CH-005

双方向言語モデルからの知識蒸留を用いた日本語情景文字認識
Japanese Scene Text Recognition Using Knowledge Distillation from Bidirectional Language Model

折橋 翔太† 山﨑 善啓† 内田 美尋† 高島 瑛彦† 東羅 翔太郎† 増村 亮†
Shota Orihashi Yoshihiro Yamazaki Mihiro Uchida
Akihiko Takashima Shotaro Tora Ryo Masumura

1. はじめに

風景を撮影することで得られる情景画像には、交通標識や広告看板など、画像の理解に必要な文字情報が多く含まれている。情景文字認識は、そのような情景画像から文字の領域を切り出した画像を入力として、そこに写る文字列を認識するタスクである。情景文字認識は、画像分類[1–3]や画像検索[1,4]、画像に対する質問応答生成[5]など多様なアプリケーションに応用可能であり、学術的にも産業的にも活発に研究開発が進められている。

深層学習技術の発展に伴い、エンドツーエンド型のニューラルネットワークを用いる情景文字認識の手法が多数提案されている[6,7]。特に近年は、自然言語処理のための深層学習技術の進展により、Transformer[8]を用いたエンコーダ・デコーダ構造を持つ自己回帰型のモデルが高い文字認識性能を達成している[9–12]。これらの手法では、Convolutional Neural Network（CNN）と Transformer エンコーダからなるエンコーダが入力画像を画像特徴にエンコードし、Transformer デコーダからなるデコーダが画像特徴を自己回帰により文字列にデコードする。

情景文字認識は、情景画像から文字領域を検出する情景文字検出の後段に実行されることが多い[6,7]。ここで、英語など欧米圏の言語では、図 1 に示されるように単語がスペースにより空間的に区切られていることから、単語単位で文字領域が検出される。一方、日本語などアジア圏の言語では、単語がスペースにより区切られていないため、複数単語や複数文が含まれる文字領域が検出される。このような言語では、情景文字認識を単語単位でなく、複数単語や複数文の単位で行う必要がある。このため、正確な文字認識には文字の順序や出現頻度に関する高度な言語的知識が必要となる。

情景文字認識において、言語的知識を強化する手法はいくつか提案されている。特に、大規模なテキストデータで学習された外部の強力な言語モデルを用いる手法の有効性が示されている。例えば、Kang らの手法[14]では、文字認識モデルが出力する文字の生成確率と、外部の言語モデルが出力する文字の生成確率をフュージョンすることで最終的な文字の生成確率を得る。また、Wick らの手法[15]では文字認識モデルから出力された複数の文字列候補を外部の言語モデルによりリスコアリングすることで、言語的に尤もらしい文字列を得る。これらの手法は、言語的知識を踏まえた文字認識を可能にするが、自己回帰型の文字認識モデルが出力する文字生成確率の修正や文字列候補のリスコアリングにとどまっている。ここで、自己回帰型の文字認識モデルは主に過去の文脈に基づき次の文字を予測するため、未来の文脈を十分に考慮することができない。そのため、上述の手法で言語モデルの知識を用いたとしても、未来の文脈を十分に考慮した文字認識は難しく、言語的知識の強化や精度の向上に限界がある。

図 1　情景文字検出の例．赤枠で囲まれた各領域が情景文字認識への入力となる．なお，本図は ICDAR2019 のコンペティションである MLT 向けのデータセット[13]の画像とアノテーションを示している．

本稿では、情景文字認識のための外部の双方向言語モデルからの知識蒸留手法を提案する。提案手法は、過去だけでなく未来の文脈も考慮する言語モデルに蓄積された知識を文字認識モデルに蒸留することで、文字認識モデルに過去だけでなく未来の文脈を積極的に考慮させる。このため提案手法は、大規模なテキストデータを用いて Masked Language Modeling（MLM）[16]により学習された過去と未来両方の文脈を考慮し文字の生成確率を導出する外部の双方向言語モデルから、文字認識モデルへの知識蒸留[17]を適用する。具体的には、双方向言語モデルを教師、文字認識モデルを生徒として、文字認識モデルの文字の生成確率を双方向言語モデルの文字の生成確率に近づけるような損失関数を与え学習を行う。双方向言語モデルが導出する文字の生成確率は未来の文脈を考慮しなければ推定できないため、その推定を文字認識と同時に学習することで、未来の文脈を積極的に考慮した文字認識モデルの構築が可能になる。また、提案手法は学習時の損失関数を変更するのみで実現され、推論時に言語モデルを使用しないことから、推論時の計算量が増加しないという利点もある。同様の考え方は音声認識の改善にも導入されているが[18–20]、情景文字認識に特化して外部の言語モデルの知識を蒸留する手法は、本手法が初めてである。日本語の情景文字認識の実験により、提案手法の有効性を示す。

