

深層学習を用いた 学術用語解説ウェブページの見易さの自動評定

塩川 隼人[†] 春日 孝秀^{††} 韓 炳材[†] 宇津呂武仁^{†††} 河田 容英^{††††}

[†] 筑波大学 大学院システム情報工学研究科 〒305-8573 茨城県つくば市天王台 1-1-1

^{††} 筑波大学 理工学群 工学システム学類 〒305-8573 茨城県つくば市天王台 1-1-1

^{†††} 筑波大学 システム情報系 知能機能工学域 〒305-8573 茨城県つくば市天王台 1-1-1

^{††††} (株) ログワークス 〒151-0053 東京都渋谷区代々木 1-3-15 天翔代々木ビル 6F

あらまし 本論文では、ウェブ上で学術用語を解説するページ群を対象として、それらのページ群における用語解説ウェブページの見易さを自動評定する手法を提案する。本論文では、特に、用語解説ウェブページを画像化し、用語解説が見やすいウェブページ、用語解説が見にくいウェブページの二種類に分類した参照用教師事例を利用して、畳み込みニューラルネットの訓練を行い、用語解説ウェブページの見易さを識別する方式を提案する。

キーワード 学術用語解説ウェブページ, 深層学習, 見易さ, 学習教材読解支援システム, ウェブ学習, 検索エンジン, 畳み込みニューラルネット

Automatic Recognition of Visual Intelligibility of Web Pages explaining Academic Concepts by Deep Learning

Hayato SHIOKAWA[†], Takahide KASUGA^{††}, Bingcai HAN[†], Takehito UTSURO^{†††}, and Yasuhide KAWADA^{††††}

[†] Grad. Sch. of Systems and Information Engineering, University of Tsukuba, Tsukuba 305-8573 Japan

^{††} College Eng. Sys., School Sci. and Eng., University of Tsukuba, Tsukuba, 305-8573, Japan

^{†††} Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba 305-8573 Japan

^{††††} Logworks Co., Ltd. Tokyo 151-0053, Japan

1. はじめに

インターネットが普及した現在、ウェブ上には多くの学術関連のコンテンツが存在する。このコンテンツを使って学術用語を学ぶことは、学校で使用される教科書・参考書による学習とは別に、学習手段の一つとしてたびたび用いられている。インターネットで学術用語を学ぶ際、ウェブ検索に頼れば関連ページは容易に見つかる。しかし、「分かり易い用語解説」を見つけるには、検索上位ページを一つずつ見比べ読み進める非効率な作業が必要となる。これは、ウェブ検索では、分かり易さをベースにした解説ページ群の体系化ができないことが原因である。そこで、本研究では、検索上位ページの中で分かり易さを充足する必要十分な数の用語解説ページを見つけ体系化することを目的とする。さらに、この研究においては、最終的に、用語解説を提示してウェブ学習を促進する学習教材読解支援システムを実現する。この背景のもとで、本研究では、図 1 に示す

ように、6 個の個別因子を手がかりとして、分かり易さの全体評定を行う手法を構築することを目的とする。

以上の目的をふまえて、本論文では、6 個の個別因子の中から「レイアウトの見易さ」についての因子を対象とし、深層学習を用いてこの因子を自動評定することを目的とする。理工系学術分野における学術用語を対象クエリとし、それらのクエリを検索エンジンによって検索した上位 10 位以内のページのスクリーンショットを画像化したものに対して、「レイアウトの見易さ」の適否を手動で付与した参照用事例を作成する。この参照用事例に対して畳み込みニューラルネットワーク (CNN) を適用して、「レイアウトの見易さ」の適否の自動判定を行うモデルの訓練・評価を行った。

