# TDP-MAT に基づく実画像を対象とした物体操作指示理解

○小槻誠太郎，石川慎太朗，杉浦孔明（慶應義塾大学）

## 1. はじめに

高齢化が進展する現代社会において，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．一方，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．そこで本研究では，生活支援ロボットにおける物体操作指示理解タスクを扱う．

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ことを目的とする．例えば “Look in the vase that is next to the potted plant” という命令文及び観葉植物と 3 つの花瓶が写った画像が与えられた場合，観葉植物に一番近い花瓶を対象物体として予測する事が望ましい．一方，人間の発する命令はしばしば曖昧であり，その内容を正確に理解することは容易ではない．例えば上述の例では参照表現を正確に解釈し，3 つの花瓶から観葉植物に一番近いものを選ぶ必要がある．こうした参照表現を含む命令文では対象物体の特定が難しく，実際に誤る例が報告されている [1].

既存手法である Target-Dependent UNITER [2] は物体操作指示理解タスクにおいて UNITER [3] を使用して高い精度を達成したモデルである．一方で Target-Dependent UNITER は精度がばらつきやすく，複雑なシーンを扱う際に十分な性能が安定して得られないという問題がある．

本研究では敵対的学習を導入し，Perceiver [4] によって特徴量間の関係のモデル化を行う物体操作指示理解モデルである Target-Dependent Perceiver with Moment Adversarial Training(TDP-MAT) を提案する．敵対的学習は頑健なモデルを獲得するための一般的な手法であり，これによって提案手法はモデルの頑健性を向上させると期待される．また，Perceiver によって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ことで，計算量の入力の長さへの依存を抑えつつ，モデルの性能が向上することが期待される．提案手法と既存手法との違いは，視覚特徴量に対して，MAT [5] に基づく敵対的摂動を加える点，また Perceiver によって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点にある．

本研究の貢献を以下に示す.

- Perceiver によって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 を提案する.
- Perceiver の出力に対して MAT に基づく敵対的摂動を加える構造を導入する.
- REVERIE データセット [6] から命令文とそれに対応する実画像を収集し，REVERIE-fetch データセットを作成する.

## 2. 関連研究

マルチモーダル言語処理分野のサーベイ論文として，[7] が挙げられる．マルチモーダル言語処理分野のうち言語と画像を扱う分野に分類されるものとして Vision-and-Language Navigation(VLN) が挙げられる.

Multimodal Target-source Classifier Model [8] は自然言語による命令文と視覚的な情報から対象物体を特定するモデルである．HLSM-MAT [5] は VLN において，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，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た手法である.

VLN 分野における有名なデータセットとして，Room-to-Room (R2R) [9]，MatterPort3D [10]，REVERIE [6] が挙げられる．R2R は実際の室内環境における視覚情報に基づいた自然言語による移動タスクのためのデータセットである．MatterPort3D は室内環境におけるシーン理解を行うための大規模な深度付き画像データセットである．REVERIE データセットは Remote Embodied Visual Referring Expression in Real Indoor Environments(REVERIE) タスクを行うためのデータセットである．この REVERIE タスクは，環境中に存在する対象物体に関連した自然言語による命令文が与えられ，始めに対象物体が存在する場所をゴールとする複数ステップの行動選択サブタスクを行い，その後対象物体を特定するサブタスクを行う.

## 3. 問題設定

本論文では以下のように用語を定義する.

- **対象物体:** 命令文が対象としている物体
- **候補物体:** 対象物体であるか否かの候補となる物体
- **コンテキスト物体:** 物体検出器で検出される物体

なお，対象領域，候補領域，コンテキスト領域はそれぞれ対象物体，候補物体，コンテキスト物体を囲むバウンディングボックスを指すものとする.