以降、2 章では関連研究として情景文字認識および知識蒸留について述べる。3 章では準備として、Transformer を用いた情景文字認識と MLM を定式化する。4 章では提案手法について述べ、5 章では評価実験とその結果について述べる。最後に、6 章でまとめを述べる。

† 日本電信電話株式会社,
NTT コンピュータ＆データサイエンス研究所
NTT Computer and Data Science Laboratories,
NTT Corporation

2. 関連研究

2.1 情景文字認識

情景文字認識は、情景画像から文字の領域を切り出した画像を入力として、そこに写る文字列を認識するタスクである。情景文字認識の実現のため、エンドツーエンド型のニューラルネットワークを用いる手法が複数提案されている[6,7]。初期の手法では、入力画像を CNN と双方向 Long Short-Term Memory（LSTM）により画像特徴に変換し、得られた画像特徴を Connectionist Temporal Classification（CTC）[21]や注意機構を用いた自己回帰型のデコーダにより文字列に変換する[22–25]。近年は、Transformer[8]を用いた自然言語処理のための深層学習技術の進展により、CNN と Transformer エンコーダからなるエンコーダが入力画像を画像特徴にエンコードし、Transformer デコーダからなるデコーダが画像特徴を自己回帰により文字列にデコードするモデル化手法が高い認識性能を達成している[9–12]。本研究では、このような Transformer に基づくエンコーダ・デコーダ構造による自己回帰型モデルをベースラインとし、言語的知識を強化する手法を提案する。

情景文字認識において、言語的知識を強化する手法はいくつか提案がある。Wang らの手法[26]や Lyu らの手法[27]では、Bidirectional Encoder Representations from Transformers（BERT）[16]の学習方法に着想を得て、学習時に入力画像の一部をマスクし、マスクされた領域を含めた文字列を推定することで、文字認識モデルの言語的知識を強化する。しかし、この手法は大量の画像と対応する文字列のペアデータを用いることでしか十分に学習することができない。これと異なり、提案手法はテキストのみで学習された外部の言語モデルの知識を用いることで、効率的に文字認識モデルを学習する。

言語的知識を高めるため、外部の言語モデルを利用する手法も提案されている。例えば Kang らの手法[14]では、文字認識モデルが出力する文字の生成確率と外部の言語モデルが出力する文字の生成確率をフュージョンすることにより、最終的な文字の生成確率を導出する。Sabir らの手法[28]や Wick らの手法[15]では、文字認識モデルが出力する複数の文字列候補を、外部の言語モデルを用いてリスコアリングすることで、言語的に尤もらしい文字列を得る。Qiao らの手法[29]では、言語モデルが出力する潜在表現に近い中間表現を予測するモジュールを文字認識モデルに追加し、推論時には予測された中間表現を参照しながら文字列をデコードすることで、言語モデルの知識を踏まえた文字認識を行う。Fang らの手法[30]では、デコード時に誤りを訂正する言語モデルを繰り返し適用することで、文字認識の誤りを訂正する。これらの手法は、過去の文脈に基づいて次の文字を予測する自己回帰型の文字認識モデルから得られた文字列やその生成確率を言語モデルにより変更するため、主に過去の文脈を考慮した文字認識の結果しか得ることができない。提案手法はこれと異なり、双方向言語モデルの知識を蒸留することで、文字認識モデル自体に未来の文脈を積極的に考慮させる。