2. 用語解説ウェブページの「分かり易さ」評定の因子

本節では、用語解説ウェブページの個別因子および全体評定

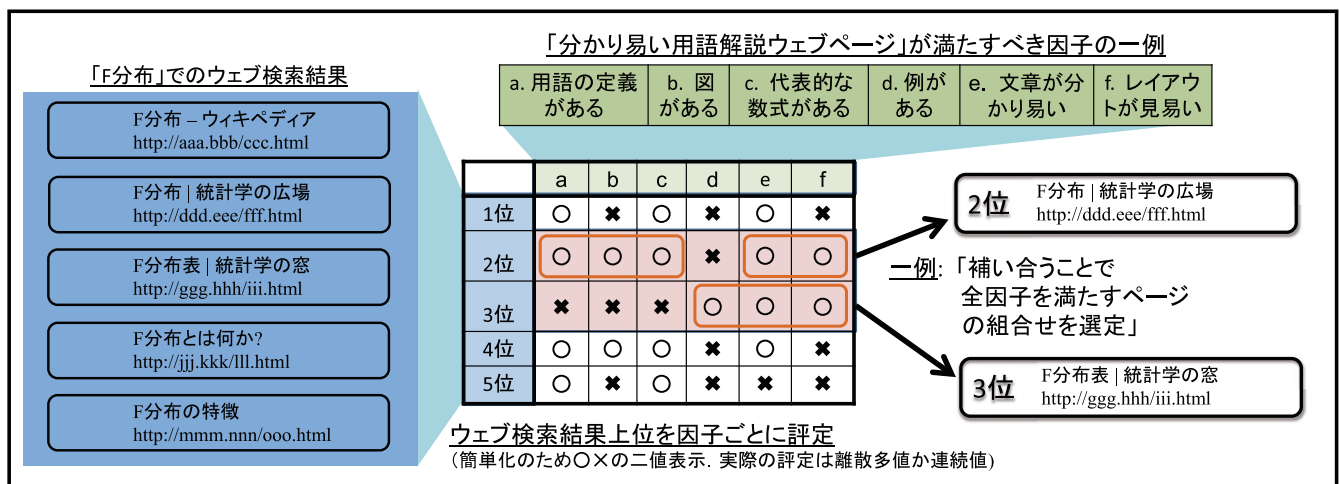


図 1 用語解説ウェブページの「分かり易さ」評価の因子

の評価基準について述べる。これらの評価においては、本論文の著者が二値での評価を行った。

2.1 個別因子

(a) 用語の定義がある

必要十分な用語の定義を述べている。例えば、統計分野の用語を検索クエリとしたとき、エクセル操作の解説ページがあり、学術用語に関する説明がされず、応用的な内容のみであれば定義無しとする。

(b) 図がある

図 2 の例に示すように、用語解説や例示の為に使用されている図を一つ以上含む。文字のみの表などは対象としない。

(c) 数式がある

図 3 の例に示すように、一般的に数式と呼べるものを含む。アルファベット一文字等は含めない。

(d) 例がある

図 4 の例に示すように、例題や、「例えば、～のように」などの例示を含む。例示については、用語を説明するための短すぎない例示に絞る。一つでも例題、例示と判断できるものがあれば良いものとした。

(e) レイアウトが見易い

ページのレイアウトの見易さ。学術用語解説ウェブページの見易さに特化する、というよりは、一般的なウェブページの見易さに準ずる。例えば、シンプルさ、改行や章題による構成の工夫、色遣い等が挙げられる。

(f) 文章が分かり易い

用語解説ページの文章の分かり易さ。そのページで扱う用語と、対象とするページ閲覧者層を考慮し、総合的な評価で行う。具体的には、例えば、専門用語と平易な用語のバランス、情報量、文章量等を基準に判断を行った。

2.2 全体評価

6 個の個別因子を元に、ウェブページ全体の評価付けを行う。

全体評価の評価は、個別因子の評価がどの程度満たされているかを基準として判断しているが、個別因子全てが充足されている必要はなく、ある程度充足されていれば、全体評価も充足すると評価している。例えば、「図あり」や「式あり」などの因子は必ずしも必要ではなく、これらの因子が欠けていたとしても全体評価が充足されないわけではない。6 個の因子の中で特に重要なのは「レイアウトが見易い」と「文章が分かり易い」のため、この両方が欠けている場合、あるいはどちらか片方が欠けていて、かつ、その他の個別因子が複数欠けている場合は、全体評価を「充足しない」と評価している。

