マルチモーダル言語理解タスクにおける Dual ProtoNCE に基づく転移学習

A Transfer Learning Method for the Multimodal Language Understanding Based on Dual ProtoNCE

小槻 誠太郎 *1　石川 慎太朗 *1　杉浦 孔明 *1
Seitaro Otsuki　Shintaro Ishikawa　Komei Sugiura

*1慶應義塾大学
Keio University

We focus on the task of identifying target objects in domestic environments according to free-form natural language instructions. In this study, we propose a novel transfer learning approach for multimodal language understanding, Prototypical Contrastive Transfer Learning (PCTL) which uses a new contrastive loss, Dual ProtoNCE. Our experiment demonstrated that PCTL outperformed existing methods.

1. はじめに

高齢化が進展する現代社会において，在宅介助者の不足が深刻な社会問題となっている．こうした社会問題に対する解決策の一つとして生活支援ロボットが挙げられるが，生活支援ロボットが人間と自然な対話をする能力は未だ不十分である．こうしたロボットの言語理解モデルの学習には，実際の環境で収集されたデータを利用することが望ましい．しかし，実世界データの収集及びアノテーションには大きなコストがかかる．それに対し，シミュレーション上であれば，実世界で行うよりも低いコストでデータを収集できるため，シミュレーションデータを用いた学習は便利である．

本研究では自然言語による命令文と状況を説明する画像が与えられた際，正しく命令内容を解釈し，対象物体を特定するタスクを扱う．例えば“Bring me the book closest to the lamp.”という命令文及びいくつかの本と一つのランプが置いてある画像が与えられた際，その中からランプに最も近い本を対象物体として予測する事が望ましい．

本研究において，我々はシミュレーションデータから獲得した経験を転移させ，実環境データにおけるモデルの性能を向上させることを目的とする．物体操作に関するマルチモーダル言語理解タスクにおいて，多くの既存手法は実環境データのみを用いて学習されている [Hatori 18, Magassouba 19, Ishikawa 21]．しかし，こうした手法においてデータセットを大規模化することは，実環境データの収集コストが高いため容易ではない．それに対して，シミュレータを利用すると，実環境データの収集に比べて非常に低いコストで学習用データを収集することができる．結論として転移学習によってシミュレーションデータを利用することにより，モデルの効率的な性能向上が期待される．

我々はマルチモーダル言語理解タスクにおける新しい転移学習手法として Prototypical Contrastive Transfer Learning (PCTL) を提案する．PCTL は新たに設計された対比損失である Dual ProtoNCE によって，転移元ドメインのデータと転移先ドメインのデータの間で対照学習を行う．そのため，Dual ProtoNCE の最小化によって，転移元ドメインと転移先ドメインの間の差異の影響を低減することが期待される．

2. 関連研究

Uppal ら [Uppal 22] はマルチモーダル言語理解言語における最近の動向を紹介しており，タスクの定式化，評価指標，モデル構造，そしてバイアスや公平性，敵対的攻撃などのその他の話題を説明している．マルチモーダル言語理解タスクの一つに referring expression comprehension (REC) があり，既存研究で取り組まれている [Yu 18]．REC タスクにおいて，モデルは参照表現で説明された物体を接地することを求められる．一方，我々のタスクは REC タスクよりも少し柔軟に定式化される．具体的には，候補物体が対象物体であるかという 2 値分類として定式化されるため，対象物体が複数与えられる場合や一つも与えられない場合を扱うことが出来る．

画像と自然言語の命令文から対象物体を特定するモデルの研究は複数存在する [Magassouba 19,Ishikawa 21]．特に Target-Dependent UNITER [Ishikawa 21] (TDU) は画像と自然言語の関係性の学習に UNITER 型注意機構を導入し，汎用事前学習モデルの利用を可能にしたモデルである．また，小槻ら [小槻 22] は画像中の領域間の関係性のモデル化，及び画像中の領域群と命令文の特徴量の関係のモデル化を行う TDP-MAT を提案している．HLSM-MAT [Ishikawa 22] は，敵対的な摂動を特徴量空間に加える Moment-based Adversarial Training (MAT) を提案し，vision-and-language navigation (VLN) において，サブゴールおよび状態表現の特徴量空間に対して適用している．