2.2 知識蒸留

知識蒸留は、教師モデルである学習済みモデルの出力する分布に近い分布を生徒モデルが出力できるよう、生徒モデルを学習する手法である[17]。大規模で高性能なモデルを教師モデル、小規模なモデルを生徒モデルとし、生徒モデルが教師モデルの出力を模倣するように学習することで、モデルを軽量化することができる[17,31,32]。また、モデルの軽量化だけでなく、教師モデルに学習されている知識を別のタスク向けの生徒モデルに転移することで、生徒モデルの性能を向上させる手法も存在する[18–20]。提案手法はこれを応用し、文字認識モデルを教師、双方向言語モデルを生徒として、双方向言語モデルの知識を文字認識モデルに与えることで認識精度を向上させる。

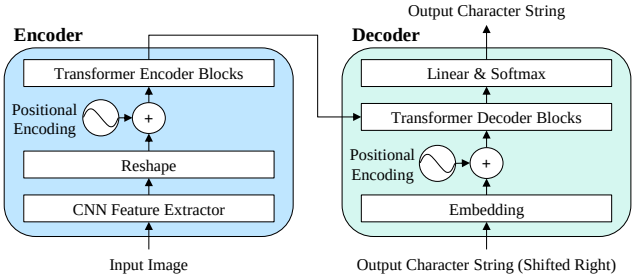

図 2　Transformer に基づく情景文字認識モデル

3. 準備

3.1 Transformer に基づく情景文字認識

情景文字認識は、文字が写る画像$\boldsymbol{I}$に対応する文字列$\boldsymbol{C} = \{c_1, \cdots, c_T\}$を認識するタスクである。ここで、$c_t$は$t$番目の文字であり、$T$は文字数である。Transformer に基づくエンコーダ・デコーダ構造による自己回帰型の情景文字認識モデルでは、$\boldsymbol{I}$に対する$\boldsymbol{C}$の生成確率を(1)式の通りモデル化する。

$$P(\boldsymbol{C}|\boldsymbol{I}; \boldsymbol{\Theta}_{\mathrm{STR}}) = \prod_{t=1}^{T} P(c_t|\boldsymbol{I}, c_{1:t-1}; \boldsymbol{\Theta}_{\mathrm{STR}}) \tag{1}$$

ここで、$c_{1:t-1} = \{c_1, \cdots, c_{t-1}\}$であり、$\boldsymbol{\Theta}_{\mathrm{STR}} = \{\boldsymbol{\theta}_{\mathrm{enc}}, \boldsymbol{\theta}_{\mathrm{dec}}\}$は学習可能なモデルのパラメータセットを表す。学習では、データ数Nの画像と文字列のペアによるデータセット$\mathcal{D} = \{(\boldsymbol{I}^1, \boldsymbol{C}^1), \cdots, (\boldsymbol{I}^N, \boldsymbol{C}^N)\}$を用い、(2)式により表されるクロスエントロピーの最小化によりパラメータを最適化する。

$$\begin{aligned} \mathcal{L}_{\mathrm{STR}} &= -\sum_{n=1}^{N} \log P(\boldsymbol{C}^n|\boldsymbol{I}^n; \boldsymbol{\Theta}_{\mathrm{STR}}) \\ &= -\sum_{n=1}^{N}\sum_{t=1}^{T^n} \log P(c_t^n|\boldsymbol{I}^n, c_{1:t-1}^n; \boldsymbol{\Theta}_{\mathrm{STR}}) \end{aligned} \tag{2}$$

ここで、T^nは$\boldsymbol{C}^n$の長さであり、c_t^nはn番目のデータにおけるt番目の文字である。また、$c_{1:t-1}^n = \{c_1^n, \cdots, c_{t-1}^n\}$である。

図 2 に、Transformer に基づく情景文字認識モデル[9]の構造を示す。図 2 に示される通り、モデルはエンコーダとデコーダで構成される。

3.1.1 エンコーダ

エンコーダは、(3–4)式に従い入力画像$\boldsymbol{I}$をベクトル系列$\boldsymbol{R}$に変換する。

$$\boldsymbol{Q} = \mathrm{CNNFeatureExtractor}(\boldsymbol{I}; \boldsymbol{\theta}_{\mathrm{enc}}) \tag{3}$$

$$\boldsymbol{R} = \mathrm{Reshape}(\boldsymbol{Q}) \tag{4}$$

ここで、CNNFeatureExtractor()は CNN により画像の特徴を抽出する関数であり、Reshape()は 3 次元の画像特徴（幅×高さ×チャネル）を 2 次元のベクトル系列（幅×（高さ×チャネル））に変換する関数である。さらに、ベクトル系列$\boldsymbol{R}$を、(5–6)式に従いベクトル系列$\boldsymbol{U}$に変換する。

$$\boldsymbol{S} = \mathrm{AddPosEnc}(\boldsymbol{R}) \tag{5}$$

$$\boldsymbol{U} = \mathrm{TransformerEnc}(\boldsymbol{S}; \boldsymbol{\theta}_{\mathrm{enc}}) \tag{6}$$