本論文で扱うタスクは，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 値分類を行うというものである．以降，本タスクを REVERIE-fetch タスクと呼ぶ．REVERIE-fetch タスクにおいて期待される出力は，候補領域が対象領域である確率の予測値 $p(\hat{y})$ である．候補領域と対象領域が一致しているときは $p(\hat{y}) = 1$，異なるときは $p(\hat{y}) = 0$ と出力することが望ましい．図 1 に REVERIE-fetch タスクの代表例を示す．図 1 においては，画像中央奥に位置する赤いバウンディングボックスで示す枝編み細工の花瓶が対象物体となる．例えば候補領域として緑色のバウンディングボックスが与えられた時は $p(\hat{y}) = 0$，候補領域として対象領域に対応する赤いバウンディングボックスが与えられたときは $p(\hat{y}) = 1$ と出力するこ

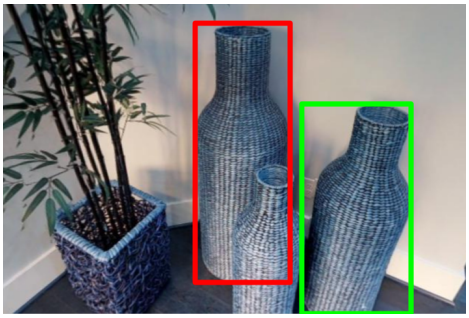

図 1 REVERIE-fetch タスクの例.
命令文: “Look in the left wicker vase that is next to the potted plant on the second floor at the foot of the stairs”

とが期待される．本タスクの評価尺度には，候補領域が対象領域であるか否かを分類する精度 Acc を使用する．なお本研究では，物体検出誤りが十分に少ない物体検出器を利用できることを前提とし，命令文に対応する画像から候補領域，及びコンテキスト領域を抽出する過程はこの物体検出器によって行われることを仮定する．

## 4. 提案手法

提案ネットワークの構造を図 2 に示す．図 2 において，Instruction は命令文，Context Regions はコンテキスト領域群，Candidate Region は候補領域を表す．

### 4.1 入力

ネットワークの入力 $\boldsymbol{x}$ を以下のように定義する．

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

$$\boldsymbol{X}_{\text{inst}} = \{\boldsymbol{x}_{\text{inst}}, \boldsymbol{x}_{\text{pos}}\} \tag{2}$$

$$\boldsymbol{X}_{\text{cont}} = \left\{\boldsymbol{x}^{(i)}_{\text{cont}}, \boldsymbol{x}^{(i)}_{\text{contloc}} | i = 1, \ldots, N_{\text{FRCNN}}\right\} \tag{3}$$

$$\boldsymbol{X}_{\text{cand}} = \{\boldsymbol{x}_{\text{cand}}, \boldsymbol{x}_{\text{candloc}}\} \tag{4}$$

ここで，$\boldsymbol{X}_{\text{inst}}$ は命令文，$\boldsymbol{X}_{\text{cont}}$ はコンテキスト領域群の特徴量，$\boldsymbol{X}_{\text{cand}}$ は候補領域の特徴量を表す．$\boldsymbol{x}_{\text{inst}}$ は命令文のトークン列，$\boldsymbol{x}_{\text{pos}}$ は命令文の各トークンの先頭からの順序であり，$\boldsymbol{x}^{(i)}_{\text{cont}}$ は $i$ 番目の物体のコンテキスト領域の特徴量，$\boldsymbol{x}^{(i)}_{\text{contloc}}$ はそのコンテキスト領域の位置埋め込み，$\boldsymbol{x}_{\text{cand}}$ は候補領域の特徴量，$\boldsymbol{x}_{\text{candloc}}$ は候補領域の位置埋め込みを表す．また，$N_{\text{FRCNN}}$ は Faster R-CNN [11] によって検出した画像中の領域の数である．