3. 問題設定

本論文で用いる用語として対象物体，候補物体，コンテキスト物体，対象領域，候補領域，コンテキスト領域の定義は [小槻 22] に従う．本論文では，Multimodal Language Understanding for Fetching Instruction (MLU-FI) タスクに取り組む．本タスクでは，命令文と，それに対応する画像から物体検出器によって抽出された候補領域，コンテキスト領域群が与えられ，候補物体が対象物体であるか否かの 2 値分類を行うことが求められる．MLU-FI における入力と出力は以下である．

- 入力： 命令文，候補領域，コンテキスト領域
- 出力： 候補物体が対象物体である確率の予測値 $p(\hat{y}=1)$

ここで y はラベル，$\hat{y}$ はその予測であり，$y=1$ は候補物体が対象物体であることを示す．

図 2 に対象タスクで与えられる画像の代表例を示す．命令文は“Look in the left wicker vase next to the potted plant on the second floor at the foot of the stairs.”である．赤と緑のバウンディングボックスはそれぞれ候補領域と対象領域であり，青いバウンディングボックスは残りのコンテキスト領域群である．ここでは対象物体は緑色のバウンディングボックスに囲まれた花瓶であり，与えられた候補領域が対象物体を囲んでいないため，$p(\hat{y}=1)=0$ となることが望ましい．

本研究では候補領域及びコンテキスト領域を物体検出器によって抽出することを前提とする．また，本タスクでは評価尺度として分類精度を使用する．

連絡先: 小槻誠太郎，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，otsu8sei14@keio.jp

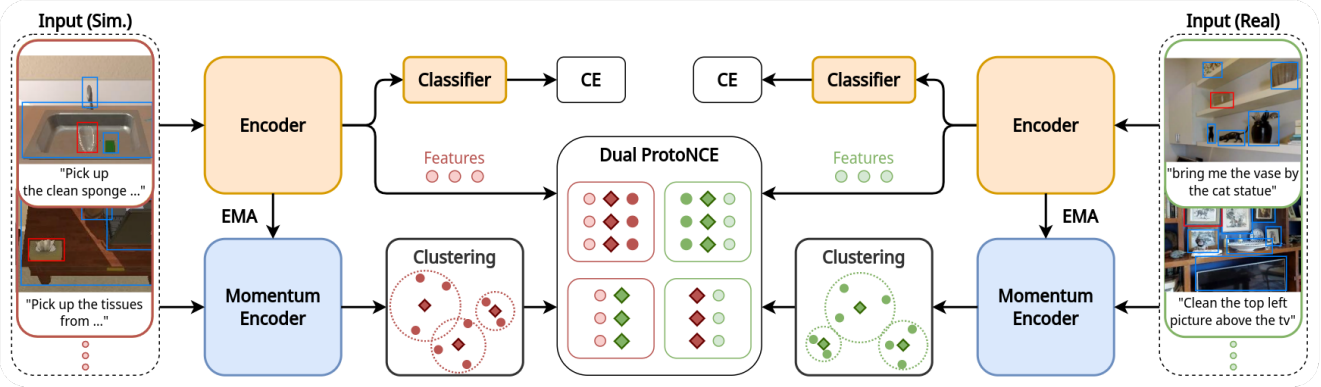

図 1: 学習フレームワーク図

本研究において，転移学習を行うにあたって転移先ドメインのデータは実世界で，転移元ドメインのデータはシミュレーションによって収集される．これらのデータはいずれも自然言語による命令文と視覚的データのペアで構成されており，室内環境における物体操作に関するマルチモーダル言語理解タスクに向けて収集されている．

4. 提案手法

本研究ではマルチモーダル言語理解タスクにおける Dual ProtoNCE に基づく転移学習手法，PCTL を提案する．本研究では提案手法を物体操作に関するマルチモーダル言語理解タスクに対して適用するが，提案手法はより広くマルチモーダル言語理解タスクにおける転移学習に応用できると考えられる．図 1 に PCTL の学習フレームワークの外観を示す．提案フレームワークは Encoder，Momentum Encoder，Clustering Module の 3 つのモジュールに大別される．