ここで、AddPosEnc()は位置埋め込みの関数であり、TransformerEnc()は Transformer エンコーダブロックである[8]。

3.1.2 デコーダ

デコーダは、エンコーダから出力されたベクトル系列を参照し、自己回帰により文字列の生成確率を求める。t番目の文字c_tの生成確率は、(7)式により得る。

$$P(c_t|\boldsymbol{I}, c_{1:t-1}; \boldsymbol{\Theta}_{\mathrm{STR}}) = \mathrm{Softmax}(\boldsymbol{v}_{t-1}; \boldsymbol{\theta}_{\mathrm{dec}}) \tag{7}$$

ここで、Softmax()はソフトマックス関数を活性化関数にもつ線形変換関数である。Transformer デコーダは、エンコーダの出力$\boldsymbol{U}$と$\boldsymbol{W}_{1:t-1} = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{t-1}\}$から、(8)式に従いベクトル$\boldsymbol{v}_{t-1}$を導出する。

$$\boldsymbol{v}_{t-1} = \mathrm{TransformerDec}(\boldsymbol{W}_{1:t-1}, \boldsymbol{U}; \boldsymbol{\theta}_{\mathrm{dec}}) \tag{8}$$

ここで、TransformerDec()は Transformer デコーダブロックである[8]。ベクトル$\boldsymbol{w}_{t-1}$は文字の埋め込み表現であり、(9–10)式により得る。

$$\boldsymbol{w}_{t-1} = \mathrm{AddPosEnc}(\boldsymbol{c}_{t-1}) \tag{9}$$

$$\boldsymbol{c}_{t-1} = \mathrm{Embedding}(c_{t-1}; \boldsymbol{\theta}_{\mathrm{dec}}) \tag{10}$$

ここで、Embedding()は文字を連続ベクトルに埋め込む線形変換関数である。

3.2 Masked Language Modeling

MLM は、BERT[16]の事前学習手法として提案されたものである。MLM では、入力された文の一部のトークンをマスクトークン<MASK>に置換し、過去だけでなく未来のトークンを含む周囲の文脈を考慮してマスクトークンに対応するトークンを予測する学習を、大規模なラベルなしのテキストコーパスを用いて行う。なお、本稿ではトークンを文字として定義する。この学習により、MLM で学習された双方向言語モデルは、t番目の文字がマスクされた文字列$\boldsymbol{C}_{\setminus t} = \{c_1, \cdots, c_{t-1}, \text{<MASK>}, c_{t+1}, \cdots, c_T\}$から t番目の文字c_tを(11)式のように推定することが可能である。

$$P(c_t|\boldsymbol{C}_{\setminus t}; \boldsymbol{\Theta}_{\mathrm{BLM}}) = P(c_t|c_{1:t-1}, c_{t+1:T}; \boldsymbol{\Theta}_{\mathrm{BLM}}) \tag{11}$$

ここで、$\boldsymbol{\Theta}_{\mathrm{BLM}}$は双方向言語モデルのパラメータである。

4. 提案手法

提案手法は、大規模なテキストコーパスで学習した過去と未来の両方の文脈を考慮する外部の双方向言語モデルの知識を、文字認識に活用する。このため、提案手法は双方向言語モデルを教師、文字認識モデルを生徒として、双方向言語モデルから文字認識モデルへの知識蒸留[17]を適用

