合、348 対 148 エポック)。元の BU-TD モデルでは、TD 指示はすでに構成形式で与えられていた: 指示された人物と抽出すべき特性は、それぞれ独自のワンホットベクトルによって別々に与えられる(図6 提供指示の Iarg, Itask)。難易度が高くなるにつれて、モデルの精度も低下するが、依然として高い精度を維持している: 低排除条件、中排除条件、高排除条件では、それぞれ93%, 91%, 80% である。

その結果、組み合わせ汎化には TD 指示の構成形式が重要であることが示された。TD 命令が記号的なワンホット形式で提供された場合、汎化は、構成要素である人物と特性が埋め込み課題表現から読み出される限りにおいてのみ可能であった。組合せによる汎化は、既存の構成要素の新しい組合せを扱う能力に依存するため、このような構成表現の利点は予想される。簡単な課題(ほとんどすべての可能な組み合わせで学習)では、ワンホット表現から構成表現への移行は学習中に勝手に現れた。しかし、課題が難しくなると、移行は失敗し、埋め込み表現から構成要素を読み出すことができなくなった。この発見の意味と、構成性を利用した生得的な構造との関係については、最後の考察でさらに後述する。

6. 人間の視覚との関係

本節では、モデルと人間の視覚との関係について簡単に説明し、今後の研究の方向性について述べる。

6.1 靈長類の視覚における反対流路

フィード・フォワード型ディープ・ネットワークと靈長類の視覚野の顕著な違いは、視覚野では、皮質領域の階層構造の高位から低位へと向かう大規模なトップダウン接続が存在することである。つまり、視覚野 A がその出力をより高次の B 野に送る場合、B 野から A 野へとする相互結合も存在する(Van_Essen&Maunsell1983, Markov+2013)。解剖学的データと生理学的データに基づいて、視覚野における BU と TD の流れのモデル(Ullman1995)は、図 29(反対流路皮質構造)に示す概略的な結合図を提案し、(Markov+2014)に基づいて拡張した。この図は、相互に接続された 2 つの皮質領域(I)と高次領域(II)の間の接続性を示している。BU 流路は I 野の第 4 層と第 3 層 B を通り、高次野の

第4層へ向かう。TDの流れは、II野の2/3A層と6層を通って、下位野の2/3A層と6層へ向かう。表層における上昇集団と下降集団の分離は、(Ullman1995)で予測され、(Markov+2014)で実証された。

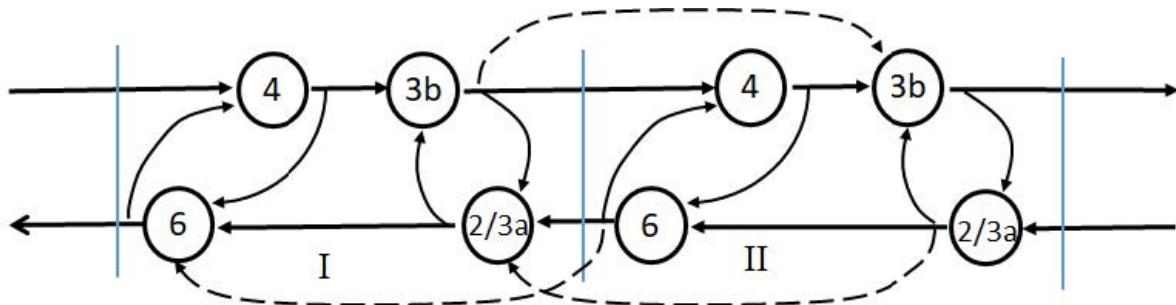


図29: 反対流皮質構造(Ullman1995から引用, Markov+2014からの追加)。基本的な反対流構造は、腹側流路の皮質結合にどのように具現化されるであろうか。I, IIとラベル付けされた連続する領域間の主要な接続。上向きの BU 経路は、第4層から I 野の第3B 層へと進み、次の領域へと向かう。下降する TD 経路は、II 野の 2/3A 層と 6 層から、下の野の 2/3A 層と 6 層へと向かう。この 2 つの流れは、4 層と 6 層の間、および表層 2, 3 層内の両方向で相互に連結している。破線の矢印は、流れのステップをスキップする経路である。

5.1節(マルチタスクと容量)で述べたように、反対流構造は、TD命令がツリーに沿った様々な分岐の先頭に入力を提供するカウンタツリー形式にも拡張できる。ツリーの枝は、その枝のTD部分によって制御される一連の関連課題をグループ化する。霊長類の視覚系では、視覚流路が、顔、物体、場所などの大カテゴリや、自然物と人工物(Chao&Martin2000), 屋内と屋外の情景(Henderson+2007)などの小カテゴリに、何らかの形で分岐していることを示唆する証拠がある。このような階層的なツリー構造におけるBU-TD処理を、計算的にも霊長類の視覚系をモデル化する上でも、さらに探求することは興味深い。

6.2 記号的表現と構成的／組込み的表現

5.3節(記号的命令表現から組込み命令表現へ)では、「記号的」TD命令と「構成的」TD命令の使い方を比較した。構成型では、実行する課題と、それを何に適用するかが別々に与えられ、「記号型」ではそれらが1つの命令、例えば1つのワンホットベクトルにまとめられる(Zhao+2018)。その結果、構成形式の明らかな優位性が示された。また、合成形式は訓練中に学習できるが、合成形式から訓練を開始した方が、組み合わせ汎化の点で実質的に有利であることも示された。組合せ的汎化の重要性を考えると、各課題について個別にそのような表現を学習するよりも、異なる視覚的課題を学習する最初から、TD指示をその構成形式で使用する方が非常に有利であるようと思われる。

BU-TDモデルにおける構成的表象の使用は、知覚と認知における構成的表象の広範な使用の一例に過ぎないが、生得的表象と学習された表象に関するさらなる問題を研究するのに役立つ(Ullman2019)。一般的に2つの可能性がある(その中間の選択肢もある)。1つは、TDの指示は非構成的な形で与えられ、構成構造は各課題ごとに学習されるというものである。もう1つの可能性は、視覚流路を誘導する構成表現の利点は進化の過程で発見されたものであり、学習をより効率的にする生得的な側面を持っているというものである。人間の知覚と認知の観点からは、発達初期の視覚知覚における組み合わせ汎化を研究し、その生得的な側面と学習された側面を明らかにすることは興味深い。コンピュータビジョンの観点からは、ネットワークモデルの構造に構成的表現を取り入れることの利用を探求し、一般的な視覚課題の学習、特に組合せ汎化への効果を研究することが興味深いと思われる。

6.3 TD ガイダンスの機能と利点

このモデルにおけるTD流路の主な機能は、次に何を実行するかの選択とガイダンスである。「認知的」指示を使用することで、実行する課題と、課題を適用する引数を選択することが可能になり、TD流路は、選択された課題を実行するためにBU流路を修正するという意味で、課題をガイドすることができる。この機能を得るために、モデルは入力画像と対応する命令の組み合わせで学習される。BU-TD構造の研究は、関連する情景構造を抽出する目的で、関連する多くの利点を示している。第5節(容量と汎化)で議論される2つの有用な成果は、組合せ汎化と容量制限への対処である。容量に関しては、BU-TD構造の主な貢献は、個々の課題に対して埋め込み課題表現が学習されることである。その結果、各課題は独自の埋め込みパラメータのセットを持ち、メインのBU-TDバックボーンのパラメータに追加される。人間の知覚理論では、能力の限界は注意過程による早期選択と関連している。一般的な考え方として、視覚情景の中には、視覚系の処理能力に限界があるため、一度に処理できる量よりも多くの項目が含まれていることが多い。そのため選択過程は、関連する情報を選択することで、この処理能力の問題に対処するために用いられる(Van_Essen&Maunsell1983, Maunsell2015)。例えば、視覚的注意の役割は、神経科学百科事典(McMains&Kastner2008)で次のように定義されている: 視覚的注意とは、関連する情報を選択し、無関係な情報をフィルタリングすることで、この容量問題に効率的に対処できるようにする認知操作のことである、と定義されている。5節(容量と汎化)の研究は、組み合わせ汎化を可能にするという点で、早期選択の使用に関するもう1つの主要な機能を初めて示している。最後に、冒頭で述べたように、「認知的」指示の使用は、情景分析処理を関連情報の抽出に導くのにも役立つ。この利点を得るために、我々は、情景解釈は、スキームの認知側で「認知的補強」段階を必要とすることを示唆した。最後に、4節(BU-TD系列を使用した情景構造の抽出)でさらに議論するように、TD命令の系列に導かれた複数のBU-TDサイクルを使用することにより、複雑な視覚的課題、特に関心のある構造記述の抽出を実行することが可能になる。また、既存のTD命令の新たな構成を生成し、新たな「視覚プログラム」を作成することも可能となり、、のようにして「有限の手段を無限に利用する」(von_Humboldt1836, Chomsky1965)ことで、新規の視覚課題を実行することができる。