ネットワークへの入力 $\boldsymbol{x}$ を次のように定義する．

$$\boldsymbol{x} = \{\boldsymbol{x}_{\text{inst}}, \boldsymbol{x}_{\text{cand}}, \boldsymbol{X}_{\text{cont}}\} \tag{1}$$

$$\boldsymbol{X}_{\text{cont}} = \{\boldsymbol{x}_{\text{cont}}^{(i)} | i = 1, \ldots, N_{\text{det}}\}. \tag{2}$$

上式において $\boldsymbol{x}_{\text{inst}}$，$\boldsymbol{x}_{\text{cand}}$ そして $\boldsymbol{x}_{\text{cont}}^{(i)}$ はそれぞれ命令文，候補領域の特徴量，i 番目のコンテキスト領域の特徴量を表し，N_{det} は Faster R-CNN [Ren 17] によって入力画像から検出された領域の数を表す．なお $\boldsymbol{x}_{\text{cand}}$，$\boldsymbol{X}_{\text{cont}}$ 及び $\boldsymbol{x}_{\text{inst}}$ から抽出した言語特徴量に対しては，positional encoding を付与する．$\boldsymbol{x}_{\text{cand}}$ 及び $\boldsymbol{X}_{\text{cont}}$ に対する positional encoding は，バウンディングボックスの左上の頂点の座標を (x_1, y_1)，右下の頂点の座標を (x_2, y_2) とするとき，$[x_1, y_1, x_2, y_2, x_2 - x_1, y_2 - y_1, (x_2 - x_1) \cdot (y_2 - y_1)]^T$ という 7 次元ベクトルを使用した．ここで (x_1, y_1) 及び (x_2, y_2) は入力画像の幅及び高さによって正規化されていると仮定する．

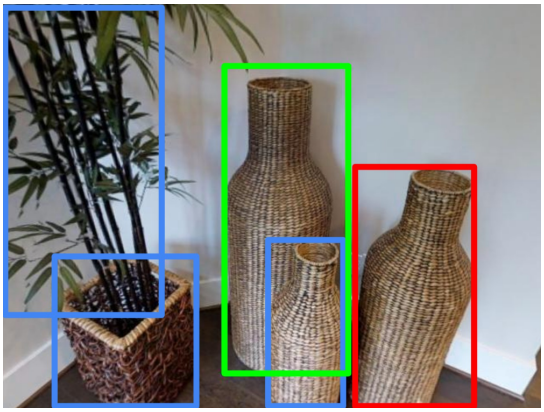

図 2: MLU-FI タスクの代表例

4.1 Model Architecture

Encoder $f_{\boldsymbol{\theta}}$ は $\boldsymbol{x}$ を入力にとり 768 次元の特徴量を出力するモデルであり，$f_{\boldsymbol{\theta}}$ の構造は Target-Dependent UNITER [Ishikawa 21] (TDU) と同一である．$f_{\boldsymbol{\theta}}$ の出力は TDU の Multi-Layer Transformer の最終層の出力のうち，$\boldsymbol{x}_{\text{cand}}$ を入力した位置に対応する特徴量である．

本章以降，転移元ドメインのサンプルを $(\boldsymbol{x}_{\text{s}}, y_{\text{s}})$ と表し，同様に転移先ドメインのサンプルを $(\boldsymbol{x}_{\text{t}}, y_{\text{t}})$ と表す．我々は本モジュールの入出力を以下のように記す．

$$\boldsymbol{u} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{s}}) \in \mathbb{R}^{768}, \ \boldsymbol{v} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{t}}) \in \mathbb{R}^{768}. \tag{3}$$

