マルチモーダル言語理解タスクにおける Dual ProtoNCE に基づくドメイン適応と 大規模言語モデルを用いた指示文理解

○松田一起，小槻誠太郎，杉浦孔明 (慶應義塾大学)

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．例えば，“Change the light bulb of the middle light in the office” という命令文が与えられたときに，ロボットが中央の照明を特定することは困難である．こうした物体操作指示理解タスクを解決する上で，転移学習を当該タスクに導入することで精度の向上を図った手法も存在する．しかし，この転移学習においてシミュレーションデータと実環境データのドメイン間の差異が大きい場合，転移学習の性能が低下し当該タスクの精度が低下してしまうという問題があった．本研究では，物体操作指示理解タスクにおける転移学習手法に対して Paraphraser を導入する．本研究の新規性である Paraphraser は，大規模言語モデルを用いて自然言語の命令文に対するドメイン間の差異を埋める言い換えを行うモジュールであり，既存の転移学習手法における精度の向上を目的とする．候補物体が対象物体であるか否かに関する精度において実世界データセットで性能評価を行った結果，提案手法はベースライン手法を上回った．

1. はじめに

現代社会では高齢化が進行し，在宅介助者の不足は深刻な社会問題となっている．この社会問題に対する解決策として，生活支援ロボットが有望視されているが，生活支援ロボットの人間との自然な対話能力はまだ不十分である。こうしたロボットの訓練には，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が，そのためには膨大なコストがかかり困難である．一方，シミュレーション上であれば，実世界でのデータ収集よりも低コストでデータを収集することができる．そのため，シミュレーションデータを用いて訓練を行うことができると便利である．

本研究では，画像と自然言語の命令文が与えられた際，ロボットが画像中の対象物体を特定することを目的とする．例えば，“Head up to the bathroom on level 3 with the raised sink and turn off the light above it” という命令文が与えられた場合には，画像内からシンクの上にある照明を対象物体として予測することが望ましい．我々が取り組む Multimodal Language Understanding for Fetching Instruction (MLU-FI) [1,2] というマルチモーダル言語理解タスクにおいて，Otsuki らは [2]，シミュレーションデータと実環境データの間で転移学習を行う手法である PCTL を提案し，良好な結果を報告している．一方，当該手法は転移学習の性能が不十分であり，MLU-FI の精度には改善の余地があった．

本研究では，マルチモーダル言語理解タスクにおける転移学習手法である PCTL [2] に対して Paraphraser を導入する．Paraphraser は，大規模言語モデルを用いて自然言語の命令文に対するドメイン間の差異を埋める言い換えを行うモジュールであり，既存の転移学習手法における精度の向上を目的とする．例えば，“Make your way down the hall to the second floor office kitchen and turn off the lights” という命令文が与えられた場合，Paraphraser は当該文を “Turn off the lights in the second floor office kitchen.” に言い換える．

提案手法が PCTL と異なる点は，大規模言語モデルで構成される Paraphraser を導入し，ドメイン間の差異を埋めるように入力命令文の言い換えを行った点である．PCTL では，転移学習によりドメインの差異を軽減したが，シミュレーションデータと実環境データにおいてドメイン間の差異が大きい場合，転移学習の性能が低下する．ここで，Paraphraser を導入することにより，入力命令文にドメイン間の差異を埋める言い換えを施すことで性能向上が期待できる．

図 1: MLU-FI におけるタスクの代表例．命令文：“Get me the picture furthest on the left.”