文献

- Bahdanau, D., Noukhovitch, M. and Courville, A. (2019) ‘Systemic generalization: what is required and can it be learned?’, in ICLR arXiv:1811.12889, pp. 1–16.
- Battaglia, P. W. et al. (2018) ‘Relational inductive biases, deep learning, and graph networks’, arXiv.
- Buschman, T. J. and Miller, E. K. (2010) ‘Shifting the spotlight of attention: Evidence for discrete computations in cognition’, *Frontiers in Human Neuroscience*, 4(November), pp. 1–9. doi: 10.3389/fnhum.2010.00194.
- Chao, L. L. and Martin, A. (2000) ‘Representation of manipulable man-made objects in the dorsal stream’, *NeuroImage*. doi: 10.1006/nimg.2000.0635.
- Chen, Z. et al. (2018) ‘GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks’, in 35th International Conference on Machine Learning, ICML 2018.
- Chomsky, N. (1965) *Aspects of the Theory of Syntax*. Cambridge, Massachusetts London, England: MIT Press.
- Cohen, G. et al. (2017) ‘EMNIST: Extending MNIST to handwritten letters’, Proceedings of the International Joint Conference on Neural Networks, 2017-May, pp. 2921–2926. doi: 10.1109/IJCNN.2017.7966217.
- Van Essen, D. C. and Maunsell, J. H. R. (1983) ‘Hierarchical organization and functional streams in the visual cortex’, *Trends in Neurosciences*. doi: 10.1016/0166-2236(83)90167-4.
- Feldman, J. (2013) ‘The neural binding problem(s)’, *Cognitive Neurodynamics*. doi: 10.1007/s11571-012-9219-8.
- Fink, G. R. et al. (1997) ‘Space-base and object-based visual attention: Shared and specific neural domains’, *Brain*, 120(11), pp. 2013–2028. doi: 10.1093/brain/120.11.2013.
- Grill-Spector, K. and Kanwisher, N. (2005) ‘As Soon as You Know It Is There, You Know What It Is’, 16(2), pp. 1–27.
- He, K. et al. (2016) ‘Deep residual learning for image recognition’, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- He, K. et al. (2017) ‘Mask R-CNN’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2), pp. 386–397. doi: 10.1109/TPAMI.2018.2844175.
- Henderson, J. M., Larson, C. L. and Zhu, D. C. (2007) ‘Cortical activation to indoor versus outdoor scenes: An fMRI study’, *Experimental Brain Research*. doi: 10.1007/s00221-006-0766-2.
- von Humboldt, W. (1836) *On Language: On the diversity of human language construction and its influence on the mental development of the human species*. Cambridge University Press.
- Ioffe, S. and Szegedy, C. (2015) ‘Batch normalization: Accelerating deep network training by reducing internal covariate shift’, in 32nd International Conference on Machine Learning, ICML 2015.
- Johnson, J. et al. (2017) ‘CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning’, Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua, pp. 1988–1997. doi: 10.1109/CVPR.2017.215.
- Kafle, K. and Kanan, C. (2017) ‘Visual question answering: Datasets, algorithms, and future challenges’, *Computer Vision and Image Understanding*. doi: 10.1016/j.cviu.2017.06.005.
- Kingma, D. P. and Ba, J. L. (2015) ‘Adam: A method for stochastic optimization’, in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.
- Kriete, T. et al. (2013) ‘Indirection and symbol-like processing in the prefrontal cortex and basal ganglia’, *Proceedings of the National Academy of Sciences of the United States of America*, 110(41), pp. 16390–16395. doi: 10.1073/pnas.1303547110.
- Krishna, R. et al. (2018) ‘Referring Relationships’, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. doi: 10.1109/CVPR.2018.00718.
- Lake, B. and Baroni, M. (2018) ‘Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks’, in 35th International Conference on Machine Learning, ICML 2018.
- Lake, B. M. et al. (2011) ‘One shot learning of simple visual concepts’, in In {Proceedings of the 33rd Annual Conference of the Cognitive Science Society}.
- LeCun, Y. et al. (1998) ‘Gradient-based learning applied to document recognition’, *Proceedings of the IEEE*, 86(11), pp. 2278–2323. doi: 10.1109/5.726791.
- Lecun, Y., Bengio, Y. and Hinton, G. (2015) ‘Deep learning’, *Nature*, 521(7553), pp. 436–444. doi: 10.1038/nature14539.
- Levi, H. (2021) Combining bottom-up and top-down computations for full image interpretation: computations and network models. PhD. thesis, Weizmann Institute of Science.
- Levi, H. and Ullman, S. (2020) ‘Multi-Task Learning by a Top-Down Control Network’. Available at: <http://arxiv.org/abs/2002.03335>.
- Liu, R. et al. (2019) ‘CLEVR-REF+: Diagnosing visual reasoning with referring expressions’, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. doi: 10.1109/CVPR.2019.00431.
- Loshchilov, I. and Hutter, F. (2019) ‘Decoupled weight decay regularization’, in 7th International Conference on Learning Representations, ICLR 2019.
- Lu, C. et al. (2016) ‘Visual relationship detection with language priors’, in European conference on computer vision, pp. 852–869.
- Luo, R. and Shakhnarovich, G. (2017) ‘Comprehension-guided referring expressions’, in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. doi: 10.1109/CVPR.2017.33.
- Malsburg, C. von der (1995) ‘Binding in models of perception and brain function’, *Current Opinion in Neurobiology*. doi: 10.1016/0959-4388(95)80014-X.
- Mao, J. et al. (2016) ‘Generation and comprehension of unambiguous object descriptions’, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. doi: 10.1109/CVPR.2016.9.
- Markov, N. T. et al. (2013) ‘Cortical high-density counterstream architectures’, *Science*, 342(6158). doi: 10.1126/science.1238406.

- Markov, N. T. et al. (2014) ‘Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex’, *Journal of Comparative Neurology*, 522(1), pp. 225–259. doi: 10.1002/cne.23458.
- Maunsell, J. H. R. (2015) ‘Neuronal Mechanisms of Visual Attention’, *Annual Review of Vision Science*. doi: 10.1146/annurev-vision-082114-035431.
- McMains, S. A. and Kastner, S. (2008) ‘Visual Attention’, in *Encyclopedia of Neuroscience*. doi: 10.1007/978-3-540-29678-2_6344.
- Mitchell, M., Van Deemter, K. and Reiter, E. (2013) ‘Generating expressions that refer to visible objects’, in NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference.
- Netanyahu, A. (2018) Cyclical Bottom-Up Top-Down Neural Networks for Relational Reasoning. MSc. thesis, Weizmann Institute of Science.
- P.J. Werbos (1990) ‘Backpropagation Through Time: What It Does and How to Do It’, *Proceedings of the IEEE*, pp. 1550–1560. Available at: <http://ieeexplore.ieee.org/document/58337/?reload=true>.
- Papadimitriou, C. H. and Vempala, S. S. (2015) ‘Cortical learning via prediction’, in *Journal of Machine Learning Research*.
- Redmon, J. et al. (2016) ‘You only look once: Unified, real-time object detection’, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem, pp. 779–788. doi: 10.1109/CVPR.2016.91.
- Ren, S. et al. (2015) ‘Faster R-CNN: Towards real-time object detection with region proposal networks’, in *Advances in Neural Information Processing Systems*.
- Rosch, E. et al. (1976) ‘Basic objects in natural categories’, *Cognitive Psychology*, 8(3). doi: 10.1016/0010-0285(76)90013-X.
- Sabour, S., Frosst, N., and Hinton, G. E. (2017) ‘Dynamic routing between capsules’, in *In Advances in neural information processing systems*, pp. 3856–3866.
- Santoro, A. et al. (2017) ‘A simple neural network module for relational reasoning’, in *Advances in Neural Information Processing Systems*.
- Sener, O. and Koltun, V. (2018) ‘Multi-task learning as multi-objective optimization’, *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS), pp. 527–538.
- Shelhamer, E., Long, J. and Darrell, T. (2017) ‘Fully Convolutional Networks for Semantic Segmentation’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), pp. 640–651. doi: 10.1109/TPAMI.2016.2572683.
- Smolensky, P. (1990) ‘Tensor product variable binding and the representation of symbolic structures in connectionist systems’, *Artificial Intelligence*. doi: 10.1016/0004-3702(90)90007-M.
- Strezoski, G., Noord, N. and Worring, M. (2019) ‘Many task learning with task routing’, in *Proceedings of the IEEE International Conference on Computer Vision*. doi: 10.1109/ICCV.2019.00146.
- Thomas McCoy, R. et al. (2019) ‘RNNS implicitly implement tensor-product representations’, in *7th International Conference on Learning Representations*, ICLR 2019.
- Tokmakov, P., Wang, Y. X. and Hebert, M. (2019) ‘Learning compositional representations for few-shot recognition’, *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob, pp. 6371–6380. doi: 10.1109/ICCV.2019.00647.
- Treisman, A. M. and Gelade, G. (1980) ‘A feature-integration theory of attention’, *Cognitive Psychology*. doi: 10.1016/0010-0285(80)90005-5.
- Ullman, S. (1984) ‘Visual routines’, *Cognition*. doi: 10.1016/0010-0277(84)90023-4.
- Ullman, S. (1995) ‘Sequence seeking and counter streams: A computational model for bidirectional information flow in the visual cortex’, *Cerebral Cortex*. doi: 10.1093/cercor/5.1.1.
- Ullman, S. (2019) ‘Using neuroscience to develop artificial intelligence’, *Science*. doi: 10.1126/science.aau6595.
- Valiant, L. G. (1994) *Circuits of the Mind*. USA: Oxford University Press, Inc.
- Valiant, L. G. (2005) ‘Memorization and association on a realistic neural model’, *Neural Computation*. doi: 10.1162/0899766053019890.
- Vankov, I. I. and Bowers, J. S. (2020) ‘Training neural networks to encode symbols enables combinatorial generalization’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1791), pp. 1–16. doi: 10.1098/rstb.2019.0309.
- Vatashsky, B. Z. and Ullman, S. (2020) ‘VQA with no questions-answers training’, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. doi: 10.1109/CVPR42600.2020.01039.
- Wagner, S. (2020) Learning and executing multiple tasks together vs. one task at a time. MSc. thesis, the Weizmann Institute of Science.
- Wah, C. et al. (2011) ‘The Caltech-UCSD Birds-200-2011 Dataset’. Available at: http://www.vision.caltech.edu/visipedia/papers/CUB_200_2011.pdf.
- Wu, Y. and He, K. (2020) ‘Group Normalization’, *International Journal of Computer Vision*, 128(3). doi: 10.1007/s11263-019-01198-w.
- Xu, D. et al. (2017) ‘Scene graph generation by iterative message passing’, in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2017. doi: 10.1109/CVPR.2017.330.
- Yang, J. et al. (2018) ‘Graph R-CNN for Scene Graph Generation’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-030-01246-5_41.
- Zhang, M. R. et al. (2019) ‘Lookahead Optimizer: K steps forward, 1 step back’, arXiv.
- Zhao, X. et al. (2018) ‘A Modulation Module for Multi-task Learning with Applications in Image Retrieval’.