$\boldsymbol{u}$ 及び $\boldsymbol{v}$ はそれぞれ転移元ドメイン及び転移先ドメインの特徴量ベクトルである．これらはクラスタリングや損失に利用される．Classifier g は 2 層の MLP 及び softmax 関数で構成されるモデルである．g は $f_{\boldsymbol{\theta}}$ の出力を受け取り，予測確率 $p(\hat{y} = 1) = g(f_{\boldsymbol{\theta}}(x))$ を計算する．

Momentum Encoder $f_{\boldsymbol{\theta}'}$ は $f_{\boldsymbol{\theta}}$ と同じ構造を持ち，$f_{\boldsymbol{\theta}}$ のパラメータ $\boldsymbol{\theta}$ の指数移動平均 $\boldsymbol{\theta}'$ をパラメータとして持つモデルである．式 (3) と同様に入力 $\boldsymbol{x}_{\text{s}}$ 及び $\boldsymbol{x}_{\text{t}}$ に対する $f_{\boldsymbol{\theta}'}$ の出力を $\boldsymbol{u}'$，$\boldsymbol{v}'$ と記す．

4.2 Clustering

Clustering Module は各エポックの直前に $\boldsymbol{v}'$，$\boldsymbol{u}'$ に対して k 近傍法によるクラスタリングを，クラスタ数を変えて M 回行う．m 回目のクラスタリングにおけるクラスタ数を $k^{(m)}$ と記す．また，m 回目の $\boldsymbol{v}'$，$\boldsymbol{u}'$ に対するクラスタリングで得た i 番目のクラスタの重心をそれぞれ $\boldsymbol{c}_i^{(m)}$，$\boldsymbol{d}_i^{(m)}$ とし，これを i 番目のクラスタのプロトタイプと呼ぶ．

4.3 Contrastive Transfer Learning

Contrastive Transfer Learning では転移学習のために一般化された対比損失である Dual ProtoNCE を計算し，転移元ドメインと転移先ドメインの差異を考慮した学習を行う．

4.3.1 InfoNCE

教師なし表現学習は，学習データ集合 $\boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}$ が与えられた際，$\boldsymbol{A}$ を $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$ に写像するモデル f を学習することを目的とする．このとき，$\boldsymbol{z}_i = f(\boldsymbol{a}_i)$ が $\boldsymbol{a}_i$ を最もよく表現するように学習されることが望ましい．対照学習は InfoNCE [Oord 18, He 20] に代表される対比損失を最小化することでこの目的を達成する手法である．InfoNCE は次のように定式化される．

$$\mathcal{L}_{\text{InfoNCE}} = -\sum_{i=1}^{n} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_i'/\tau)}{\sum_{j \in J} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_j'/\tau)} \tag{4}$$

ここに $J = \{i, n+1, n+2, \ldots, n+r\}$ である．上式において $\boldsymbol{z}_i$ はアンカーの，$\boldsymbol{z}'_i$ は正例の特徴量であり，$\{\boldsymbol{z}'_j \mid j = n+1, n+2, \ldots, n+r\}$ は r 個の負例の特徴量である．また，τ は温度パラメータである．我々は [Radford 21] に従い τ を学習可能パラメータとし，$1/\tau = 0.07$ となるように初期化する．さらに学習を安定させるため，$1/\tau$ が 100 より大きくならないよう数値をクリップする．

4.3.2 ProtoNCE

ProtoNCE は特徴量の表現をクラスタのプロトタイプに近づけるように InfoNCE を一般化した対比損失である．特徴量 $\boldsymbol{z}'$ に関する m 回目のクラスタリングによって得られた i 番目のプロトタイプを $\boldsymbol{h}_i^{(m)}$ として，ProtoNCE は次のように定式化される．

$$\begin{aligned}\mathcal{L}_{\text{ProtoNCE}} = -\sum_{i=1}^{n}\Bigg(&\log\frac{\exp(\boldsymbol{z}_i\cdot\boldsymbol{z}'_i/\tau)}{\sum_{j\in J}\exp(\boldsymbol{z}_i\cdot\boldsymbol{z}'_j/\tau)} \\ &+\frac{1}{M}\sum_{m=1}^{M}\log\frac{\exp(\boldsymbol{z}_i\cdot\boldsymbol{h}_s^{(m)}/\phi_s^{(m)})}{\sum_{j\in J'}\exp(\boldsymbol{z}_i\cdot\boldsymbol{h}_j^{(m)}/\phi_j^{(m)})}\Bigg), \\ J'\subset\{1,2,\ldots,k^{(m)}\}, &s\in J', r' = |J'\setminus\{s\}|.\end{aligned} \tag{5}$$