本研究の新規性は以下である．

- 大規模言語モデルを用いて，自然言語の命令文に対してドメイン間の差異を埋める言い換えを行う Paraphraser を導入する．

2. 関連研究

マルチモーダル分野では近年多くの研究が行われている [3]．Uppal ら [3] は，マルチモーダル言語理解タスクについてサーベイを行い，視覚と言語のモダリティに関連する研究の最新動向について詳しく紹介している．

参照表現理解タスクは，マルチモーダル言語理解タスクの一つであり，広く研究が行われている．[4–6]．参照表現理解タスクは，参照表現を用いて表された対象物体を画像内から特定するというタスクである．一方，我々が取り組むタスクである MLU-FI [1,2] は参照表現理解タスクと比べ柔軟な問題設定がなされている．MLU-FI に対して，Magassouba らは MCTM [7] を，石川らは UNITER [4] をベースモデルとした Target-Dependent UNITER [1] を提案している．小槻らは，シミュレーションデータと実環境データの間での転移学習を本タスクに適用する手法である PCTL [2] を提案している．

3. 問題設定

本研究では MLU-FI に取り組む．本タスクは，まず画像とそれに対応する命令文，それらに加え画像から

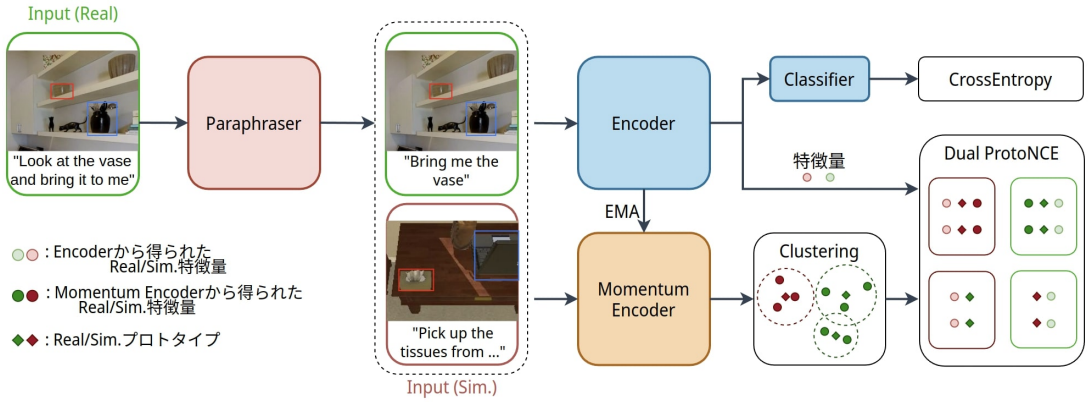

図 2: 提案手法のフレームワーク図

物体検出器によって得られた一つの候補領域とそれを含むコンテキスト領域群が与えられ，候補領域が対象領域であるか否かの 2 値分類を行うものである．

MLU-FI において期待される出力は，候補物体が対象物体と一致している確率の予測値 $P(\hat{y}=1)$ である．候補物体と対象物体が一致しているときは $P(\hat{y}=1)=1$，異なるときは $P(\hat{y}=1)=0$ と出力することが望ましい．ここで y はラベル，$\hat{y}$ はその予測であり，$\hat{y}=1$ は候補物体が対象物体と一致していることを示す．

MLU-FI におけるタスクの代表例を図 1 に示す．このサンプルでは画像とともに “Get me the picture furthest on the left.” という命令文が与えられている．対象領域，候補領域，コンテキスト領域群をそれぞれ緑色，赤色，黄色の矩形領域で示す．この例の場合，命令文に対応する左側の絵が対象物体であるのに対し，上側の絵が候補物体であるため，期待される出力は $P(\hat{y}=1)=0$ となる．一方，対象物体と候補物体が一致している場合は $P(\hat{y}=1)=1$ と出力することが期待される．

- **対象物体:** 命令文が指し示す物体
- **候補物体:** 対象物体と同じ物体か否かを判断する物体
- **コンテキスト物体:** 画像から物体検出器によって検出されたすべての物体

対象領域，候補領域，コンテキスト領域群はそれぞれ，対象物体を囲む矩形領域，候補物体を囲む矩形領域，コンテキスト物体を囲む矩形領域群のことを示す．また，候補領域はコンテキスト領域群から選ばれる．