補足資料

補足資料目次

- S1 The psychophysical timeline study
- S2 The BU-TD structure
- S3 The Persons data set
- S4 The Actors data set

- S5 Classification by location
- S6 Spatial relations
- S7 Extracting full scene structure
- S8 Guided extraction of scene structure
- S9 Selecting the next TD instruction: composing visual routines
- S10 Combinatorial generalization

S1. 心理物理学的タイムライン研究

現在進行中の研究 (S. Ullman, D. Harari, H. Benoni) では、画像からの構造化情報(情景構成要素、特性、関係)の抽出を時間(50-2000 ミリ秒)で追跡した。この研究は、Amazon Mechanical Turk プラットフォームを通じて被験者を募集し、オンラインで実施された。研究の各画像は、7つの提示時間(50, 75, 100, 125, 200, 500, 2000 msec.)のうちの1つで被験者に提示され、その後にマスクが提示された。各被験者は、これらの提示時間のうち1つのみで提示された画像を見た。被験者には、各画像の提示後に、その画像に写っているすべての物体、人物、それらの特性、相互作用を列挙した情景の詳細な説明を作成するよう求めた。下図は、解釈のタイムライン例である。1つ目は、1節(一般的な背景と目標: 視覚と認知の融合)において、視覚と認知の統合について議論した「盗む」シーンのタイムラインを示す。2つ目は、3.3節(関係性)で説明した、参照関係を説明するための2つの画像を示している。

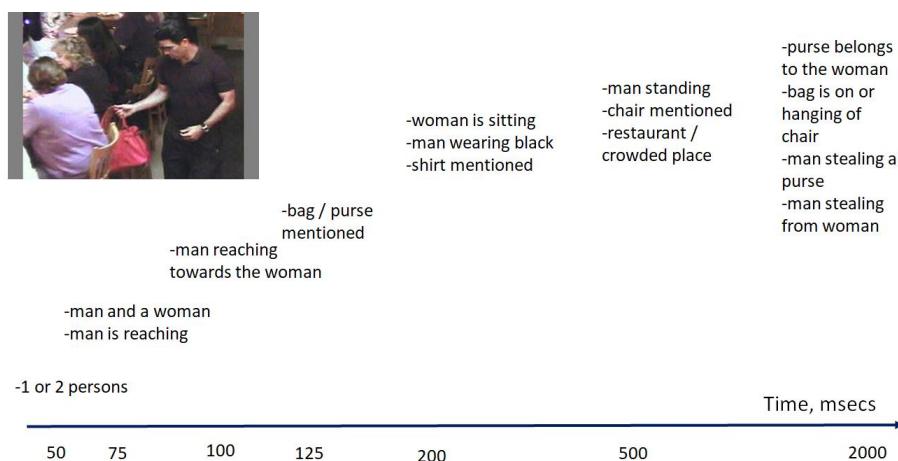


図30：盗みのタイムライン。時間経過に伴う情景構成要素、その特性、関係の抽出。この図は、各時間ビンについて、25人の観察者のうち少なくとも5人が情景記述に追加した情報を示している。

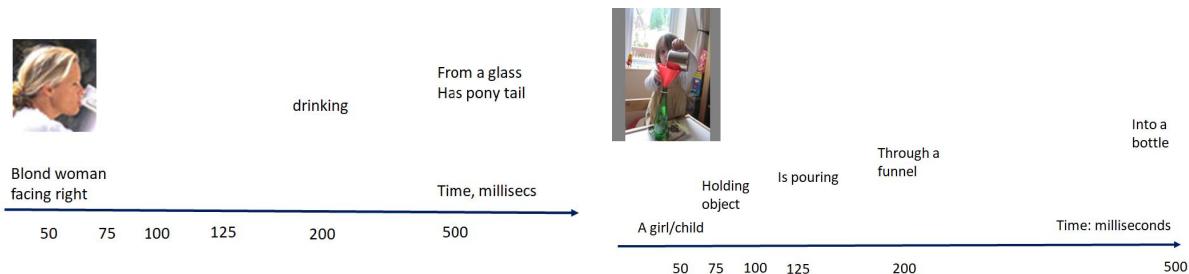


図31: 関係タイムライン図は、各時間ビンについて、情景記述に追加された情報(25人中少なくとも5人の観察者による)を示している。人物が行動している画像では、人物と行動が、行動で使われた物体よりも先に認識されている。

S2. BU-TD 構造

BU-TD 逆行流モデルは、TD 逆行流と横方向の接続を追加することで、異なる BU モデルから BU-TD モデルを生成できるという意味で、一般的なスキームである。我々が使用した主な構造はボトムアップの Resenet 18 ネットワークに基づくものであったが、Persons データセット、EMNIST データセット、Actor データセットに関する課題の BU-TD ネットワークなど、課題ごとに浅いネットワークモデルと深いネットワークモデルの両方を使用した。単一の BU-TD ネットワークを複数の異なる課題に対して学習させることもできるが、複数のネットワークを使用することも可能である。モデルに提供される TD 命令は、`<task, argument>` 対の構成である。例えば、EMNIST の場合は`<right-of, 5>` または`<left-of, w>`、Persons 課題の場合は`<glasses, person-1>` または`<mustache, person-5>` である。課題と引数はそれぞれワンホットベクトルとして別々に符号化され、学習された重みによって埋め込まれた形に変換される。この埋め込みベクトルは TD 流の最上層への入力となる。

交差流側方結合： BU-TD スキームは、BU->TD 方向と TD->BU 方向の2組の交差流側方結合を含む。我々の実験では、より多くの代替バージョン間の計算比較に基づき、横方向の接続に2つの代替スキームを使用した。代替案の1つは、標準的なディープネットワークモデル

で使用される接続に似た、加法的な横方向の接続を使用する。TD->BU 方向に使用した 2 つ目の代替案では、以下に述べる乗法的相互作用を適用した。

TD 流上の層の i 番目のチャネルと k 番目の位置にある単位 $\bar{X}_{i,k}$ を考える。(\bar{X} のような変数の上の上の棒は、TD 流上の変数を示すのに使われる)。このユニットへの総入力 I は、 $\bar{I}_S + I_L$ で構成される。ここで、 \bar{I}_S は TD 流に沿った標準入力であり、 I_L は BU 流からの横方向の寄与である。我々が使用した I_L の 1 つの形式は、畳み込み式であった: $I_{Lk} = \sum_c w_{i,c} x_{c,k}$ 。つまり、TD 流の $\bar{X}_{i,k}$ への側方の入力は、BU 流の対応する層の対応する位置にあるすべてのユニット $x_{c,k}$ から、すべてのチャンネルにわたって入力される。重みは位置 k に依存しない(BU 側では「1*1 畳み込み」とも呼ばれる)。しかし、重みはターゲットチャネル i に依存する。単純化したケースとして、TD 流のターゲットチャネル i への入力が、全チャネルからではなく、BU 流の单一チャネル i からのみである場合、同様のバージョンを使用した。乗算スキーム(「ゲーティング」スキームとも呼ばれる)では、横方向寄与 I_L は $\sum_c w_{i,c} x_{c,k}$ の畳み込み形式を使用するが、乗算形式で使用される: $\bar{X}_{i,k} = \bar{I}_S * I_L = \bar{I}_S * \sum_c w_{i,c} x_{c,k}$ 。この形式は、加法項 $\bar{X}_{i,k} = \bar{I}_S * I_L + \alpha I_L$ (α は学習パラメータ) と組み合わせることもできる。

複数サイクルの訓練: BU-TD モデルはリカレントネットワークであり、複数サイクルの学習が可能である。BU-TD モデルはリカレントネットワークであり、複数サイクルの学習が可能である(2.1 節(ネットワーク構造))。複数のサイクルから構成される処理を訓練するために、我々は 2 つのオプションを使用した。1 つ目は、複数の時間ステップに対してアンフォールドネットワークを使用する方法である。もう 1 つは、2 つのネットワーク(1 つは BU、もう 1 つは TD 部分)を使って多数のサイクルを訓練する方法である。これは、2 つのネットワーク間で活性化値を保存し、受け渡すことで実現される。BU 流の活性化値は記録され、次のサイクルで TD ネットワークの入力として使用され、同様に TD 流の活性化レベルは記録され、次のサイクルの訓練で BU ネットワークの入力として使用される。