上式において $\boldsymbol{h}_s$ は $\boldsymbol{z}_i$ に最も近い正のプロトタイプを表す．$\left\{\boldsymbol{h}_j^{(m)} \mid j \in J\setminus\{s\}\right\}$ は $\boldsymbol{h}_s$ を除いたプロトタイプからランダムに選出された r' 個の負のプロトタイプである．ϕ はクラスタ中のインスタンスがプロトタイプ $\boldsymbol{h}$ に集中している度合いを表現する．ϕ の定義は [Li 21] を参照されたい．以降 ϕ を concentration factor と呼ぶ．

4.3.3 Dual ProtoNCE

我々が提案する Dual ProtoNCE $\mathcal{L}_{\text{DualProtoNCE}}$ は，Intra-Domain 損失 $\mathcal{L}_{\text{Intra}}$ と Inter-Domain 損失 $\mathcal{L}_{\text{Inter}}$ の和として定義される．

我々はまず転移元ドメインのデータ及び転移先ドメインのデータに対して独立に ProtoNCE を適用し，$\mathcal{L}_{\text{Intra}} = \mathcal{L}_{\text{Target}} + \mathcal{L}_{\text{Source}}$ を計算する．ここに $\mathcal{L}_{\text{Target}}$及び $\mathcal{L}_{\text{Source}}$ は次のように定義される．

$$\begin{aligned}\mathcal{L}_{\text{Target}} = \sum_{i=1}^{n}-\Bigg(&\log\frac{\exp({}^{\dagger}\boldsymbol{v}_i\cdot{}^{\dagger}\boldsymbol{v}'_i/\tau)}{\sum_{j\in J}\exp({}^{\dagger}\boldsymbol{v}_i\cdot{}^{\dagger}\boldsymbol{v}'_j/\tau)} \\ &+\frac{1}{M}\sum_{m=1}^{M}\log\frac{\exp({}^{\dagger}\boldsymbol{v}_i\cdot{}^{\dagger}\boldsymbol{c}_s^{(m)}/\phi_s^{(m)})}{\sum_{j\in J'}\exp({}^{\dagger}\boldsymbol{v}_i\cdot{}^{\dagger}\boldsymbol{c}_j^{(m)}/\phi_j^{(m)})}\Bigg)\end{aligned} \tag{6}$$

$$\begin{aligned}\mathcal{L}_{\text{Source}} = \sum_{i=1}^{n}-\Bigg(&\log\frac{\exp({}^{\dagger}\boldsymbol{u}_i\cdot{}^{\dagger}\boldsymbol{u}'_i/\tau)}{\sum_{j\in J}\exp({}^{\dagger}\boldsymbol{u}_i\cdot{}^{\dagger}\boldsymbol{u}'_j/\tau)} \\ &+\frac{1}{M}\sum_{m=1}^{M}\log\frac{\exp({}^{\dagger}\boldsymbol{u}_i\cdot{}^{\dagger}\boldsymbol{d}_s^{(m)}/\varphi_s^{(m)})}{\sum_{j\in J'}\exp({}^{\dagger}\boldsymbol{u}_i\cdot{}^{\dagger}\boldsymbol{d}_j^{(m)}/\varphi_j^{(m)})}\Bigg).\end{aligned} \tag{7}$$