MLU-FI は，画像中の全ての物体から一つの物体を選択する多クラス分類ではなく，ある候補物体が対象物体と一致しているか否かを出力する 2 値分類である．2 値分類として定式化することによって，複数の物体が対象物体となる場合や対象物体が画像中に存在しない場合を考慮できる．また，MLU-FI タスクでは評価尺度として精度を用いる．データの大規模化による性能向上を考えた際に，実機を用いたデータ収集はコストが高い．そのため，本研究では実機で得られたデータに加えシミュレーションで得られたデータも使用する．

4. 提案手法

本手法は，MLU-FI における既存手法の PCTL を拡張した手法である．PCTL は，対照学習を利用した転移学習の手法であり，本研究では PCTL に大規模言語モデルによる入力命令文の言い換えを行うモジュールを適用する．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である MLU-FI に適用するが，大規模言語モデルによる入力命令文に対する言い換えは広くマルチモーダル言語理解タスクに応用可能である．

図 2 に提案手法のフレームワーク図を示す．提案手法は大きく分けて Paraphraser，Encoder，Momentum Encoder，Clustering Module，Dual Contrastive Transfer Learning Module という 5 つのモジュールから構成される．

モデルの入力を次のように定義する．

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

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

上式において，$\boldsymbol{x}_{\text{inst}}$ は命令文，$\boldsymbol{x}_{\text{cand}} \in \mathbb{R}^{2048}$, $\boldsymbol{X}_{\text{cont}}^{(i)} \in \mathbb{R}^{2048}$ は，候補領域とコンテキスト領域それぞれの特徴量を表し，$\boldsymbol{X}_{\text{cont}}$ と N_{det} はそれぞれ，コンテキスト領域群と Faster R-CNN [8] によって入力画像から検出された領域数を表す．なお，$\boldsymbol{x}_{\text{inst}}$ から抽出する言語特徴量及び $\boldsymbol{x}_{\text{cand}}$，$\boldsymbol{X}_{\text{cont}}^{(i)}$ には positional encoding が付与される．

Paraphraser は，$\boldsymbol{x}_{\text{inst}}$ をドメイン間の差異を埋める言い換えを行った命令文 $\boldsymbol{x}_{\text{pinst}}$ へと変換する．本モジュールでは，OpenAI の提供する大規模言語モデルである gpt-3.5-turbo [9] を用いる．この大規模言語モデルを用いた言い換えには以下のプロンプトを用いた．“[instructions]. Remove the clause giving the instruction regarding movement beginning with ’Go to’ and extract the clause giving the instruction regarding action after ’and’ from each of the above sentences. Beginning and end of the each output should be enclosed in #. Output the information only.” 例として，$\boldsymbol{x}_{\text{inst}}$ が “Go to the familyroom and tidy up the end table by the sofa” であるとき，gpt-3.5-turbo から “#Tidy up the end table by the sofa in the familyroom.#” という文を得る．その後文頭と文末の “#” を除去することで，最終的に出力 $\boldsymbol{x}_{\text{pinst}}$ として “Tidy up the end table by the sofa in the familyroom.” を得る．

Encoder f は，$\boldsymbol{x}_{\text{pinst}}$ を命令文にもつサンプルを入力とし，768 次元の特徴量を出力するモジュールであり，f の構造は Target-Dependent UNITER [1] と同一である．本章以降，転移元ドメインのサンプル及びラ

ベルをそれぞれ x_s，y_s と表し，同様に転移先ドメインのサンプル及びラベルを x_t，y_t と表す．本モジュールの入出力を $\boldsymbol{u} = f_\theta(x_\mathrm{s}) \in \mathbb{R}^{768}$, $\boldsymbol{v} = f_\theta(x_\mathrm{t}) \in \mathbb{R}^{768}$ と表す．$\boldsymbol{u}$ 及び $\boldsymbol{v}$ はそれぞれ 768 次元の，転移元ドメイン及び転移先ドメインの特徴量である．これらはクラスタリングや損失の計算に利用される．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 の出力を受け取り予測確率を出力する．