$\boldsymbol{X}_{\text{inst}}$ に関しては，SentencePiece [12] によるトークン化で $\boldsymbol{x}_{\text{inst}}$ を獲得する．$\boldsymbol{x}_{\text{cont}}$ に関しては，全体画像を Faster R-CNN に入力し，Faster R-CNN における ResNet50 の fc6 層の出力によって $\boldsymbol{x}_{\text{cont}}$ を獲得する．一方で $\boldsymbol{x}_{\text{contloc}}$ は Faster R-CNN によって得たバウンディングボックスの位置情報をエンコードして得る．$\boldsymbol{X}_{\text{cand}}$ は $\boldsymbol{X}_{\text{cont}}$ の中から判定対象として選択する．なお，$\boldsymbol{X}_{\text{cont}}$ 及び $\boldsymbol{X}_{\text{cand}}$ における位置埋め込み $\boldsymbol{x}_{\text{contloc}}$ 及び $\boldsymbol{x}_{\text{candloc}}$ 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する．

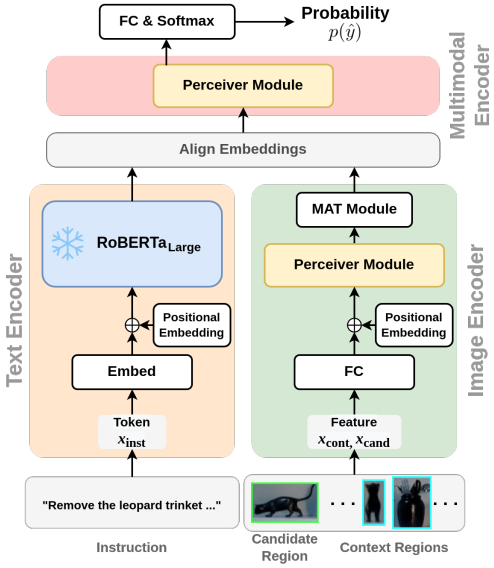


図 2 提案手法のネットワーク構成

### 4.2 Text Encoder

Text Encoder は $\boldsymbol{X}_{\text{inst}}$ を入力として特徴量 $E_{\text{T}}$ を出力する．始めに $\boldsymbol{x}_{\text{inst}}$ と $\boldsymbol{x}_{\text{pos}}$ をそれぞれ 1024 次元のベクトル表現に埋め込んでから足し合わせ，次にパラメータを固定した事前学習済み RoBERTa [13] に入力することで $E_{\text{T}}$ を得る．なお，RoBERTa は事前学習に特化した最終層を取り除いた 23 層の transformer layer で構成される．

### 4.3 Image Encoder

Image Encoder は 2 つの全結合層，Perceiver モジュール，そして MAT モジュールから構成される．Image Encoder は $\boldsymbol{X}_{\text{cand}}$ 及び $\boldsymbol{X}_{\text{cont}}$ を入力とし，それぞれを全結合層によって 1024 次元のベクトル表現に埋め込んで足し合わせる．次に足し合わせて得た特徴量を Perceiver モジュールに入力して各ベクトル表現の関係性をモデル化し，さらに MAT モジュールによって敵対的摂動を加え，特徴量 $E_{\text{cand}}$ 及び $E_{\text{cont}}$ を出力する．

#### 4.3.1 MAT モジュール

本モジュールは [5] で提案された Moment-based Adversarial Training に基づく敵対的摂動 $\boldsymbol{\delta}$ を加えるモジュールである．ここで敵対的摂動 $\boldsymbol{\delta}$ の更新方法を示す．始めに，交差エントロピー誤差の $\boldsymbol{\delta}$ に関する勾配 $\nabla_{\boldsymbol{\delta}}$ を計算する．続いて，Adam をもとにした 2 種類の移動平均 $\boldsymbol{m}_t$，$\boldsymbol{v}_t$ を導入する．

$$\boldsymbol{m}_t = \rho_1 \boldsymbol{m}_{t-1} + (1 - \rho_1) \nabla_{\boldsymbol{\delta}} E(\boldsymbol{\delta}_t), \tag{5}$$

$$\boldsymbol{v}_t = \rho_2 \boldsymbol{v}_{t-1} + (1 - \rho_2) (\nabla_{\boldsymbol{\delta}} E(\boldsymbol{\delta}_t))^2 \tag{6}$$