上式において ${}^{\dagger}\boldsymbol{a}$ は $\boldsymbol{a}/\|\boldsymbol{a}\|_2$ を表す．また，ϕ は c の，φ は d の concentration factor である．さらに我々は 2 つのドメイン間のギャップを解決するため，$\mathcal{L}_{\text{Inter}} = \mathcal{L}_{\text{S2T}} + \mathcal{L}_{\text{T2S}}$ を計算する．$\mathcal{L}_{\text{S2T}}$ は転移元ドメインの特徴量 $\boldsymbol{u}$ と転移先ドメインのプロトタイプ $\boldsymbol{c}$ の間に定義される対比損失，$\mathcal{L}_{\text{T2S}}$ は転移先ドメインの特徴量 $\boldsymbol{v}$ と転移元ドメインのプロトタイプ $\boldsymbol{d}$ の間に定義される対比損失であり．それぞれ次のように定式化される．

$$\mathcal{L}_{\text{S2T}} = -\frac{1}{M}\sum_{i=1}^{n}\sum_{m=1}^{M}\log\frac{\exp({}^{\dagger}\boldsymbol{u}_i\cdot{}^{\dagger}\boldsymbol{c}_s^{(m)}/\phi_s^{(m)})}{\sum_{j\in J'}\exp({}^{\dagger}\boldsymbol{u}_i\cdot{}^{\dagger}\boldsymbol{c}_j^{(m)}/\phi_j^{(m)})},$$

$$\mathcal{L}_{\text{T2S}} = -\frac{1}{M}\sum_{i=1}^{n}\sum_{m=1}^{M}\log\frac{\exp({}^{\dagger}\boldsymbol{v}_i\cdot{}^{\dagger}\boldsymbol{d}_s^{(m)}/\varphi_s^{(m)})}{\sum_{j\in J'}\exp({}^{\dagger}\boldsymbol{v}_i\cdot{}^{\dagger}\boldsymbol{d}_j^{(m)}/\varphi_j^{(m)})}.$$

我々が最終的に最小化する損失関数 $\mathcal{L}$ は以下である．

$$\mathcal{L} = \lambda\mathcal{L}_{\text{DualProtoNCE}} + \mathcal{L}_{\text{t}} + \mathcal{L}_{\text{s}} \tag{8}$$

$$\begin{aligned}\mathcal{L}_{\text{t}} = \sum_{i=1}^{n}\Bigg(&\mathcal{L}_{\text{CE}}(g(f_\theta(\boldsymbol{x}_{\text{t}}^{(i)})), y_{\text{t}}^{(i)}) \\ &+\frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_{\text{CE}}(g(\boldsymbol{c}_s^{(m)}), y_{\text{t}}^{(i)})\Bigg)\end{aligned} \tag{9}$$

$$\begin{aligned}\mathcal{L}_{\text{s}} = \sum_{i=1}^{n}\Bigg(&\mathcal{L}_{\text{CE}}(g(f_\theta(\boldsymbol{x}_{\text{s}}^{(i)})), y_{\text{s}}^{(i)}) \\ &+\frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_{\text{CE}}(g(\boldsymbol{d}_s^{(m)}), y_{\text{s}}^{(i)})\Bigg).\end{aligned} \tag{10}$$

なお $\mathcal{L}_{\text{CE}}$ は交差エントロピー誤差を表し，λ はハイパーパラメータである．また，$\boldsymbol{c}_s^{(m)}$ 及び $\boldsymbol{d}_s^{(m)}$ は特徴量 $f_\theta(\boldsymbol{x}_{\text{t}}^{(i)})$ 及び $f_\theta(\boldsymbol{x}_{\text{s}}^{(i)})$ にそれぞれ最も近いプロトタイプである．

5. 実験

5.1 データセット

本研究では多様な実世界の室内環境における，自由視点の MLU-FI タスクに焦点を当てる．一方このタスクの標準データセットとして実環境で収集されたものは存在しないため，我々は必要なデータを REVERIE データセット [Qi 20] より収集することで新しく REVERIE-fetch データセットを作成した．

本研究では転移元ドメインのデータを ALFRED [Shridhar 20] データセットから抽出し，ALFRED-fetch-b データセットを作成した．ALFRED データセットは，シミュレーション環境で収集された Vision-and-Language Navigation (VLN) タスクのためのデータセットである．ALFRED-fetch-b データセットを構成するデータは ALFRED データセットの訓練集合から抽出された命令文及び画像である．