中間損失の使用: BU-TD ネットワークは、各流れの終わりに中間出力と損失を自然に使用することができる。BU2 の先頭の 1 つの損失で学習は十分であることが多いが、中間的な損失の使用は、学習処理を加速し、TD 流の下部に検出された物体の切り出し地図のような有用な追加出力を提供するという点で有用である。例えば、我々は EMNIST データセットに対して、異なる損失数で実験を行った。課題は、24 文字を含む画像の空間的関係を計算することであった(3.3 節(関係)、図 14(空間-2D) と同様)。ネットワークは初期ランダム重みから学習され、2 つの損失が設定された。BU1 の先頭では、2 値交差エントロピーを用いて、ネットは画像に存在する全ての文字を識別する必要があった。2 つ目の損失は最終出力で、交差エントロピー損失を用いて 1 文字を生成することであった。訓練セットには 10,000 枚の画像が含まれ、各訓練エポックでは、各画像が 10 種類の TD 命令で 10 回表示され、合計 100,000 回の單一文字出力が行われた。この条件下で、性能は 33 回の訓練エポックで 94% の精度に達した(96% で漸近、39 エポック)。損失の数を変えると、精度と学習速度に影響した。損失が 1 つの場合、この課題では end-to-end 学習は失敗し、わずか 4% の精度で漸近した。3 つの損失を使用すると、TD 流の最後に指示された文字の切り出し損失が追加され、33 エポックではなく 13 エポックで 94% の精度を達成した。我々がテストした他の課題では、最終的な損失は 1 つで十分であることが多かったが、訓練時間が長くなかった。

事前訓練済モデルの使用: 複数課題を学習する場合、課題は何らかの自然な順序で順次習得されることがある。例えば、個々の物体を認識することを、物体間の空間的関係を認識することを学習する前に学習することができる。ある課題 T の学習は、関連する課題についてすでに学習済みのネットワークで学習すれば、より容易になる可能性がある。例として、EMNIST 文字間の空間的関係(right-of, left-of)の課題について、モデルを単独で、ランダムな重みから学習させる場合と、個々の文字認識について既に学習済みのネットワークから学習させる場合を比較した。文字認識課題の学習のために、BU-TD ネットワークは 24 文字の画像を与えられ、文字の序列位置(1 から 24 の間)と、画像内のその文字のグランドトゥルースラベルを指示された。このネットワークは、BU1 の先頭で出現損失(画像に存在する全ての文字)を使用し、最終出力で分類損失を使用した。このネットワークは認識課題を学習するまで訓練された。このネットワークをランダムな重みではなく、事前に訓練された重みとして使用すると、空間関係の課題に対して、26 エポック後に 1 つの損失を使用して 94% の性能を達成し、(事前訓練なしの上述の訓練失敗と比較して)最後まで訓練が成功した。2 つの損失を使用した場合、ランダムな重みで開始した 33 エポックと比較して、11 エポックで同じ性能に達した。初期学習は高速で、3 エポック後に 87% の精度に達する。これに対し、事前学習なしでの 3 エポック後の精度は 4% であった。

S3. The Persons data set

一般に入手可能なデータセットに加え、本節と次節で説明する Persons と Actors データセットを使用した。

Persons データセットは、グラフィックパッケージ <http://avatarmaker.com/> を用いて作成された。各人物は、定常と可変という 2 種類の特徴から構成される。定数特徴は、目、肌の色、顔の形など、このアバターをユニークにする特徴である。髪型、メガネの形、口ひげ、服装などの可変特徴は、アバターの同一性に影響を与えることなく変更できる。以下の表 3 は、一定特徴および可変特徴と、各特徴の変種数を示している。人物画像はカラーで作成することもできるが、我々の実験では、以下に示すように濃淡レベルのものを使用した。

最終的な画像は、サイズ 448X224 の濃淡背景画像上に多数の人物を配置することで構成されている。人物の分類を多少難しくするため、我々の画像では、人物は通常、他の各人格と少なくとも 1 つの一定の特徴を共有しており、1 つの特徴に基づいて人物を識別することはできない。

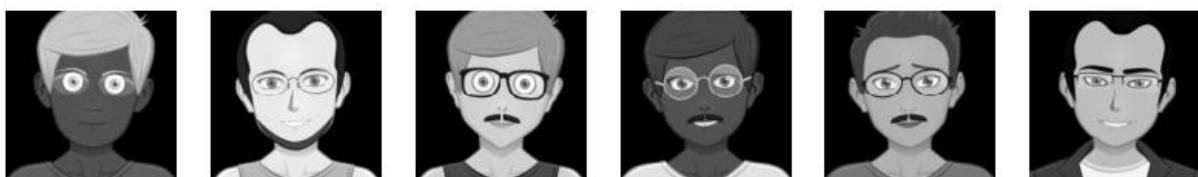


図 32 人物データセット さまざまな人物画像の例。

表3 : Persons データセット 上: 定数パラメーター。下: 可変パラメータ

Parameter	Face	Skin color	Lips	Lips color	nose	ears	eye front	iris	eyebrows	iris color	hair color	eyebrows color	beard color	beard
Variations	15	20	15	20	15	7	15	10	15	20	20	20	20	13
Parameter	Face Tilt		Clothes		Glasses		Hair		Mustache					
Variations	2		5		8		3		2					

S4. The Actors data set

Actors データセットの領域は、 Makehuman というオープンソースのキャラクタ作成ツールを使って作成された、 人体(actors) のオブジェクトと道具を持つコンピュータ生成シーンで構成されている: <http://www.makehumancommunity.org/>。 我々が使用するアクターは、 4 クラス(女性、 男性、 少女、 少年) から成り、 異なるポーズをとることができ、 異なる特性でデザインすることができる。 可能性の幅は広く、 主実験では、 4 つの異なる衣装、 オプションのサングラス、 オプションの帽子、 2 つの髪型、 そして場合によっては物(リュックサック、 ショルダーバッグ) を持ったり、 物(紙袋、 バット、 弓、 ハンマー、 テニスラケット、 信号、 剣) を持ったりすることもできる。 ベンチ、 椅子、 ゴミ箱、 街灯、 木など、 さまざまな種類と色の物体を風景に配置できる。

シーン内の俳優間の関係をいくつか定義し、 訓練時に使用するために注釈を付けた。 例えば、 2 人の俳優が「向かい合う」関係に置かれることがあり、 モデルは新しい画像において「向かい合う」関係を認識するように訓練される。 主な実験で使用した関係には、 空間関係(左、 右、 前、 後ろ)、 対面、 接触、 最近隣(他の人物や物体に)、 妨害(与えられた距離内で、 2 番目のアクターの進路を妨害する)が含まれる。 アクターは 'holding' 関係を介して物体と相互作用することができ、 物体間の関係には空間的関係と 'on' 関係が含まれる。 画像は情景グラフのサンプリング、 すなわち物体とその属性、 情景内位置、 および物体間の関係をサンプリングすることによって生成された。 画像は Blender (<https://www.blender.org/>) を使い、 カメラとライトの位置を指定してレンダリングした。 生成された画像はすべての構成要素の注釈を持ち、 完全な実体分割地図を持つ。 我々の実験で使用された情景成分、 その特性、 および関係の概要は、 以下の表 4 (Actor データセット) に示されている。 Actor データセットで生成された情景の例を図 33 (情景例) に示す。

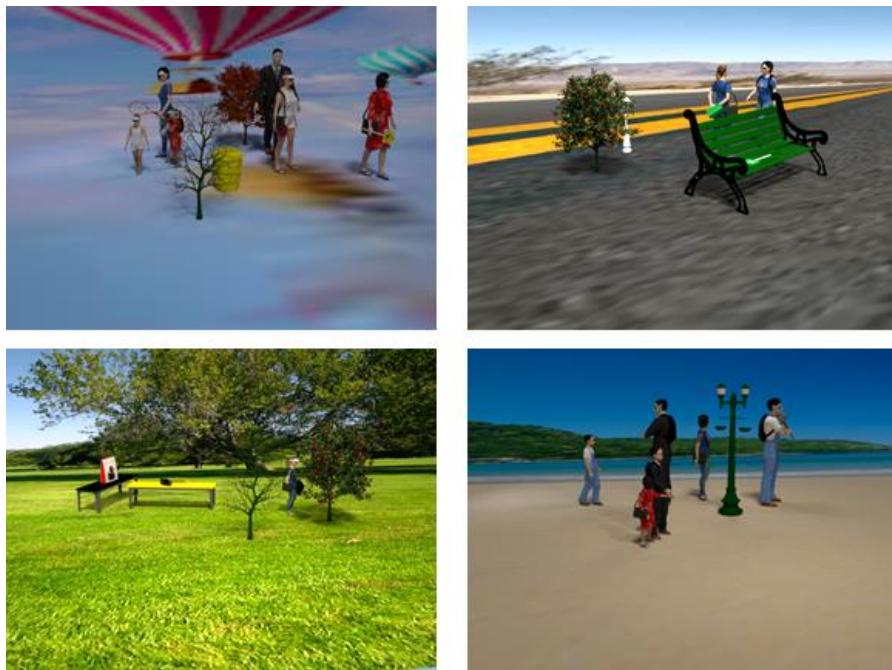


図 33 情景の例 Actors データセットで使用されている人物、 情景物体、 保持物体を示す例。

最後に、 学習データと出力精度を簡単に示す。 精度には、 平均切り出し精度 (IoU, intersection-over-union で測定) と平均分類精度 (構成要素、 特性、 関係) が含まれる。