$t$ は現在の摂動更新ステップ，$\rho_1$ と $\rho_2$ は各移動平均の平滑化係数である．これらを用いて，摂動の更新幅 $\Delta\boldsymbol{\delta}_t$ を次のように導出する．

$$\hat{\boldsymbol{m}}_t = \frac{\boldsymbol{m}_t}{1 - \rho_1^t}, \ \hat{\boldsymbol{v}}_t = \frac{\boldsymbol{v}_t}{1 - \rho_2^t}, \tag{7}$$

$$\Delta\boldsymbol{\delta}_t = \eta \frac{\hat{\boldsymbol{m}}_t}{\sqrt{\hat{\boldsymbol{v}}_t + \varepsilon}} \tag{8}$$

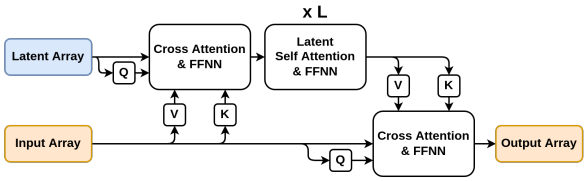


図 3　Perceiver モジュールの構造

ここで $\eta$ は MAT モジュールの学習率，$\varepsilon$ はゼロ除算を避けるための小さな値である．最後に導出した摂動の更新幅 $\Delta\boldsymbol{\delta}_t$ を利用し，[5] に従って $\boldsymbol{\delta}$ を更新する．

#### 4.3.2　Perceiver モジュール

本モジュールの構造を図 3 に示す．本モジュールは Perceiver [4] を用い，入力された特徴量間の関係性をモデル化する．Perceiver は，入力された特徴量を，潜在変数との cross attention を取ることによって固定サイズの潜在空間にマッピングし，潜在空間にマッピングした特徴量に対して self attention を繰り返す．これにより，入力の系列長に対して線形の計算量を達成しつつ入力された特徴量間の関係性をモデル化できる．出力の際には，潜在空間にマッピングされた特徴量を，入力特徴量との cross attention を取って入力と同じ空間にマッピングする．本モジュールではアテンション機構及び FFNN の初期化に DeepNet [14] で提案された DeepNorm を用い，同様に DeepNorm に従って残差接続にモジュールの深さに応じた係数を掛ける．

### 4.4　Multimodal Encoder

本モジュールは Perceiver モジュールで構成される．本モジュールは Text Encoder 及び Image Encoder の出力を連結した $[E_{\text{cand}}; E_{\text{inst}}; E_{\text{cont}}]$ を入力とし，候補領域が対象領域である確率の予測値 $p(\hat{y})$ を出力する．モデルの学習に際して，損失関数には交差エントロピー誤差を使用した．

## 5.　実験設定

### 5.1　データセット

我々は REVERIE データセット [6] から命令文及び各命令文に対応する実画像を収集し，1 枚の画像から対象物体を特定するタスクのための REVERIE-fetch データセットを作成した．なお，本実験では，REVERIE-fetch データセットの画像を Faster R-CNN [11] に入力し，各画像について候補領域及びコンテキスト領域群を抽出してサンプルを作成した．

REVERIE-fetch データセットは、画像と英語の命令文のペアからなるサンプルを 20990 サンプル含む．語彙サイズは 2576，全単語数は 399476，平均文長は 19.0 である．REVERIE-fetch データセットは訓練集合に 20440 サンプル，検証集合に 246 サンプル，そしてテスト集合に 304 サンプルのデータを含んでいる．

### 5.2　パラメータ設定

Image Encoder 及び Multimodal Encoder で使用される Perceiver [4] は，Latent Self Attention 層の数を 3，潜在空間中の特徴量の次元数を 512，Attention の Head 数を 16 とした．最適化には AdamW を使用し，学習率は $10^{-3}$，ステップ数は 32000，バッチサイズは 32，ドロップアウト率は 0.1 とした．なお，1 ステップは 1 つのバッチの処理を意味する．

TDP-MAT の総パラメータ数は 510M，学習可能パラメータ数は 168M である．学習にはメモリ 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．学習には約 3 時間を要し，推論には約 4.11ms/sample を要した．