これらの REVERIE データセット及び ALFRED データセットから抽出したデータに対する事前処理として，Faster R-CNN を使用して画像から候補領域及びコンテキスト領域群を抽出し，サンプルを作成した．またサンプルのうち，対象領域と候補領域の GIoU [Rezatofighi 19] が 0.80 より大きいものを正例，0.45 より小さいものを負例とした．ここで対象領域は元データセットで与えられている．そしてテスト集合に関しては物体検出器の検出失敗に起因する不適切な例を除去した．

REVERIE-fetch データセットは訓練集合に 8302，検証集合に 994，テスト集合に 947 サンプルの合計 10243 サンプルのデータを含む．ALFRED-fetch-b データセットは訓練集合に 27492，検証集合に 3470，テスト集合に 3324 サンプルの合計 34286 サンプルのデータを含む．

5.2 パラメータ・学習設定

PCTL のハイパーパラメータを次のように設定する．

$$(r, r') = (32, 32),\ k^{(1)} = 64,\ \lambda = 1/32,\ \tau' = 0.2,\ \alpha = 10$$

$f_{\boldsymbol{\theta}}$ が持つ Multi-Layer Transformer の層数は 12，隠れ層の次元数は 768，Attention の Head 数は 12 である．また，使用したモデルは 1 億 1 千万個の学習可能パラメータを持つ．Optimizer として SGD(momentum=0.9) を利用し，Multi-Layer-Transformer の学習率を 8×10^{-5}，その他のモジュールの学習率を 8×10^{-4}，バッチサイズを 64 とし，30 エポックの学習を行った．本実験ではメモリ 24GB 搭載の GeForce RTX 3090 及びメモリ 64GB 搭載の Intel Core i9-10900KF を使用した. 学習にはおよそ 2 時間を要し，推論には約 59.3ms/sample を要した．

学習の際，各エポック終了時に検証集合での損失関数の値を評価し, REVERIE-fetch データセットの検証集合における損失関数 $\mathcal{L}_{\text{CE}}$ の値が最も低いときの，テスト集合における精度を，最終的な精度とした．

表 1: REVERIE-fetch データセットにおける定量的結果

Method	Accuracy [%]
Target domain only	73.0 ± 1.87
Fine-tuning	73.4 ± 11.8
MCDDA+ [Saito 18]	74.9 ± 3.94
Ours	**78.1 ± 2.49**

5.3 実験結果

表 1 に REVERIE-fetch データセットにおける各手法の精度を示す．左側が手法，右側が手法に対応する精度であり，精度の欄には 5 回の試行における平均値と標準偏差を示す．

ベースラインとして以下の 3 つの手法を用意した．

- Target domain only: 転移先ドメインのデータのみで学習を行う．
- Fine-tuning: 転移元ドメインのデータによる pretraining を行った後，転移先ドメインのデータによる fine-tuning を行う．
- MCDDA+: [Saito 18] で提案された教師なし転移学習手法である MCDDA を，転移先ドメインの教師データを利用するよう拡張して適用する．

ベースライン (i) 及び (ii) は，それぞれ転移元ドメインのデータを一切使用しない場合，転移元ドメインのデータで pretraining を行う場合に対して提案手法を比較するために設定した．[Saito+, CVPR18] において MCDDA は画像分類データセットにおける転移学習手法として良好な結果が報告されている．そのため，我々は MCDDA を教師あり転移学習設定に拡張した MCDDA+をベースライン (iii) に設定した．

表 1 に示すように，提案手法の精度は全てのベースラインを上回った．具体的にはベースライン (i), (ii), (iii) の精度はそれぞれ 73.0%，73.4%，74.9%であり，提案手法の精度 78.1%はそれぞれを 5.1 ポイント，4.7 ポイント，3.2 ポイント上回った．なお，提案手法とベースライン手法 (i) との性能差は統計有意であった（$p < 0.01$）．