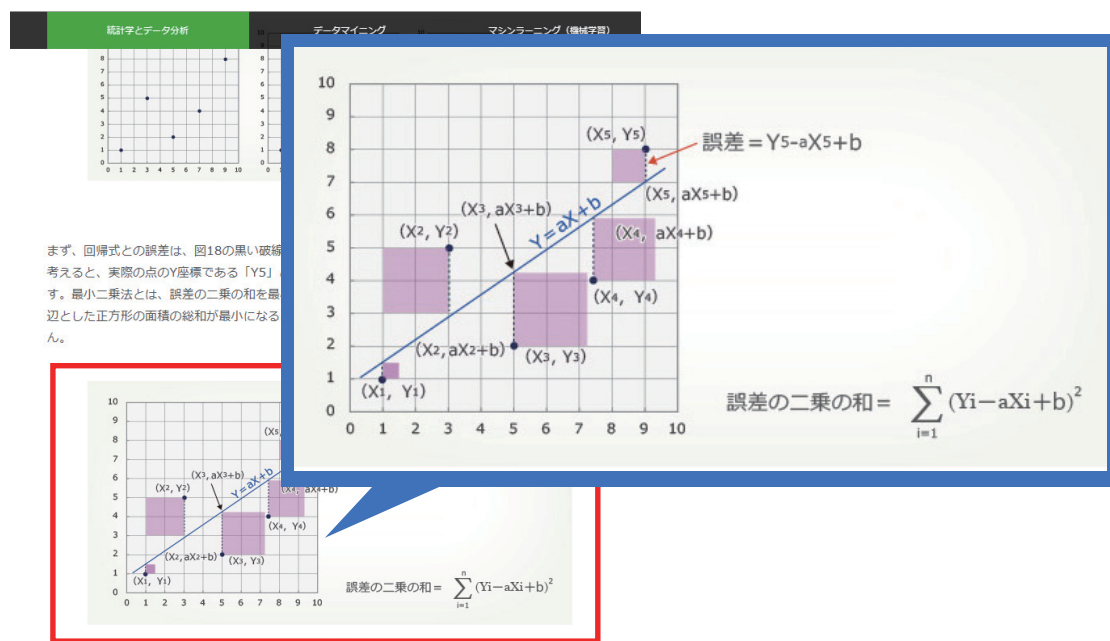
3. 参照学術用語解説ウェブページ集合の作成

3.1 対象学術分野および用語

事前調査を行い、「レイアウトの見易さ」、「文章の分かり易さ」、「全体評価」の評価基準が似ている理工系学術分野である統計・線形代数・物理・生物・化学・プログラミング・IT 分野を本研究での対象分野として選んだ。次に、各分野の用語の内、主に高校 3 年生から大学の学校教育レベルのものの中から、表 1 に示すように、統計のみ 30 語 (内 15 用語のみ対象)、物理・化学・プログラミング・IT は 15 語、線形代数のみ 14 語、生物のみ 7 語を対象学術用語として選んだ。

3.2 作成手順および作成結果

まず、前節で収集した各学術用語を検索クエリとし、検索エンジンの検索結果上位 10 件のウェブページを収集する。そして、2.2 節で述べた基準に基づき全体評価の判定を手で行い、その判定結果を付与した事例を蓄積する。以上の手順によって、全体評価が「充足する」と判定された事例を正例、「充足しない」と判定された事例を負例として、各分野ごとに表 2 に示す数の正例・負例を収集した。



回帰係数はどのように求めるか

回帰分析は予測することが目的のひとつでした。身長から体重を予測する、母親の身長から子供の身長を予測するなどです。相関関係を「 $Y = aX + b$ 」の一次方程式で表せたとすると、定数の a （傾き）

図 2 用語解説ウェブページにおける図の例 (https://www.albert2005.co.jp/knowledge/statistics_analysis/multivariate_analysis/single_regression より抜粋)