- 使用した画像数： 訓練 12,000 枚、 テスト : 3,353 枚。
 - 展開ネットワーク： 例数（「例」は画像に TD 命令を適用した結果であり、 各情景には複数の例が含まれる）： 訓練 638,825、 テスト 176,596。
 - 結果 (訓練／検証) : TD での IOU : 0.626/0.59、 BU2 での精度 : 0.855/0.823。
- 精緻化ネットワーク： 画像数： 訓練 2,484,781、 テスト 679,424
 - 結果 (訓練/検証) : TD での IOU : 0.968/0.966、 BU2 での精度 : 0.912/0.899

Objects type	classes	Properties	Interacting relation		
Humans	Man, woman, girl, boy	Clothes, hair, sunglasses, hat	Spatial, closest obstructing, facing, touching, holding		
Scene objects	Bench, chair, trash can, streetlight, tree	Size, color	Spatial, closest obstructing, on		
Held objects	Hammer, handbag, sword, racket, signal	Size, color	Holding, on		
Relation	Ref object location \vec{r}_1	Ref object Rotation θ_1	Related object location \vec{r}_2	Related object Rotation θ_2	Other conditions
Obstructing	random	Random distance r within (0.6,3.5) from \vec{r}_1 in direction θ_1 .	within range that is visible to camera		
Facing	random	Random within visibility range s.t $\theta_1 + \pi$ also visible	$r = \text{average shoulder width of the ref, and the related object direction}$ $\theta_1 \pm \frac{\pi}{2}$	$\theta_2 = \theta_1 + \pi$	
Touching	random	Random within visibility range	r within (0.4,3) from \vec{r}_1 in direction θ_1 .	$\theta_2 = \theta_1$	The arms of both actors in touching position
Right of, left of, front of, behind, closest	random	Random in range visible to camera	Random	Random in range visible to camera	Relation calculated from location

Table 4: Actor data set The table lists the scene components (people and objects), properties, and relations between components. ‘Within visibility range’ means that the person’s rotation in the scene was constraint to make the face region visible.

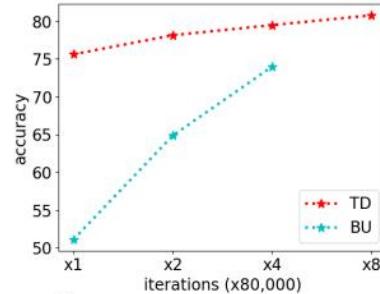
S5. 位置に基づく分類

多物体画像(3.2節 物体、図 8b)中の物体を認識する際、よく使われるアプローチ(例えばRen+2015, Redmon+2016)は、与えられた画像中の全ての物体検出し、分類することである。ここでのアプローチは、オブジェクトの位置がTD命令によって指定されるBU-TDモデルを使用して、選択された1つの物体を認識することである。選択された位置をネットワークに提供するためのいくつかの選択肢を比較した。以下の最初の比較では、TD流の入力として、粗いローカライズ地図(ストライド32)によって命令を与えた。もう一つの選択肢は、入力画像とともにBU方式で供給される空間地図として選択位置を提供することであった。どちらの場合も、物体クラスは、第2のBU流の最後に出力として生成される。

両方のBU流でResNet-50 バックボーンを使用した。全てのモデルはCOCO train2017(118,000画像)で学習され、COCO val2017(5,000画像)で評価された。モデルは80,000回の反復で訓練し、2つ目のBU流の最後に分類損失を与えた。標準的なacc#1とacc#5を分類基準として、Faster-RCNNの2つの主要な変種(C4とFPNと呼ばれる)を使用した結果と比較した。結果を図34aに示す。結果は高度に最適化されたFaster-RCNNモデルの精度に匹敵する。

分類精度は、位置情報が入力画像とともにBU方式で供給される空間地図として供給される場合のBU-TDモデルとも比較された。図34bの結果は、訓練に使用された反復回数の関数として、TD方式で提供された命令の収束が著しく速いことを示している。

		acc#1	acc#5
Faster-RCNN	FPN	83.167	96.752
Faster-RCNN	C4	80.754	96.004
Counter-Stream	TD	80.919	96.810



a

b

図34：場所による分類

- a. 位置情報による分類を用いた Faster-RCNN と BU-TD 逆行流モデルの分類精度。
- b. BU-TD モデルの分類精度は、位置情報が TD 方式(図では TD)で供給される場合と、BU 方式で空間地図として入力画像と一緒に供給される場合がある。

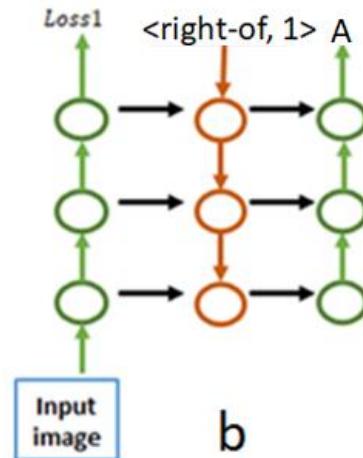
また、BU-TDモデルの結果を、選択された物体の位置を指定する他の代替方法(中心点の位置、バウンディングボックスの座標、オブジェクトマスク)と比較した。漸近結果は同様であったが、TD地図の使用率が最も高く、TD命令による学習はBU命令よりも速かった。

S6. 空間関係

空間関係を学習する例として、EMNISTデータセット(Cohen+2017)の複数の文字の画像を使用する。サイズ224*112の画像には、29個の数字と文字の合計セットから24個の文字が含まれている。入力画像の文字は非繰り返しだった。下の図34bに示すネットワークは、入力画像とトップダウン命令(<right-of, 1> 正解はA)の2つの入力を受け取る。ネットワークはBU1, TD, BU2流で構成され、それぞれ6層で構成され、BU1とBU2の間に重みが共有されている。



a



b

図35：EMNISTの空間関係

- a. サイズ224*112の入力画像。
- b. TD命令によるネットワークの概略構造。

ネットワークは初期ランダム重みから訓練され、2つの損失がある。BU1の先頭では、2値交差エントロピー損失を使って、画像に存在するすべての文字を識別するようにネットが訓練された。2つ目の損失は最終出力で、交差エントロピー損失を用いて1文字を生成した。学習セットには10,000枚の画像が含まれた。1回のエポック中に、各画像は10回のTD命令で表示され、合計100,000回の1文字出力が行われた。この条件下では、39回の訓練エポックで96%の精度で漸近した。これに対し、同じネットワークの非ガイド版(5.2節(組み合わせ汎化))を使用した場合、537エポックで漸近的性能は48%に達した。(これらの結果は非一般化テストの結果である。組み合わせ汎化の結果については補足S10で述べる)

非参照型の空間関係

3.3節(参照関係)では、参照形式と呼ばれる1つの物体を使用する関係指示の例と、2つの物体ト引数を使用する関係指示の例について述べた。関連する2つの物体を指定することで、BU-TDモデルを関係で学習させる追加的な例として、関係Occlude(x,y)がある。

Personデータセットを使って、<occluding,(person-1 id, person-2 id)>という形で指示が与えられ、課題は2つの間の咬合関係、つまり1つ目が2つ目を咬合しているのか、1つ目が2つ目に咬合されているのか、2つの間に咬合はないのかを学習することである。したがって、出力は3つのクラスに分類される。データセットには4800枚の学習画像が含まれる。各画像には4人の人物が写っており、そのうち1組がオクルージョン関係にある。各画像に対して、学習は3回のクエリを用い、毎回2人の人物を選択する。クエリの半分がオクルージョン関係を持ち、半分が持たないようにデータセットを調整した。モデルは3つの損失を使用した: BU1の終了時の出現損失(つまり、どの人物が心像性で現れたか), BU2の終了時の最終課題損失、TD切り出し損失(指示中の2人の人物の切り出し)。このモデルは99%の課題精度を得た。例を以下の図36(Occlude(x,y))に示す。

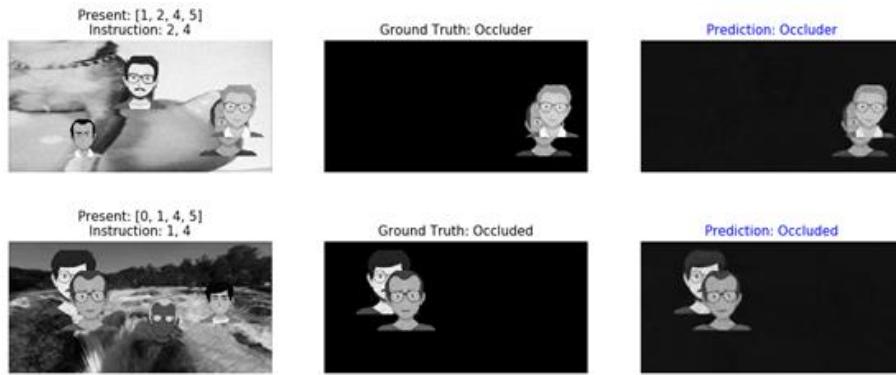


図36 Occlude(x,y)課題は、人物-1と人物-2との間の隠蔽関係を計算することである。すなわち、人物-1が人物-2を隠蔽する(隠蔽者)か、人物-2に隠蔽されるか、隠蔽関係が成立しないかである。左:入力画像とTD命令中の人、央:TD流によって生成された切り出し、右:出力予測。