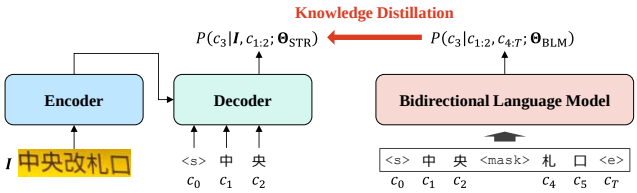

図 3　提案手法の概要

することで、文字認識モデルの言語的知識を強化する。具体的には、文字認識モデルが出力する文字の生成確率を、双方向言語モデルの出力する文字の生成確率に近づけるような損失関数を学習時に導入する。提案手法には 2 点の強みがある。1 点目は、双方向言語モデルが過去と未来の文脈を考慮して文字の生成確率を導出しており、文字認識モデルがこの生成確率を再現できるよう学習することで、文字認識モデルに過去だけでなく未来の文脈を積極的に考慮した生成確率を導出させ、未来の文脈を考慮した文字認識が可能となることである。2 点目は、提案手法が学習時の損失関数の変更のみで実現することができるため、推論時に言語モデルを用いず、計算量を増加させないことである。

図 3 に、提案手法の概要を示す。図 3 において、<s>と<e>はそれぞれ開始トークンと終了トークンを表す。提案手法は、$\mathcal{L}_{\mathrm{STR}}$に加え知識蒸留損失$\mathcal{L}_{\mathrm{KD}}$を用いて、文字認識モデルを学習する。知識蒸留損失は、(12)式に示される通り各教師データの各文字に対する文字認識モデルと双方向言語モデルの間の文字の生成確率の Kullback-Leibler 情報量の総和として定義される。

$$\mathcal{L}_{KD} = \sum_{n=1}^{N} \sum_{t=1}^{T^n} \mathrm{KL}\left(P(c_t^n|\boldsymbol{I}^n, c_{1:t-1}^n; \boldsymbol{\Theta}_{\mathrm{STR}}) \| P(c_t^n|\boldsymbol{C}_{\setminus t}^n; \boldsymbol{\Theta}_{\mathrm{BLM}})\right) \tag{12}$$

ここで、KL()は 2 つの分布間の Kullback-Leibler 情報量である。なお、$\boldsymbol{\Theta}_{\mathrm{BLM}}$は外部のラベルなしテキストを用いて事前学習済みであるものとし、$\boldsymbol{\Theta}_{\mathrm{STR}}$の学習時は固定する。全体の損失関数は、(13)式の通り定義される。

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\mathrm{STR}} + \alpha\mathcal{L}_{\mathrm{KD}} \tag{13}$$

ここで、αはハイパーパラメータである。

5. 評価実験

5.1 データセット

日本語の情景文字認識に対する評価実験のため、ICDAR2019 のコンペティションである MLT 向けのデータセット[13]と同様のフォーマットを持つ大規模な実画像の情景文字認識データセットを、Web から画像を大量に収集しアノテーションを付与することで作成した。なお、縦書きの文字列や 20 文字より長い文字列が含まれるデータは除外した。これにより作成された日本語情景文字認識データセットの詳細を、表 1 に示す。文字のクラス数は 9,649 であった。

言語モデルの学習には、Wiki-40B[33]の日本語データを用いた。学習用データの各パラグラフのテキストを収集しパラグラフごとに分割することで、約 500 万のテキストを収集し用いた。

5.2 実験条件

評価実験では、次の 5 つの手法を比較した。

- **ベースライン：**文字認識モデルを、言語モデルを用いずに学習し推論した。
- **Shallow fusion：**ベースラインの文字認識モデルが出力する文字の生成確率と、$P(c_t|c_{1:t-1})$をモデル化した単方向言語モデルが出力する文字の生成確率を、Shallow fusion[34]によりフュージョンすることで文字の生成確率を導出し推論した。ハイパーパラメータβは 0.1 に設定した。
- **Deep fusion：**各モデルの出力層の 1 層手前の中間表現を連結することにより、文字認識モデルと単方向言語モデルの Deep fusion[34]を実現するモデルを構築し、学習および推論を行った。なお、単方向言語モデルは文字認識モデルの学習時に固定するものとした。
- **リスコアリング：**ベースラインの文字認識モデルが出力する 4-best の文字列候補を、言語モデルを用いてリスコアリング[35]することで最終的な文字列を推論した。単方向言語モデルと双方向言語モデルの 2 つの言語モデルに対して評価を行った。
- **知識蒸留：**言語モデルからの知識蒸留を用いて文字認識モデルを学習し推論した。単方向言語モデルからの知識蒸留と、提案手法である双方向言語モデルからの知識蒸留を評価した。αは 0.1 に設定した。