表 1 学術分野ごとの対象学術用語一覧

分野	対象とする学術用語数	対象とする学術用語
統計	30(内 15 用語のみ対象)	正規分布, F 分布, 信頼区間, 標準偏差, 主成分分析, 回帰分析, 自己回帰モデル, ポアソン分布, マルコフ連鎖モンテカルロ法, 確率密度関数, コーシー分布, 分散, 確率, 事後分布, 事前分布, 相関係数, 共分散, 自己回帰, 中央値, ガンマ分布, ノンパラメトリック手法, パレート分析, 十分統計量, ホワイトノイズ, 独立成分分析, 多変量解析, 帰無仮説, ラプラス分布, モーメント, k 平均法 (計 30 語)
線形代数	14	階数, 共役勾配, 行列式, クラメル公式, クロネッカーのデルタ, 三角行列, 正規直交基底, 対角化, 直交行列, 特性多項式, 二次形式, ノルム, メネラウスの定理, ヤコビ行列 (計 14 語)
物理	15	電気力線, 張力, 慣性の法則, 遠心力, 電波, 電流, 万有引力, 交流, 音波, ホイットストンブリッジ, 反発係数, 相互誘導, 正電荷, 速度, 変圧器 (計 15 語)
生物	7	DNA, ショウジョウバエ, 原核生物, 減数分裂, 光合成, 細胞, 葉緑体 (計 7 語)
化学	15	イオン結合, エステル, カルボン酸, ケトン, 化学反応式, 化学平衡, 共有結合, 合成高分子, 酸化還元, 遷移元素, 典型元素, 天然高分子, 燃料電池, 物質の三態, 芳香族有機化合物 (計 15 語)
プログラミング	15	C 言語, Java, エスケープシーケンス, コマンドライン引数, スコープ, フィールド値, ポインタ, メソッド, 繰り返し処理, 構造体, 算術演算子, 条件分岐, 配列変数, 文字列, 論理演算 (計 15 語)
IT	15	API, DBMS, HTML, IP アドレス, JDBC, RDB, SDK, SQL, Unicode, URL, スコープマネジメント, ステークホルダーマネジメント, タイムマネジメント, ナレッジマネジメント, リスクマネジメント (計 15 語)
合計	96	—

4. 深層学習による学術用語解説ウェブページの見易さの自動評定

4.1 VGG16 [6]

近年、深層学習は多様な分野のタスクにおいて、その有用性

を示している。特に画像認識の分野では、畳み込みニューラルネットワーク (CNN) と ImageNet などの大規模なデータセッ

● 対角行列 Λ の導出

一般に、対角化された行列は、対角成分がもとの行列の固有値になることが知られている。よって、 A の固有値を求めて、対角成分に並べれば、対角化した行列 Λ が得られる。
 A の固有値 λ を求めるには、固有方程式

$$|\lambda I - A| = 0 \quad (1)$$

を λ について解けばよい。左辺は3行3列の行列式であるので、

$$\begin{aligned} |\lambda I - A| &= \begin{vmatrix} \lambda & 1/2 & 3/2 \\ -1 & \lambda - 3/2 & -3/2 \\ 1 & -1/2 & \lambda - 1/2 \end{vmatrix} \\ &= \lambda \left(\lambda - \frac{3}{2} \right) \left(\lambda - \frac{1}{2} \right) + \frac{1}{2} \left(-\frac{3}{2} \right) 1 + \frac{3}{2} (-1) \left(-\frac{1}{2} \right) \\ &\quad - \frac{3}{2} \left(\lambda - \frac{3}{2} \right) 1 - \frac{1}{2} (-1) \left(\lambda - \frac{1}{2} \right) - \lambda \left(-\frac{3}{2} \right) \left(-\frac{1}{2} \right) \\ &= \lambda^3 - 2\lambda^2 - \lambda + 2 \end{aligned}$$

である。よって、(1) は、

$$\lambda^3 - 2\lambda^2 - \lambda + 2 = 0$$

と表される。3次方程式であるので、解くのは簡単ではない。

$$(\lambda + 1)(\lambda - 2)(\lambda - 1) = 0$$

となるため、解は

$$\lambda = -1, 1, 2$$

である。固有値が求められたので、対角行列 Λ は、

$$\Lambda = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

$$\begin{aligned} |\lambda I - A| &= \begin{vmatrix} \lambda & 1/2 & 3/2 \\ -1 & \lambda - 3/2 & -3/2 \\ 1 & -1/2 & \lambda - 1/2 \end{vmatrix} \\ &= \lambda \left(\lambda - \frac{3}{2} \right) \left(\lambda - \frac{1}{2} \right) + \frac{1}{2} \left(-\frac{3}{2} \right) 1 + \frac{3}{2} (-1) \left(-\frac{1}{2} \right) \\ &\quad - \frac{3}{2} \left(\lambda - \frac{3}{2} \right) 1 - \frac{1}{2} (-1) \left(\lambda - \frac{1}{2} \right) - \lambda \left(-\frac{3}{2} \right) \left(-\frac{1}{2} \right) \\ &= \lambda^3 - 2\lambda^2 - \lambda + 2 \end{aligned}$$

図 3 用語解説ウェブページにおける式の例 (<http://physmath.main.jp/src/matrix-diagonalization-example-3-3.html> より抜粋)

表 2 対象学術分野ごとの正例数及び負例数

対象学術分野	正例数	負例数	合計
統計	66	84	150
線形代数	35	105	140
物理	46	104	150
生物	14	56	70
化学	52	98	150
プログラミング	58	92	150
IT	69	81	150
合計	340	620	960