S7. 全情景構造の抽出

4.1節(情景記述の抽出)で簡単に説明したように、「完全な構造」とは、すべての情景構成要素、それらの特性、および関係を抽出することを意味する。実際の自然情景の分析において、その完全な構造を抽出することは、完全な構造が完全に知られている制御された条件下で、情景の完全な構造を抽出するモデルの能力を調べるために有用である。この補足では、Actorsデータセットに適用された完全な情景構造を抽出するために使用されたアルゴリズムをより詳細に説明する。まず、モデルによって使用される2つのネットワークについて説明し、次に完全な情景構造を抽出する手順について説明する。

情景構造は2つのBU-TDネットワーク(展開ネットワークexpansionと精緻化ネットワークelaboration network)の組み合わせによって抽出される。「展開ネットワーク」は、カメラに次に近い「主」項目(人物または情景物体)に対する参照関係、または「抽出-次へ」命令となり得る命令に従って物体を抽出する。「精緻化ネットワーク」は、命令によって指定された特定の特性/関係を認識する。

展開ネットワーク: このネットワークは、参照情景の構成要素との関係に基づいて人物または物体を見つけ、それを切り出し、そのクラスを認識するように学習される(3.3節関係で説明)。参照命令は、前述の<facing,person-1>のような二項関係であることができ、これは、person-1に向いている画像の人物を切り出し、分類する。引数(person-1)は、入力画像とともに供給される切り出し地図によってネットに与えられる。参照命令は、単一の物体や人物ではなく、複数の物体や人物の集合を参照することもできる。特に、これは命令を使用して、例えばカメラからの距離に基づいて、情景記述に新しい人物や物体を追加するために使用される。命令の引数は、画像からこれまでに抽出された人物や物体を含む切り出し地図によって与えられる。

切り出し地図を作成するために、トップダウン流は、TD流の最後に、各画像の画素について、ターゲット物体に属する画像を表す得点を生成し、閾値を超える画素が選択される。最終的な物体を生成するために、雑音や冗長な領域を取り除く後処理が行われる。雑音の少ない候補をフィルタリングするために、モルフォロジー演算(浸食erosionと拡張dilation)が行われる。選択された画素が複数の物体候補(つながっていない領域)を形成する場合、主に画素の平均得点に基づいて候補をフィルタリングする。

精緻化ネットワーク: このネットワークは、TD命令に従って、特定の項目の選択された特性を認識するように学習される。同じネットワークは、R(x,y)形式の関係についても学習される。入力は2つの物体と関係で構成され、出力は関係の値である(例えば真/偽や、前/後/なしなど、2つの相補的な関係の間の選択である)。入力引数は、関係中の2つの項目を分割する2つの分割地図によって提供される(各地図は1

つの項目に対応する)。技術的には、プロパティの分類では、1つの切り出し地図はクエリされた項目に対応し、2つ目の切り出し地図は無関係(すべてゼロ)である。トップダウン流は、参加する物体切り出し(1つの切り出し地図)を提供するように学習される。

完全な構造を取得する最初の段階は、no component が返されるまで、展開ネットワークを呼び出してすべての情景成分(物体と人物)を抽出し、配列に格納することである。第2段階は、「精緻化ネットワーク」を使ってすべての構成要素のプロパティを照会し、配列内の対応する構成要素の結果を更新することである。プロパティ命令の特定のセットは、項目の型に応じて決定される(例えば「衣服」命令は人物に対してのみ起動される)。

最後の段階は、構成要素間の関係(空間的関係には「右」と「後ろ」を使い、その補完的な関係である「前」と「左」は除外する)を得ることである。各構成要素について、「拡張ネットワーク」は複数回起動され、その都度、異なる関係命令で起動される。特定の関係セットは、成分の型に応じて決定される(たとえば「対面」命令は人物に対してのみ活性化される)。「補助・物体」(ハンマーやハンドバッグなど)は、対応する関係('holding' や 'on')の命令によってのみ取り出され、'extract-next' 命令では取り出されない。取り出された 'auxiliary-object' は配列に追加され、そのプロパティが(精緻化ネットワークを用いて)照会され、同様に追加される。他の関係(「主要」成分を持つ)の場合、検索された成分は、配列で表現される成分の1つでなければならない。配列の中で、IOU の重なりが最大となる成分が選択される。最大 IOU が閾値以下であれば、新しい成分が配列に追加される。アルゴリズムの概略的な「擬似コード」を以下に示す。関数 get_item_by_flag と get_prop_by_flag は、それぞれ展開ネットワークと精緻化ネットワークを呼び出すラッパー関数であることに注意。

Algorithm 1: Image full structure extraction using counter stream estimators

```

Input: image
Result: image graph
initialization: image_graph = [], item_class = 0, items_count = 0, scene_mask = zeros;
begin
    while item_class is -None ∧ items_count < MAX_ITEMS do
        [item_class, pred_mask] = get_item_by_flag(image, scene_mask, flag = 'next_item')a;
        items_count = items_count + 1;
        if item_class is -None then
            scene_mask = scene_mask ∨ pred_mask;
            overlap_ind = get_item_ind_by_overlap(pred_mask, image_graph);
            if overlap_ind is None then
                item['class'] = item_class;
                item['mask'] = pred_mask;
                item['child_nodes'] = [], item['child_rels'] = [];
                item['parent_nodes'] = [], item['parent_rels'] = [];
                for prop_type in prop_listb do
                    item[prop_type] = get_prop_by_flag(image, pred_mask, flag = prop_type)c;
                end
                image_graph.append(item);
            end
        end
    end
    for item_ind in length(image_graph) do
        for rel in rels_listd do
            [rel_item_class, rel_pred_mask] = get_item_by_flag(image, image_graph[item_ind][mask], flag = rel)a;
            child_ind = get_item_ind_by_overlap(rel_pred_mask, image_graph);
            if rel in tool_rels ∨ child_ind is None then
                new_item['class'] = rel_item_class;
                new_item['mask'] = rel_pred_mask;
                new_item['parent_nodes'] = [], new_item['parent_rels'] = [];
                for prop_type in prop_listb do
                    new_item[prop_type] = get_prop_by_flag(image, rel_pred_mask, flag = prop_type)c;
                end
                child_ind = length(image_graph);
                image_graph.append(new_item);
            end
            image_graph[child_ind]['parent_nodes'].append(item_ind);
            image_graph[child_ind]['parent_rels'].append(rel);
            image_graph[parent_ind]['child_nodes'].append(child_ind);
            image_graph[parent_ind]['child_rels'].append(rel);
        end
    end
end

```

^aget_item_by_flag function invokes the 'expansion network'
^bproperty list is adopted according to item type (e.g. 'clothes' is relevant only for persons)
^cget_prop_by_flag function invokes the 'elaboration network'
^drelation list is adopted according to item type (e.g. 'facing' is relevant only for persons)

S8. 情景構造のガイド付き抽出

ガイド付き解釈では、与えられた構造によってガイドされる BU-TD 課題の系列を実行することによって、画像中のクエリ構造の心像性を見つける。本節では、抽出された構造とその画像における接地性を表現するためのデータ構造についてのみ簡単に述べる。目的の構造を抽出するために BU-TD ネットワークを誘導する処理については、次の補足資料(補足 S9 次の命令の選択)で説明する。

この手続きは、2つのデータ構造を維持・更新する：

1. 物体(人物を含む)とその詳細の配列で、BU-TD 活性化の出力に応じて更新される。これは、クエリ画像に接地された検索された構造を表す。
2. ノードは物体を表し、エッジは関係を表す。各ノードは、物体の配列から、このノードの候補として現在有効な対応する物体へのポインタを含む。これらのポインタはテストの結果に応じて更新される。

クエリ処理が終了すると、オブジェクトの配列には、「接地」手順で検出されたすべての検出物体とその対応情報(関連物体のインデックスを含む)が含まれる。この配列は、基本的に抽出された構造(グラフ)であり、「完全な構造」の部分集合であり、クエリ検証に成功した場

合には、クエリ構造の上位集合となる。この配列は、欠落している情報のみが計算される「フォローアップ」問い合わせに使用することができる。

S9. 次の TD 命令の選択：視覚ルーチンの構成

TD 命令を生成するために使用される手順は、一度に 1 つのノードを処理する再帰的処理である（グラフにルートノードがある場合はルートノードから開始し、そうでない場合は任意のノードから開始する）。上記のように、完全な処理は、補足資料 S7（全構造抽出）で上述した 2 つのネットワークモデルの組み合わせによって実行される。1つ目（展開ネットワーク）は、画像内で検出された新しい項目（人物または物体）を追加することで、グラフを拡張する役割を担う。2つ目（精緻化ネットワーク）は、要求された構造に従って、追加的な特性や特定の関係を抽出することにより、既存のグラフを精緻化するために使用される。Vatashsky & Ullman 2020 に詳細に記述されているアルゴリズムに基づいている。