Momentum Encoder f' は f と同じ構造を持ち，f のパラメータ θ の指数移動平均 θ' をパラメータとして持つモジュールである．本モジュールの入出力を $\boldsymbol{u}' = f_{\theta'}(x_\mathrm{s}) \in \mathbb{R}^{768}$, $\boldsymbol{v}' = f_{\theta'}(x_\mathrm{t}) \in \mathbb{R}^{768}$ と表す．またパラメータ θ' の更新において，この指数移動平均で用いられる平滑化係数を γ と定義する．

Clustering Module は，$\boldsymbol{u}'$, $\boldsymbol{v}'$ に対して k 近傍法によるクラスタリングを，クラスタ数を変えながら 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 $k(m)$ と表す．また，m 回目の，$\boldsymbol{u}'$, $\boldsymbol{v}'$ 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，$c_i(m)$, $d_i(m)$ とし，これを i 番目のクラスタのプロトタイプとする．

Contrastive Transfer Learning Module は転移学習のために一般化された対比損失である Dual ProtoNCE [2] を計算し，転移元ドメインと転移先ドメインの差異を埋めるモジュールである．$\mathcal{L}_\mathrm{DualProtoNCE}$ は，Intra-Domain 損失 $\mathcal{L}_\mathrm{Intra}$ と Inter-Domain 損失 $\mathcal{L}_\mathrm{Inter}$ の和で表される．まず，$\mathcal{L}_\mathrm{Intra}$ は，転移元ドメインデータと転移先ドメインデータに対して独立に ProtoNCE [10] を適用することで計算する．

$$\mathcal{L}_\mathrm{Intra} = \mathcal{L}_\mathrm{Target} + \mathcal{L}_\mathrm{Source}$$

$$\mathcal{L}_\mathrm{Target} = \sum_{i=1}^{n} - \left(\log \frac{\exp({}^\dagger\boldsymbol{v}_i \cdot {}^\dagger\boldsymbol{v}'_i/\tau)}{\sum_{j\in J}\exp({}^\dagger\boldsymbol{v}_i \cdot {}^\dagger\boldsymbol{v}'_j/\tau)} + \frac{1}{M}\sum_{m=1}^{M} \log \frac{\exp({}^\dagger\boldsymbol{v}_i \cdot {}^\dagger\boldsymbol{c}_s^{(m)}/\phi_s^{(m)})}{\sum_{j\in J'}\exp({}^\dagger\boldsymbol{v}_i \cdot {}^\dagger\boldsymbol{c}_j^{(m)}/\phi_j^{(m)})} \right)$$

なお，$\mathcal{L}_\mathrm{Source}$ については，$\mathcal{L}_\mathrm{Target}$ と同様の計算を転移元ドメインについて行うものとする．

ここで，$J = \{i, n+1, n+2, \ldots, n+r\}$ であり，$\boldsymbol{z}_i$ と $\boldsymbol{z}'_i$ はそれぞれアンカーと正例の特徴量，$\{\boldsymbol{z}'_j \mid j = n+1, n+2, \ldots, n+r\}$ は r 個の負例の特徴量である．${}^\dagger\boldsymbol{a}$ は，$\boldsymbol{a}/\|\boldsymbol{a}\|_2$ を表し，ϕ, φ はそれぞれ，クラスタ中のプロトタイプ $\boldsymbol{c}$, $\boldsymbol{d}$ に対して属するインスタンスが集中している度合いを表現するパラメータである．また，τ は温度パラメータである．PCTL では，CLIP [11] 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．次に，$\mathcal{L}_\mathrm{Inter}$ は 2 ドメイン間のギャップを埋めるために，次のように定式化される．

$$\mathcal{L}_\mathrm{Inter} = \mathcal{L}_\mathrm{S2T} + \mathcal{L}_\mathrm{T2S}$$

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

ここで，$\mathcal{L}_\mathrm{S2T}$ 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であり，$\mathcal{L}_\mathrm{T2S}$ について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を同様の計算で行うものとする．