なお、本評価実験は Deep fusion における連結に係る部分を除き文字認識モデルのアーキテクチャは共通とした。

文字認識モデルの入力サイズは 400×64 とし、入力画像がこれを満たすようリサイズした。辞書は、9,649 種別の文字に加え、開始トークン<s>、終了トークン<e>、マスクトークン<MASK>を登録したものを用いた。情景文字認識は 3.1 節で詳述したモデル[8]を用い、CNN 特徴抽出器として ImageNet[36]で事前学習済みの ResNet34[37]を用いた。Transformer について、Transformer エンコーダブロックの数を 4、Transformer デコーダブロックの数を 2 とし、出力次元数および全結合層における中間出力の次元数を 512、マルチヘッド注意のヘッド数を 4 とした。学習では、学習データの一部を学習の早期終了を判定するためのデータとして用い、5 エポック連続で損失関数の値が下がらないときに学習を打ち切った。最適化には学習率を 0.01 とした確率的勾配降下法を用い、バッチサイズは 8 とした。

双方向言語モデルおよび単方向言語モデルは、共通のアーキテクチャとして文字の埋め込み層、位置埋め込み、4 ブロックの Transformer エンコーダブロック、ソフトマックス層を用いた。Transformer エンコーダは出力次元数を 512、全結合層における中間出力の次元数を 2048、マルチヘッド注意のヘッド数を 4 として、文字認識モデルと共通の辞書を用いた。双方向言語モデルおよび単方向言語モデルはそれぞれ共通のテキストデータを用い 10 エポック学習した。最適化にはデフォルトパラメータの RAdam[38]を用い、バッチサイズは 128 とした。双方向言語モデルは、長さ 256 の文字列をサンプリングし、その内 12%の文字をランダムに選択しマスクトークンに置き換え、マスクトークンに対応する文字を予測する学習を行った。単方向言語モデルは、長さが 256 より短い文字列をランダムに抽出し、次の文字を予測する学習を行った。

表 1　日本語情景文字認識データセットの詳細

	画像数	文字数
学習データ	345,039	2,069,572
テストデータ	38,338	229,365

評価指標として、完全一致によるシーケンス誤り率と、文字誤り率を用いた。

5.3 実験結果

評価実験の結果を表 2 に示す。表 2 には、推論時に必要となるモデルの合計パラメータ数も示している。まず、ベースラインと Shallow fusion、Deep fusion、リスコアリングを比較すると、Shallow fusion を除き外部の言語モデルを用いた従来手法が文字認識の誤りを抑制するのに有効であることが分かる。Shallow fusion は、ベースラインが出力する次の文字の生成確率に単方向言語モデルが出力する次の文字の生成確率を加算することで、文字の生成確率を導出しているが、本評価実験では単方向言語モデルが出力する生成確率がノイズとして振る舞ったため、誤り率が上昇したものと思われる。なお、これらの外部の言語モデルを用いた従来手法は、推論時に言語モデルを繰り返し適用するため、ベースラインと比較し推論時に必要となるモデルの合計パラメータ数が増加し、計算量が増加する。一方、双方向言語モデルを用いた知識蒸留である提案手法は、ベースラインや他の従来手法を上回る性能を示していることが分かる。また、単方向言語モデルを用いた知識蒸留と、双方向言語モデルを用いた知識蒸留の比較により、双方向言語モデルを用いること、すなわち未来の文脈を考慮することの有効性が分かる。文字認識結果の例を、図 4 に示す。図 4 から、提案手法は不鮮明な画像や読みにくい文字を含む画像に対して、言語的に意味を成さない文字列の出力を防止する効果を有し、正確な文字認識を可能とすることが分かる。これらの結果から、本手法は従来手法よりも強力に言語的な知識を用いる文字認識を実現し、文字認識の性能を向上させたといえる。さらに、提案手法は推論時に言語モデルを用いないため、表 2 に示される通り、他の従来手法と異なりベースラインと同等のモデルのパラメータ数で推論することができる。以上から、提案手法は過去だけでなく未来の文脈を含む言語的知識を活用し、情景文字認識の性能を向上させる有効な手段であることが示された。