この手続きでは、展開ネットワークを用いて人物と物体を抽出する。Actors データセットでは命令を用いて、人物と情景物体（木やベンチなど）が 1 つずつ抽出される。アルゴリズムはターゲット構造のノードに沿って進み、ターゲットグラフで記述された構造に対応する人物や物体を、その属性や関係とともに画像から抽出するようモデルを導く。抽出された人物のクラスが、グラフ内の現在のノードが必要とするクラス（例えば「少女」）に対応する場合、プロセスは、プロパティと関係の観点からノードの要件をチェックする。関連する特性は、適切な TD プロパティ命令を使用して抽出および検証され、関連する関係は、グラフ内の関係で必要とされる人物およびオブジェクトを検索する拡張ネットワークを使用して抽出および検証される（「対面」など）。空間関係命令（‘right-of’、‘behind’ など）は、人物と情景物体を抽出し、「holding」や‘on’などの関係は、人物が持ち、他のオブジェクトの上に置かれた道具や物体を抽出するために使用される。精緻化ネットワークによるプロパティの抽出は、ターゲットグラフ内のプロパティを検証する（例えば、赤い物体を探す）か、プロパティを検索する（例えば、テーブル上の物体の色を見つける）どちらかに使用される。検索された情報は、各エンタリイーが 1 つの物体（人を含む）のデータを表す配列に保存され、更新される。要件を満たす検索された物体は、ターゲット・グラフの対応する物体と対にされ、後続のテストが正しい物体に適用されるようになる。必要なオブジェクトの数は、量化子（「すべて」、「3つ」など）や、オブジェクト集合全体に依存する特性（「いくつ」など）を評価する必要性によって設定される。ノードのテストがすべて完了すると、深さ優先（DFS）探索によって決定された次のノードに対して、このプロセスが繰り返される。再帰内のノードとそれに続くノードがテストされ、画像が検証された後、必要に応じて、いわゆるノードのセット要件を検証するためのテストが適用される。このようなテストの例としては、セット内の物体のカウントと量の比較がある。再帰的な処理は、グラフ全体が要求されたとおりに画像に接地されたときか、チェックすべき代替案がなくなったときに終了する。例えば、必要なクラスの物体が検出されなかった場合などである。次の TD 命令を選択し、関心のある構造を抽出するアルゴリズムの擬似コード記述を以下に示す。

Algorithm 2: Guided scene structure extraction using counter stream estimators

```

Input: query_graph, image
Result: query answer, retrieved structure graph
initialization: workMem['current_node'] = first_parent_node;
Run [success, answer] = getGraphAnswer();
begin
    current_node = workMem['current_node'];
    Node parameters: p: properties, r: relations, f: property type, g: property of a set;
    last_item, no_items, no_child_items = False; mask = zeros;
    while ~no_items & ~last_item do
        if child_node then
            item = workMem['item'];
            last_item = True;
        else if is_super_nodea then
            | [item, last_item] = get_item_from_saved_sub_nodes
        else if existsaved_detected_items then
            | [item, last_item] = get_item_from_saved_detected
        else
            | [item, no_items] = get_item_by_flag(image, mask, flag = 'next_item')b;
            mask = mask ∨ item['seg'];
        end
        if ~empty(p) then
            for p in p do
                | p̂ = get_prop_by_flag(image, item['seg'], flag = f_p)c;
                success = p̂ == p;
                if ~success then break end
            end
            if ~success then continue end
        end
        if ~empty(f) then answer = get_prop_by_flag(image, item['seg'], flag = f)c end
        if empty(r) then
            if exist(next_parent_node) then
                | workMem['current_node'] = next_parent_node;
                Run [success, answer] = getGraphAnswer();
            end
        else
            for r in r do
                child_mask = item['seg'];
                while ~no_child_items do
                    if empty(item[r_item]) then
                        | [item[r_item], no_child_items, answer] = get_item_by_flag(image, child_mask, flag = r)d;
                    end
                    child_mask = item[r_item]['seg'];
                    success = ~no_child_items;
                    if success then
                        | workMem['current_node'] = next_nodee;
                        workMem['item'] = child_item;
                        Run [success, answer] = getGraphAnswer();
                        if success ∧ (#success_child_items == #required_child_items)e then break else success = False end
                    end
                end
                if ~success then break end
            end
            if success ∧ (#success_items == #required_items)e then break end
        end
        if success ∧ ~empty(g) then answer = g(valid_items) end
        if success ∧ comp_num_en ∧ is_checked(comp_node) then answer = comp_numf(valid_items, comp_items, comp_type); end
        if success ∧ is_sub_node then save_for_super_node(success_items); end
    return [success, answer]
end

```

^aA ‘super node’ includes other nodes ('sub nodes')

^bget_item_by_flag function invokes the ‘expansion network’

^cget_prop_by_flag function invokes the ‘elaboration network’

^dEither child node or next unvisited root node of a subgraph

^eAccording to quantifiers and other requirements

^fCompare number of valid items between nodes ('same', 'fewer', 'more')

S10. 組み合わせによる汎化

組合せ汎化テストでは、2つの課題(人物特性の抽出、EMNIST 文字間の空間関係(left-of, right-of)の計算)について4つのモデルを比較した。これらの比較における4つのモデルを図37(組み合わせモデルの比較)に模式的に示す。主な比較対象は、我々の BU-TD モデル(a)と Unguided, Readout selection, (b) とラベル付けされたモデルである。この2つのモデルの構造はほぼ同じであるが、5.2節組合せ汎化で説明したように、BU-TD モデルは命令によって誘導されるのに対し、Readout selection は誘導されないという決定的な違いがある。BU-TD モデルとは異なり、選択された命令に対して正しい出力を提供するために、読み出し選択モデルは、1回のパスで画像内の人物の心像性をすべて同時に抽出しなければならない。読み出し選択は、どのプロパティを抽出するかを決定せず、最終層からの読み出しのみを選択する。

他の2つのモデルは、ガイド付きモデルの異なるバージョンである。BU 命令、ガイド付きモデル(c)は、BU-TD モデルと同じであるが、課題命令が入力画像とともに BU 方式で与えられる。これは命令から画像と同じ大きさのベクトルへの埋め込みを学習することで行われる。最後のモデルは標準的な ResNet モデル(d)で、これも BU 命令、Guided モデルと同様の方法で、入力レベルで課題命令が与えられる。ResNet は BU-TD モデルと同様の構造、ユニット数、接続を持つが、BU 流と TD 流間の横方向の接続はない。層数は BU-TD と同じだが、パラメータの共有がないため、パラメータ数は約 1.5 倍となる。他のモデルは、BU 命令、Guided モデルの命令埋め込み層のサイズが、BU-TD や Readout 命令モデルの埋め込みよりも大きくなっていることを除けば、パラメータ数は同じである。

すべてのモデルで同じ損失を使用した。1つ目は出現損失(つまり、どのキャラクターや人物が画像内にいるか)、2つ目は課題損失(キャラクターの右/左、または人物の所有物の詳細)である。

BU-TD モデルと同様に、c と d のモデルは、選択された課題を実行する課題命令によって導かれる。c の入力レベルの命令は、BU-TD モデルの命令と同様に、BU 最上層に到達するまで変化することなく BU 流に沿って伝播し、その後 TD 流に進む。ここでは、ガイド付きモデルとガイドなしモデルを区別するためではなく、ガイド付きモデルの異なる変種間の比較を行うために使用した。

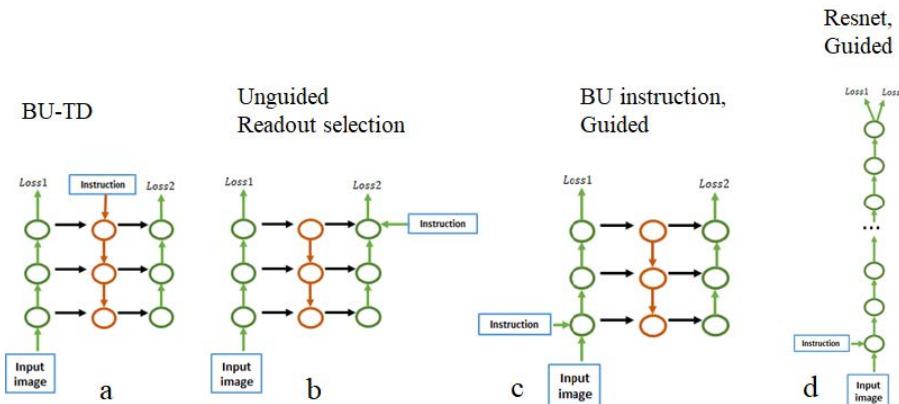


図37 組み合わせモデルの比較。各円は層を表す。

- a. BU-TD モデルのアーキテクチャ。
- b. ガイドなし、読み出し選択。
- c. BU 命令ガイド付き。
- d. ResNet、ガイドあり。主な比較は、BU-TD モデル(a)とガイドなし、読み出し選択(b)である。この2つのモデルの構造はほぼ同じだが、BU-TD モデルは命令によってガイドされ、読み出し選択はガイドされない。

BU-TD とボトムアップ(読み出し)モデルの埋め込みサイズは、EMNIST の実験では 256、Persons 実験では 512 である。BU ガイドモデルと ResNet ガイドモデル(画像サイズに依存)では、埋め込みサイズは $(W/2 \times H/2)$ であり、W と H は画像の幅と高さである。EMNIST の実験では埋め込みサイズは 6,272 であり、Persons 実験では 25,088 である。

実験の詳細

EMNIST と Persons という2つの異なる課題について組み合わせ汎化実験を行った。それぞれの課題に対して、十分サイズと拡張サイズの2つのデータセットサイズを使用した。十分なサイズとは、BU-TD が 85% 以上の精度を達成した最小のデータセットである。異なる実験では、この十分なセットで比較対象の4つのモデル全てを訓練し、異なるモデルによって得られた性能を調べた。訓練セットに加え、訓練セットと同じ分布を持つ検証セット(具体的には、5.2節組み合わせによる汎化で議論した物体-プロパティまたはオブジェクト-関係の「除外された対」を使用しない)と、新しい組み合わせへの汎化をテストするために、除外された対に焦点を当てた組み合わせテストデータを作成した。同様のスキームが拡張セット実験にも用いられた。