表 1　REVERIE-fetch データセットにおける定量的結果

| Method | Condition | Acc [%] |
|---|---|---|
| Baseline [2] | | $83.2 \pm 2.25$ |
| Ours | full | $\mathbf{85.5 \pm 1.21}$ |
| | (i)-a | $50.0 \pm 0.00$ |
| | (i)-b | $83.8 \pm 1.53$ |
| | (ii) | $82.3 \pm 1.50$ |
| | (iii) | $83.2 \pm 2.05$ |

## 6.　実験結果

### 6.1　定量的結果

表 1 に REVERIE-fetch データセットにおける各手法の精度を示す．なお，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．本実験ではベースライン手法として，類似のタスクにおいて良好な結果が報告されている Target-Dependent UNITER [2] を用いた．表 1 より，提案手法およびベースライン手法の精度はそれぞれ 85.5%および 83.2%であり，提案手法がベースライン手法を 2.3 ポイント上回っている．

### 6.2　Ablation Study

Ablation Study には，以下の 4 条件を定めた．

(i)-a W/o MAT: Image Encoder において MAT モジュールによって敵対的摂動を加える場合と加えない場合で，性能への影響を調べた．

(i)-b W/o MAT+Smaller Learning Rate: 学習率を $10^{-4}$ に設定した上で，Image Encoder において MAT モジュールによって敵対的摂動を加える場合と加えない場合で，性能への影響を調べた．

(ii) W/o Perceiver Module in Image Encoder: Image Encoder において，Perceiver モジュールによって領域間の関係性をモデル化する場合としない場合を比較し，性能への影響を調べた．

(iii) W/o RoBERTa: Text Encoder においてパラメータを固定した事前学習済み RoBERTa を使用する場合としない場合を比較し，性能への影響を調べた．

各条件での定量的結果を表 1 に示す．まず MAT モジュールを取り除いた場合，同様の学習率では学習が失敗した．一方，学習率を $10^{-4}$ に設定した上で MAT モジュールを取り除いた場合は 1.7 ポイント精度が低下した．また，Text Encoder からパラメータを固定した事前学習済み RoBERTa を取り除いた場合は 2.3 ポイント，Image Encoder 中の Perceiver モジュールを取り除いた場合は 3.2 ポイント精度が低下した．即ち Image Encoder において Perceiver モジュールによって領域間の関係性をモデル化することが性能に最も寄与していると言える．また，MAT の導入によって高い学習率の条件下で学習の発散を防げることが示された．


図 4 TP（左），FN・OOV（右）の例．緑色の枠で囲まれている領域が対象領域であり，赤色の枠で囲まれている領域が候補領域である．
左: “Go to the open living room with the white fireplace and organize the stack of books on the table with two lamps”．右: “Go to the hallway on level 1 and remove the tallest vase please”

表 2　失敗例の分類

| Error | | #Errors |
|---|---|---|
| OOV | 命令文の分割失敗 | 16 |
| CE | 視覚情報や言語情報の処理の失敗 | 11 |
| CI | 不明瞭な命令文 | 7 |
| LVR | 参照表現の視覚情報の欠如 | 5 |
| MO | 候補領域が複数の物体を含む | 3 |

### 6.3　定性的結果

定性的結果を図 4 に示す．なお，TP は True Positive，FP は False Positive，FN は False Negative，TN は True Negative を表す．図 4 左は TP の例である．この例において対象物体は図中央に位置する山積みの本である．ベースライン手法である Target-Dependent UNITER は $p(\hat{y}) = 7.48 \times 10^{-5}$ と出力しており，候補領域が対象領域ではないと予測してしまっている．一方 TDP-MAT は $p(\hat{y}) = 9.99 \times 10^{-1}$ と出力しており，正確に候補領域が対象領域であると予測している．図 4 右は FN の例である．この例において対象物体は図中左側に位置する大きな枝編み細工の花瓶であり，画像中に 3 つ写り込んでいる花瓶から，命令文中の “tallest” という表現をもとに最も大きい花瓶が対象物体であると予測する必要がある．しかし TDP-MAT は $p(\hat{y}) = 4.39 \times 10^{-8}$ と出力しており，候補領域が対象領域ではないと予測してしまっている．予測に失敗した理由としては “vase” がトークン化によって (“v”，“ase”) に分割され，予測を妨げたことが挙げられる．