6. おわりに

本稿では、高精度な情景文字認識のための知識蒸留手法を提案した。提案手法は、MLM によって学習された外部の双方向言語モデルの知識を利用することで、文字認識モデルに過去と未来の両方の文脈を考慮した認識を可能にすることを図った。そのため、双方向言語モデルを教師、文字認識モデルを生徒として知識蒸留を適用し、双方向言語モデルに獲得されている言語的知識を文字認識モデルに与えた。具体的には、文字認識モデルの出力する文字の生成確率を双方向言語モデルの出力する文字の生成確率に近づけるための損失関数を導入し学習を行った。日本語情景文字認識の評価実験により、提案手法が推論時の計算コストを増加させることなく文字認識の精度を向上させることを確認し、双方向言語モデルからの知識蒸留が情景文字認識

表 2　評価実験の結果

	シーケンス誤り率 (%)	文字誤り率 (%)	推論時のパラメータ数
(a) ベースライン	21.03	10.22	**69.4M**
(b) Shallow fusion w/ 単方向言語モデル	21.03	10.30	91.9M
(c) Deep fusion w/ 単方向言語モデル	20.29	9.99	91.9M
(d) リスコアリング w/ 単方向言語モデル	20.51	10.08	91.9M
(e) リスコアリング w/ 双方向言語モデル	19.94	9.98	91.9M
(f) 知識蒸留 w/ 単方向言語モデル	20.28	10.07	**69.4M**
(g) 知識蒸留 w/ 双方向言語モデル（提案手法）	**19.86**	**9.83**	**69.4M**

図 4　認識結果の例．(GT)は Ground Truth を表す．他は表 2 と同様に，(a)ベースライン，(b)Shallow fusion，(c)Deep fusion，(d)単方向言語モデルによるリスコアリング，(e)双方向言語モデルによるリスコアリング，(f)単方向言語モデルによる知識蒸留，(g)双方向言語モデルによる知識蒸留（提案手法）である．

における言語的知識を強化する有効なアプローチであることが示された。

参考文献

[1] S. Karaoglu, R. Tao, T. Gevers, and A. W. M. Smeulders, “Words matter: Scene text for image classification and retrieval,” IEEE Transactions on Multimedia, Vol.19, No.5, pp.1063–1076 (2017).

[2] S. Karaoglu, R. Tao, J. C. v. Gemert, and T. Gevers, “Con-Text: Text detection for fine-grained object classification,” IEEE Transactions on Image Processing, Vol.26, No.8, pp.3965–3980 (2017).

[3] X. Bai, M. Yang, P. Lyu, Y. Xu, and J. Luo, “Integrating scene text and visual appearance for fine-grained image classification,” IEEE Access, Vol.6, pp.66322–66335 (2018).

[4] L. Gomez, A. Mafla, M. Rusinol, and D. Karatzas, “Single shot scene text retrieval,” in Proceedings of the European Conference on Computer Vision (ECCV), pp.728–744 (2018).

[5] A. F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusinol, E. Valveny, C. Jawahar, and D. Karatzas, “Scene text visual question answering,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.4291–4301 (2019).

[6] S. Long, X. He, and C. Yao, “Scene text detection and recognition: The deep learning era,” International Journal of Computer Vision, Vol.129, No.1, pp.161–184 (2021).

[7] X. Chen, L. Jin, Y. Zhu, C. Luo, and T. Wang, “Text recognition in the wild: A survey,” ACM Computing Surveys, Vol.54, No.2 (2021).

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Łukasz, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems (NIPS), pp.5998–6008 (2017).

[9] F. Sheng, Z. Chen, and B. Xu, “NRTR: A no-recurrence sequence-to-sequence model for scene text recognition,” in Proceedings of the IEEE International Conference on Document Analysis and Recognition (ICDAR), pp.781–786 (2019).

[10] D. Yu, X. Li, C. Zhang, J. Han, J. Liu, and E. Ding, “Towards accurate scene text recognition with semantic reasoning networks,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.12113–12122 (2020).