我々は、30 以上の最適化手法の組み合わせで、すべてのモデルに対して広範なハイパーパラメータ探索を行った: SGD(確率的勾配降下法)with モメンタム、ADAM(Kingma&Ba2015)、学習率(6 値: 0.0001, 0.001, 0.002, 0.05, 0.1, 0.2)、バッチサイズ(10, 32, 48)、ウェイト減衰(0.0001, 0.0002)。検索は、比較に参加するすべてのモデルに同じハイパーパラメータのセットを使用するという意味で「公平」であった。比較に参加したモデルが BU-TD モデルの性能に達しなかった場合、我々はさらにそのようなモデルについて、バッチ正規化(Ioffe&Szegedy2015)や群正規化(Wu&He2020)を含む追加学習の選択肢を検索し、群正規化のハイパーパラメータを検索した。また、Lookahead 最適化(Zhang+2019), AdamW, Adamax (Loshchilov&Hutter2019)などの様々な最適化手法も試した。

組み合わせ汎化 - Persons データセット

Persons データセットでは、各画像に 5 つの属性を持つ 2 人の人物がいる。従って、1 つの画像につき複数の訓練例を生成し、人物と特性のペアに 1 つずつ対応させた。Sufficient データセットには 1600 の学習画像があり、そこから 22,400 の学習例を生成した。さらに 800 (非組み合わせ) の検証例と 1176 (組み合わせ) のテスト例があった。Extended データセットでは、4800 の訓練画像と 6 万 7144 の例があり、さらに 1198 の(非組み合わせ的な) 検証例と 1176 のテスト例があった。

結果。各モデルについて、精度、エポック数、表示例数(データセット中の例数×エポック数、括弧内は千単位)を報告する。精度は、以下の収束基準に達した後に測定される：エポック E における精度 A を報告し、その後の 50 エポックにおける精度が最大 2% 増加する。以下の表 5 に Persons 実験の結果を示す。

	BU-TD	Unguided	BU Instructed	ResNet, Guided
	Readout selection	on, Guided		
Sufficient data, Non generalization	89% 187 (4,188)	68% 48 (1,075)	55% 13 (291)	80% 263 (5,891)
Sufficient data, Generalization	89% 181 (4,054)	31% 71 (1,590)	22% 47 (1,052)	65% 573 (12,835)
Extended data, Non generalization	97% 53 (3,558)	96% 108 (7,251)	98% 91 (6,110)	98% 100 (6,714)
Extended data, Generalization	93% 71 (4,767)	29% 213 (14,301)	88% 94 (6,311)	96% 121 (8,124)

表 5 : Persons データセットにおける組み合わせ汎化。各テストについて、精度を一番上の行に、エポック数と表示例数(括弧内)を二番目の行に示す。主な結果は、非ガイド付きモデル(読み出し選択)は BU-TD モデルに比べて組み合わせ汎化が悪いということである。対照的に、非ガイド付きモデルは、拡張データによる非組合せ汎化では同等の精度に達する。追加された 2 つのガイド付きモデルは、拡張データセットで学習した場合、高い組み合わせ汎化に達する。

組合せ汎化 - EMNIST データセット

EMNIST Sufficient データセットでは、1 つの画像につき 10 個の例を生成し、100,000 個の学習例と、さらに 500 個の検証例と 500 個のテスト例を生成した。Extended データセットでは、各画像に 24 文字が含まれる。1 つの画像につき、right-of 課題用に 20 例、left-of 課題用に 20 例、合計 40 例を生成した。10,000 の画像から 400,000 の学習例、さらに 500 の検証例と 500 のテスト例が得られた。結果を以下の表 6 に示す。

	BU-TD	Unguided	BU Instruction, Readout selection	ResNet, Guided
Sufficient data, Non generalization	96% 39 (3,900)	48% 537 (53,700)	5% 2 (200)	1% 1 (100)
Sufficient data, Generalization	95% 42 (4,200)	14% 635 (63,500)	8% 81 (8,100)	4% 1 (100)
Extended data, Non generalization	95% 18 (7,200)	75% 294 (117,600)	95% 136 (54,400)	95% 15 (6,000)
Extended data, Generalization	91% 15 (6,000)	10% 135 (54,000)	96% 145 (58,000)	95% 17 (6,800)

表6：EMNIST-24 データセットにおける組み合わせ汎化 各テストについて、精度を一番上の行に、エポック数と表示例数(括弧内)を二番目の行に示す。主な結果は、非ガイド付きモデル(読み出し選択)は、BU-TD モデルと比較して組み合わせ汎化が悪いということである。追加された 2 つのガイド付きモデルは、拡張データセットで訓練した場合、組み合わせ汎化において高い精度に達する。

結果のパターンは 2 つのデータセットで類似している。主な比較対象は、課題指示によって誘導される BU-TD モデルと、類似しているが誘導されない BU モデルである。主な結果は、非ガイドの BU モデルは組み合わせ汎化が著しく制限されていることである。他の形式のガイド付きモデルも組み合わせ汎化において高い精度を達成するが、BU-TD モデルに比べて学習に時間がかかる。

最後に、EMNIST-6 と呼ばれる、1 画像あたり 6 文字の、より単純な課題の汎化実験も行った。この実験の目的は、課題が単純化され、訓練が十分に延長されたときに、組み合わせによる汎化が、偶然をはるかに超えて、無ガイドモデルで出現し始めるかどうかをテストすることである。文字数を 24 文字から 6 文字に減らした EMNIST 課題の結果を以下に示す。EMNIST-6 Sufficient データセットでは、1 つの画像につき 5 つの例を生成した(右端の文字を除く)。2,000 枚の画像を使用し、10,000 個の学習例と、さらに 500 個の検証例と 500 個のテスト例を作成した。拡張データセットでは、10,000 枚の画像を使用し、50,000 個の学習例と、500 個の検証例と 500 個のテスト例を追加した。表 7 に見られるように、課題が単純でデータセットが十分大きい場合、ガイドなしモデルは組み合わせ汎化を達成できる(精度は落ちるが)。

	BU-TD	Unguided	BU	ResNet, Guided
	Readout selection	Instructi on, Guided		
Sufficient data, Non generalization	84% 95 (950)	8% 15 (150)	8% 15 (150)	81% 815 (8,150)
Sufficient data, Generalization	89% 142 (1,420)	13% 194 (1,940)	18% 36 (360)	89% 923 (9,230)
Extended data, Non generalization	95% 50 (2,500)	80% 203 (10,150)	95% 38 (1,900)	95% 25 (1,250)
Extended data, Generalization	94% 43 (2,150)	77% 551 (27,550)	95% 45 (2,250)	96% 56 (2,800)

表 7 : EMNIST-6 データセットにおける組合せ汎化 各テストについて、精度を一番上の行に、エポック数と表示例数(括弧内)を二番目の行に示す。

非ガイドモデルでは、Readout 選択の代わりに並列学習モデルもテストした。EMINST-6 を標準的な BU モデルで学習し、すべての右隣と一緒に予測した。 $M(i,j)$ は文字 j が文字 i の右隣であることを意味する。 $M(i,j)$ は文字 j が文字 i の右隣であることを意味する。この表現では、訓練から除外されたペアを表す出力ユニットが訓練中に活性化されることはなく、したがって正しい出力ユニットの活性化は期待できない。我々は、この最後の層からの読み出しを訓練することで、正解が実際にネットワークの出力層に符号化されているかどうかをテストした。読み出し訓練は、メインネットワークの完全な訓練の後にも行われた。この読み出しネットワークを用いて、並列学習バージョンは逐次選択読み出しと同じ汎化精度に達することがわかった。この比較から、選択的読み出しネットワークは、すべての異なる課題の予測を並列に生成するように訓練されたネットワークと同等であるが、訓練中に除外された対の結果を、より標準的なテーブル表現や多分岐アーキテクチャでは不可能な方法で読み出すことができるという利点があることがわかる。

さまざまな課題における組み合わせ汎化の結果から、非ガイドモデルではガイドモデルに比べて組み合わせ汎化が著しく制限されることがわかった。空間関係課題では、文字数を 24 文字から 6 文字に減らすと、組み合わせ汎化が可能になった。文字数を 6 文字のままにして課題を複雑にすると、非ガイドモデルではガイドモデルに比べて組み合わせ汎化が減少した。例えば、left-of 関係(選択された文字の right-of か left-of のどちらか)を追加すると、非ガイド付きモデルの組み合わせ汎化は 77% に減少した(BU-TD の 95%)。除外される対の割合を 63% に増やすと、非ガイドモデルは 3% に激減した(対 BU-TD の 82%)。

上記のすべての実験から、非ガイドモデルでは組み合わせによる汎化が不可能ではないが、ガイドモデルに比べて大きく制限されるという結論に達した。ガイド付きモデルとガイドなしモデルの間のギャップは、課題がより複雑になったとき(例えば、追加の関係、より多くの物体)、または学習中に除外される組み合わせの割合が減少したときに増大する。情景の解釈という課題において、学習された組み合わせの数が全ての可能な組み合わせのほんの一部にしかなり得ない場合、ガイド付きモデルが唯一の実現可能な選択肢となる可能性がある。