### 6.4　エラー分析

提案手法のテスト集合における予測結果の分類は，TP が 115 例，TN が 147 例，FP が 37 例，FN が 5 例であった．失敗例を分類した結果を表 2 に示す．失敗例は OOV, CE, LVR, CI, MO の 5 種類に大別された．OOV は命令文が語彙外の単語を含み命令文の分割に失敗したケースである．例えば，図 4 右に示した例は OOV である．CE はモデルが視覚情報や言語情報の処理に失敗したケースである．LVR は命令文中に含まれる参照表現の視覚情報が画像中に存在しなかったケースである．CI は不明瞭な命令文が与えられたケースである．例えばテーブルが複数写っている画像に対してテーブルを特定できない命令文が与えられる例が存在した．MO は候補物体とは関係ない物体の画素が候補領域に多く含まれるケースである．

表 2 より，提案手法の ボトルネックは命令文が語彙外の単語を含むケースであると言える．これについてはさらに語彙数を増やす，あるいはデータセットに含まれる単語の分布によってトークナイザを再度学習させることで影響を低減できると考えられる．

## 7.　結論

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モデルの開発に取り組んだ．

本研究の貢献を以下に示す．

- Perceiver [4] によって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 を提案した．
- Perceiver の出力に対して MAT に基づく敵対的摂動を加える構造を導入した．
- REVERIE データセット [6] から命令文と各命令文に対応する実画像を収集し，REVERIE-fetch データセットを作成した．
- REVERIE-fetch データセットにおいて提案手法がベースライン手法を上回る精度を記録した．


### 謝辞

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


### 参　考　文　献


[1] A. Magassouba, K. Sugiura, and H. Kawai, “A Multimodal Target-Source Classifier With Attention Branches to Understand Ambiguous Instructions for Fetching Daily Objects,” RA-L, vol.5, no.2, pp.532–539, 2020.
[2] S. Ishikawa and K. Sugiura, “Target-dependent UNITER: A Transformer-Based Multimodal Language Comprehension Model for Domestic Service Robots,” IROS, 2021.
[3] Y.-C. Chen, et al., “Uniter: Universal image-text representation learning,” ECCV, pp.104–120, 2020.
[4] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, “Perceiver: General perception with iterative attention,” ICML, pp.4651–4664, 2021.
[5] S. Ishikawa and K. Sugiura, “Moment-based Adversarial Training for Embodied Language Comprehension,” arXiv preprint arXiv:2204.00889, 2022.
[6] Y. Qi, Q. Wu, P. Anderson, et al., “Reverie: Remote embodied visual referring expression in real indoor environments,” CVPR, pp.9982–9991, 2020.
[7] A. Mogadala, et al., “Trends in integration of vision and language research: A survey of tasks, datasets, and methods,” JAIR, vol.71, pp.1183–1317, 2021.
[8] 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.
[9] P. Anderson, Q. Wu, et al., “Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments,” CVPR, pp.3674–3683, 2018.
[10] A.X. Chang, et al., “Matterport3D: Learning from RGB-D Data in Indoor Environments,” 3DV, pp.667–676, 2017.
[11] S. Ren, K. He, et al., “Faster R-CNN: towards real-time object detection with region proposal networks,” IEEE Trans. PAMI, vol.39, no.6, pp.1137–1149, 2016.
[12] T. Kudo and J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing,” EMNLP, pp.66–71, 2018.
[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., “Roberta: A robustly optimized bert pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.
[14] H. Wang, S. Ma, L. Dong, S. Huang, D. Zhang, and F. Wei, “Deepnet: Scaling transformers to 1,000 layers,” arXiv preprint arXiv:2203.00555, 2022.