トの登場により、様々なタスクにおいて驚異的な性能を発揮している。また、ImageNet のような大規模なデータセットを用いて学習を行った CNN のパラメータは、その汎用性の高さから高性能な特徴抽出器として異なるドメインのタスクにおいても活用できることが知られている。

そこで、本論文では、用語解説ウェブページの「レイアウトの見易さ」を評定するタスクにおいて、VGG16 モデル [6] を基盤の特徴抽出器として用いる。VGG16 は ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014 [4] の「画像分類タスク (Image classification)」で準優勝、「位置特定タスク (Single-object localization)」で優勝したモデルであり、このモデルは 13 層の畳み込み層、その途中に挟まれている 5 層

の最大プーリング層、3 層の全結合層 (1,000 値分類用)、および、出力層から成り立っている。訓練済みの VGG16 としては、ImageNet2014 のデータセットによって 1,000 値分類用に訓練済みのモデルが一般に公開されており、この訓練済みモデルは他のタスクにも広く転用可能なことが知られている。本論文では、Python の深層学習ライブラリ Keras^(注1) 用に公開されているモデル^(注2) を利用して評価実験を行った。

4.2 訓練・評価手順

本論文では、ImageNet データセットを用いて訓練済みの

(注1) : <https://keras.io/ja/>

(注2) : <https://github.com/keras-team/keras/blob/master/keras/applications/vgg16.py>

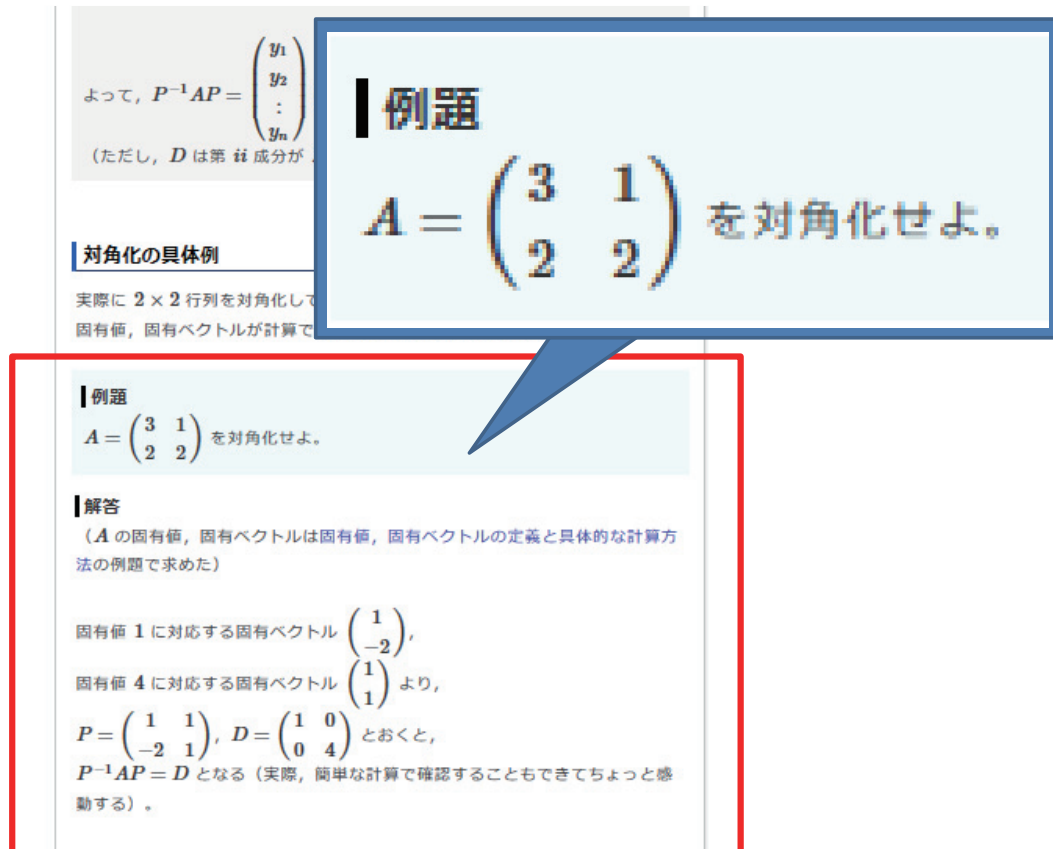


図 4 用語解説ウェブページにおける例示の例 (<https://mathtrain.jp/diagonalization> より抜粋)