損失関数の定義を次に示す．

$$\mathcal{L} = \lambda\mathcal{L}_\mathrm{DualProtoNCE} + \mathcal{L}_\mathrm{t} + \mathcal{L}_\mathrm{s}$$

$$\mathcal{L}_\mathrm{t} = \sum_{i=1}^{n}\left(\mathcal{L}_\mathrm{CE}(g(f_\theta(\boldsymbol{x}_\mathrm{t}^{(i)})), y_\mathrm{t}^{(i)}) + \frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_\mathrm{CE}(g(\boldsymbol{c}_s^{(m)}), y_\mathrm{t}^{(i)}) \right)$$

なお，$\mathcal{L}_\mathrm{s}$ については，$\mathcal{L}_\mathrm{t}$ と同様の計算を転移元ドメインについて行うものとする．ここで，$\mathcal{L}_\mathrm{CE}(\cdot,\cdot)$，$\lambda$ はそれぞれ，交差エントロピー誤差とハイパーパラメータである．

5. 実験設定

本研究では，REVERIE-fetch [2] データセットを用いる．このデータセットは REVERIE [12] データセットから MLU-FI に必要なデータを抽出したものである．訓練集合はモデルの学習に，検証集合はハイパーパラメータの調整に，テスト集合はモデルの性能の評価に使用した．また，本研究では，転移学習のために与えられるデータとして ALFREAD-fetch-b [2] データセットを用いた．

本手法では，層数 12，隠れ層の次元数 768，Attention Head 数 12 の transformer を用いた．最適化手法は momentum = 0.9 とした SGD with momentum を用い，Classifier の学習率と Multi-Layer Transformer の学習率をそれぞれ，8×10^{-4}，8×10^{-5} とした．バッチサイズは 64，エポック数は 40 とした．また，使用したモデルの総学習可能パラメータ数は約 2.2 億である．学習には VRAM24GB 搭載の GeForce RTX 3090 及び RAM64GB 搭載の Intel Core i9-12900K を使用した．訓練には約 2 時間を要し，1 サンプルあたりの推論には約 100ms を要した．本研究では早期終了を用いた．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算出し，その値が最も小さいエポックでのテスト集合における精度を最終的な精度とした．

6. 実験結果

REVERIE-fetch データセットにおける各手法の精度を表 1 に示す．なお，左列が手法，右列が精度を示しており，精度は 5 回の試行によって得られた平均値と標準偏差を用いた．ベースラインとして以下の 4 つの手法を用いた．

- (i) Target domain only: 転移先ドメインのデータセットのみで学習を行う．

表 1: REVERIE-fetch データセットにおける定量的結果．

手法	精度 [%]
(i) Target domain only	73.0 ± 1.87
(ii) Fine-tuning	73.4 ± 11.8
(iii) MCDDA+ [13]	74.9 ± 3.94
(iv) PCTL [2]	78.1 ± 2.49
Ours	$\mathbf{78.6 \pm 1.87}$

(a) “Bring me the solid white towel.”

(b) “Clean out the toaster oven in the hallway by the indoor grill.”

図 3: REVERIE-fetch データセットにおける定性的結果.

- (ii) Fine-tuning: 転移元ドメインのデータセットによる事前学習を行った後，転移先ドメインのデータセットによるファインチューニングを行う．
- (iii) MCDDA+: MCDDA [13] という教師なし転移学習手法を，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．
- (iv) PCTL: PCTL [2] で提案された，Dual ProtoNCE という対比損失を用いた転移学習を行う．

本研究で用いる REVERIE-fetch データセットには，クラスごとのサンプル数における不均衡がないため評価尺度として精度を用いた．表 1 より，提案手法の精度が 78.6%，ベースライン (i)，(ii)，(iii)，(iv) の精度がそれぞれ，73.0%，73.4%，74.9%，78.1%であり，精度において，提案手法がベースラインを全て上回った．