[11] L. Yang, P.Wang, H. Li, Z. Li, and Y. Zhang, “A holistic representation guided attention network for scene text recognition,” Neurocomputing, Vol.414, pp.67–75 (2020).
[12] N. Lu, W. Yu, X. Qi, Y. Chen, P. Gong, R. Xiao, and X. Bai, “MASTER: Multi-aspect non-local network for scene text recognition,” Pattern Recognition, Vol.117, pp.107980 (2021).
[13] N. Nayef, Y. Patel, M. Busta, P. N. Chowdhury, D. Karatzas, W. Khlif, J. Matas, U. Pal, J.-C. Burie, C. Liu, and J.-M. Ogier, “ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition,” in Proceedings of the IEEE International Conference on Document Analysis and Recognition (ICDAR), pp.1582–1587 (2019).
[14] L. Kang, P. Riba, M. Villegas, A. Fornes, and M. Rusinol, “Candidate fusion: Integrating language modelling into a sequence-to-sequence handwritten word recognition architecture,” Pattern Recognition, Vol.112, pp.107790 (2021).
[15] C. Wick, J. Zollner, and T. Gruning, “Rescoring sequence-to-sequence models for text line recognition with CTC-prefixes,” in Proceedings of the IAPR International Workshop on Document Analysis Systems (DAS), pp.260–274 (2022).
[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp.4171–4186 (2019).
[17] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in Proceedings of the Deep Learning and Representation Learning Workshop, NIPS (2015).
[18] Y. Bai, J. Yi, J. Tao, Z. Tian, and Z. Wen, “Learn spelling from teachers: Transferring knowledge from language models to sequence-to-sequence speech recognition,” in Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp.3795–3799 (2019).
[19] H. Futami, H. Inaguma, S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, “Distilling the knowledge of BERT for sequence-to-sequence ASR,” in Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp.3635–3639 (2020).
[20] R. Masumura, N. Makishima, M. Ihori, A. Takashima, T. Tanaka, and S. Orihashi, “Hierarchical transformer-based large-context end-to-end ASR with large-context knowledge distillation,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5879–5883 (2021).
[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in Proceedings of the International Conference on Machine Learning (ICML), pp.369–376 (2006).
[22] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, “Robust scene text recognition with automatic rectification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4168–4176 (2016).
[23] B. Shi, X. Bai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.39, No.11, pp.2298–2304 (2016).
[24] F. Borisyuk, A. Gordo, and V. Sivakumar, “Rosetta: Large scale system for text detection and recognition in images,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), pp.71–79 (2018).
[25] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S.J. Oh, and H. Lee, “What is wrong with scene text recognition model comparisons? Dataset and model analysis,” Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.4714–4722 (2019).
[26] Y. Wang, H. Xie, S. Fang, J. Wang, S. Zhu, and Y. Zhang, “From two to one: A new scene text recognizer with visual language modeling network,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.14194–14203 (2021).
[27] P. Lyu, C. Zhang, S. Liu, M. Qiao, Y. Xu, L. Wu, K. Yao, J. Han, E. Ding, and J. Wang, “MaskOCR: Text recognition with masked encoder-decoder pretraining,” arXiv preprint, arXiv:2206.00311, (2022).
[28] A. Sabir, F. Moreno-Noguer, and L. Padr´o, “Visual re-ranking with natural language understanding for text spotting,” in Proceedings of the Asian Conference on Computer Vision (ACCV), pp.68–82 (2018).
[29] Z. Qiao, Y. Zhou, D. Yang, Y. Zhou, and W. Wang, “SEED: Semantics enhanced encoder-decoder framework for scene text recognition,” in Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR), pp.13528–13537 (2020).
[30] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, “Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.7098–7107 (2021).
[31] C. Bucilua, R. Caruana, and A. Niculescu-Mizil, “Model compressions,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp.535–541 (2006).
[32] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, “Improved knowledge distillation via teacher assistant,” in Proceedings of the AAAI conference on Artificial Intelligence (AAAI), Vol.34, No.4, pp.5191–5198 (2020).
[33] M. Guo, Z. Dai, D. Vrandecic, and R. Al-Rfou, “Wiki-40B: Multilingual language model dataset,” in Proceedings of the Conference on Language Resources and Evaluation (LREC), pp.2440–2452 (2020).
[34] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, “On using monolingual corpora in neural machine translation,” arXiv preprint, arXiv:1503.03535 (2015).
[35] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, “Masked language model scoring,” in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp.2699–2712 (2020).
[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.248–255 (2009).
[37] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.770–778 (2016).
[38] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, “On the variance of the adaptive learning rate and beyond,” in Proceedings of the International Conference on Learning Representations (ICLR) (2019).