定性的結果を図 3 に示す．図において，赤と緑のバウンディングボックスはそれぞれ候補領域と対象領域である．命令文は ”Go down the stairs to the lower balcony area and turn off the lamp on the dresser ”であり，対象物体は衣装棚の上に位置するランプである．ベースライン手法 (i) は候補物体が対象物体ではないと予測してしまっている一方，提案手法は正確に候補物体が対象物体であると予測している．

6. おわりに

本研究ではマルチモーダル言語理解タスクにおける Dual ProtoNCE に基づく転移学習手法を提案した．本研究の貢献は以下である．

- 物体操作に関するマルチモーダル言語理解タスクに転移学習を導入した．
- ProtoNCE [Li 21] を拡張した転移学習手法である PCTL を提案した．
- PCTL において転移学習のために一般化された対比損失である Dual ProtoNCE を提案した．
- REVERIE-fetch データセットにおける MLU-FI の精度において，PCTL はベースライン手法を上回った．

謝辞

本研究の一部は，JSPS 科研費 20H04269，JST ムーンショット，NEDO の助成を受けて実施されたものである．

参考文献

[Hatori 18] Hatori, J., Kikuchi, Y., Kobayashi, S., et al.: Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions, in *ICRA*, pp. 3774–3781 (2018)

図 3: REVERIE-fetch データセットにおける定性的結果

[He 20] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R.: Momentum Contrast for Unsupervised Visual Representation Learning, in *CVPR*, pp. 9729–9738 (2020)

[Ishikawa 21] Ishikawa, S. and Sugiura, K.: Target-dependent UNITER: A Transformer-Based Multimodal Language Comprehension Model for Domestic Service Robots, *RA-L*, Vol. 6, No. 4, pp. 8401–8408 (2021)

[Ishikawa 22] Ishikawa, S. and Sugiura, K.: Moment-based Adversarial Training for Embodied Language Comprehension, in *ICPR*, pp. 4139–4145 (2022)

[Li 21] Li, J., Zhou, P., et al.: Prototypical Contrastive Learning of Unsupervised Representations, in *ICLR* (2021)

[Magassouba 19] Magassouba, A., Sugiura, K., et al.: Understanding Natural Language Instructions for Fetching Daily Objects Using GAN-Based Multimodal Target–Source Classification, *RA-L*, Vol. 4, No. 4, pp. 3884–3891 (2019)

[Oord 18] Oord, A. v. d., Li, Y., and Vinyals, O.: Representation Learning with Contrastive Predictive Coding, *arXiv preprint arXiv:1807.03748* (2018)

[Qi 20] Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W. Y., et al.: Reverie: Remote embodied visual referring expression in real indoor environments, in *CVPR*, pp. 9982–9991 (2020)

[Radford 21] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., et al.: Learning Transferable Visual Models From Natural Language Supervision, in *ICML*, pp. 8748–8763 (2021)

[Ren 17] Ren, S., He, K., et al.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *IEEE Trans. PAMI*, Vol. 39, No. 6, pp. 1137–1149 (2017)

[Rezatofighi 19] Rezatofighi, H., Tsoi, N., Gwak, J., et al.: Generalized intersection over union: A metric and a loss for bounding box regression, in *CVPR*, pp. 658–666 (2019)

[Saito 18] Saito, K., Watanabe, K., Ushiku, Y., and Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation, in *CVPR*, pp. 3723–3732 (2018)

[Shridhar 20] Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., and Fox, D.: ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks, in *CVPR*, pp. 10740–10749 (2020)

[Uppal 22] Uppal, S., Bhagat, S., Hazarika, D., Majumder, N., Poria, S., Zimmermann, R., and Zadeh, A.: Multimodal research in vision and language: A review of current and emerging trends, *Information Fusion*, Vol. 77, pp. 149–171 (2022)

[Yu 18] Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., et al.: MAttNet: Modular Attention Network for Referring Expression Comprehension, in *CVPR*, pp. 1307–1315 (2018)

[小槻 22] 小槻 誠太郎, 石川慎太朗, 杉浦孔明：TDP-MAT に基づく実画像を対象とした物体操作指示理解, 第 40 回日本ロボット学会学術講演会, 4I1-02 (2022)