提案手法が成功した例を図 3 に示す．図 3(a) は True Positive の例である．Paraphraser を通した後の命令文は “Bring me the solid white towel.” であり，対象物体と候補物体はともに白いタオルである．この例において提案手法のモデルは $P(\hat{y}=1)=0.999$ と出力しており，候補物体が対象物体と一致していると正しく予測した．図 3(b) は True Negative の例である．Paraphraser を通した後の命令文は “Clean out the toaster oven in the hallway by the indoor grill.” であり，対象物体は机の上にあるオーブン，候補物体は白い小物である．この例において提案手法のモデルは $P(\hat{y}=1)=0.095$ と出力しており，候補物体が対象物体と一致していないと正しく予測した．

以下の 3 つの条件で Ablation study を行った．

- (i) w/o Paraphraser: Paraphraser を適用しない．

表 2: Ablation study の定量的結果.

条件	精度 [%]
(i)	75.4 ± 2.43
(ii)	73.5 ± 2.60
(iii)	$\mathbf{78.6 \pm 1.87}$

- (ii) w/ Paraphraser in training: Paraphraser を訓練集合の命令文のみに適用する．
- (iii) w/ Paraphraser: Paraphraser を訓練集合，検証集合，テスト集合全ての命令文に適用する．

Ablation study の定量的結果を図 2 に示す．図 2 に示すように，提案手法である条件 (iii) の精度は 78.6%であった．これは，条件 (i) の精度 75.4%を 3.2 ポイント上回り，条件 (ii) の精度 73.5%を 5.1 ポイント上回った．この結果，提案手法が Ablation study の条件の中で最も高い精度を示した．

7. おわりに

本研究では，マルチモーダル言語理解タスクである MLU-FI に取り組んだ．本研究の貢献を次に示す．

- 大規模言語モデルを用いて，自然言語の命令文に対してドメイン間の差異を埋める言い換えを行う Paraphraser 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，提案手法がベースライン手法を上回った．

将来研究として，シミュレータによってさらに多くのデータを収集することや，提案手法で学習したモデルを実機へ応用することが挙げられる．

謝辞
本研究の一部は，JSPS 科研費 23H03478，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

[1] S. Ishikawa and K. Sugiura, “Target-dependent UNITER: A Transformer-Based Multimodal Language Comprehension Model for Domestic Service Robots,” RA-L, vol.6, no.4, pp.8401–8408, 2021.
[2] S. Otsuki, S. Ishikawa, and K. Sugiura, “Prototypical Contrastive Transfer Learning for Multimodal Language Understanding,” IROS, 2023. to appear.
[3] S. Uppal, et al., “Multimodal Research in Vision and Language: A Review of Current and Emerging Trends,” Information Fusion, vol.77, pp.149–171, 2022.
[4] C. Chen, L. Li, L. Yu, A. El, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, “UNITER: UNiversal Image-TExt Representation Learning,” ECCV, pp.104–120, 2020.
[5] A. Kamath, M. Singh, Y. LeCun, et al., “MDETR - Modulated Detection for End-to-End Multi-Modal Understanding,” ICCV, pp.1780–1790, 2021.
[6] P. Wang, et al., “OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework,” ICML, pp.23318–23340, 2022.
[7] A. Magassouba, K. Sugiura, 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.
[8] S. Ren, 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.
[9] https://platform.openai.com/docs/models/gpt-3-5
[10] J. Li, P. Zhou, C. Xiong, and S. Hoi, “Prototypical Contrastive Learning of Unsupervised Representations,” ICLR, 2021.
[11] A. Radford, J. Kim, C. Hallacy, Ramesh, et al., “Learning Transferable Visual Models from Natural Language Supervision,” ICML, pp.8748–8763, 2021.
[12] Y. Qi, et al., “REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments,” CVPR, pp.9982–9991, 2020.
[13] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, “Maximum Classifier Discrepancy for Unsupervised Domain Adaptation,” CVPR, pp.3723–3732, 2018.