VGG16 モデルを基盤に用いて用語解説ウェブページの「レイアウトの見易さ」の判定の評価実験を行った。参照用学術用語解説ウェブページデータとして、3. 節で作成した用語解説ウェブページ事例を用い、各ウェブページのトップページを画像化したものを実際の事例として用いる。この訓練・評価事例集合の正例・負例数は、表 2 の通りである。

本論文において VGG16 を利用して「レイアウトの見易さ」を評定するモデルを訓練する際には、まず、VGG16 の 1,000 値分類用の 3 層の全結合層と出力層を取り払い、2 クラス分類用の全結合層 3 層と出力層を接続し、以下の二種類のモデルを用意する。

(1) 「fine-tuning-VGG16 ありのモデル」

VGG16 の畳み込み層 10 層、および、最大プーリング層第 4 層目までのパラメータを VGG16 の訓練済みのパラメータに固定し、それ以降の畳み込み層 3 層、最大プーリング層 1 層、全結合層 3 層に対して、本論文の「レイアウト見易さ」判定タスク用の訓練事例を用いて、再度パラメータの訓練を行う。

(2) 「fine-tuning-VGG16 なしのモデル」

VGG16 の畳み込み層 13 層、および、最大プーリング層 5 層のパラメータを VGG16 の訓練済みのパラメータに固定し、それ以降の全結合層 3 層に対して、本論文の「レイアウト見易さ」判定タスク用の訓練事例を用いて、再度パラメータの訓練を行う。

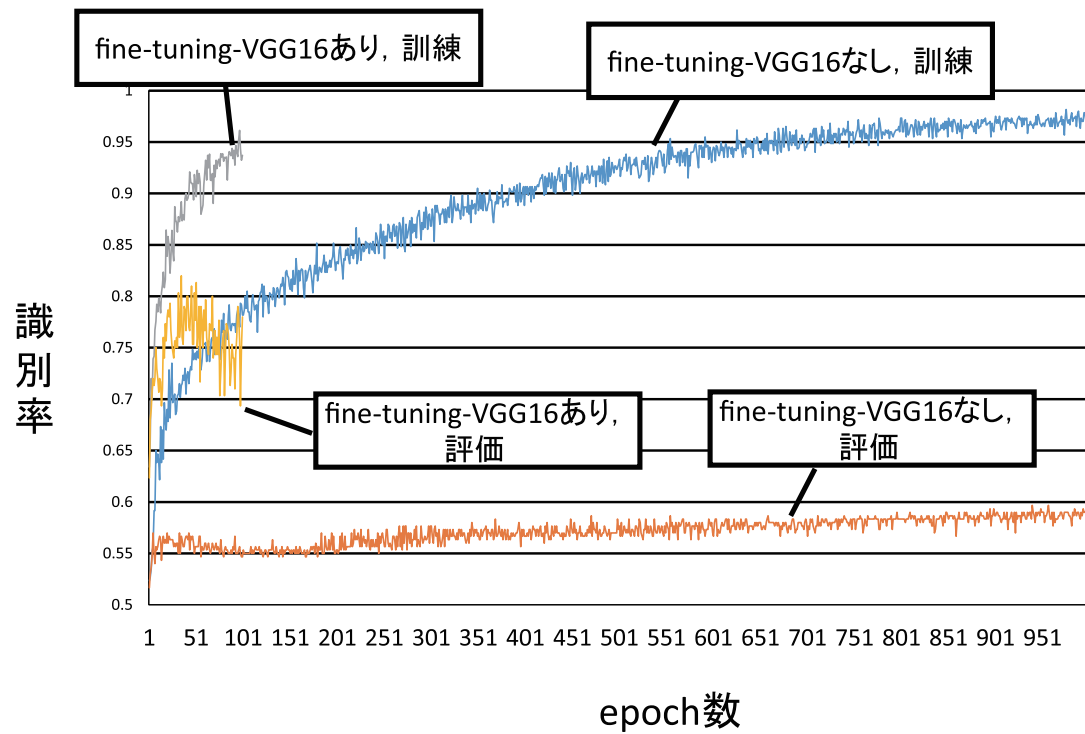
例を用いて、パラメータの訓練を行う。

4.3 評価結果

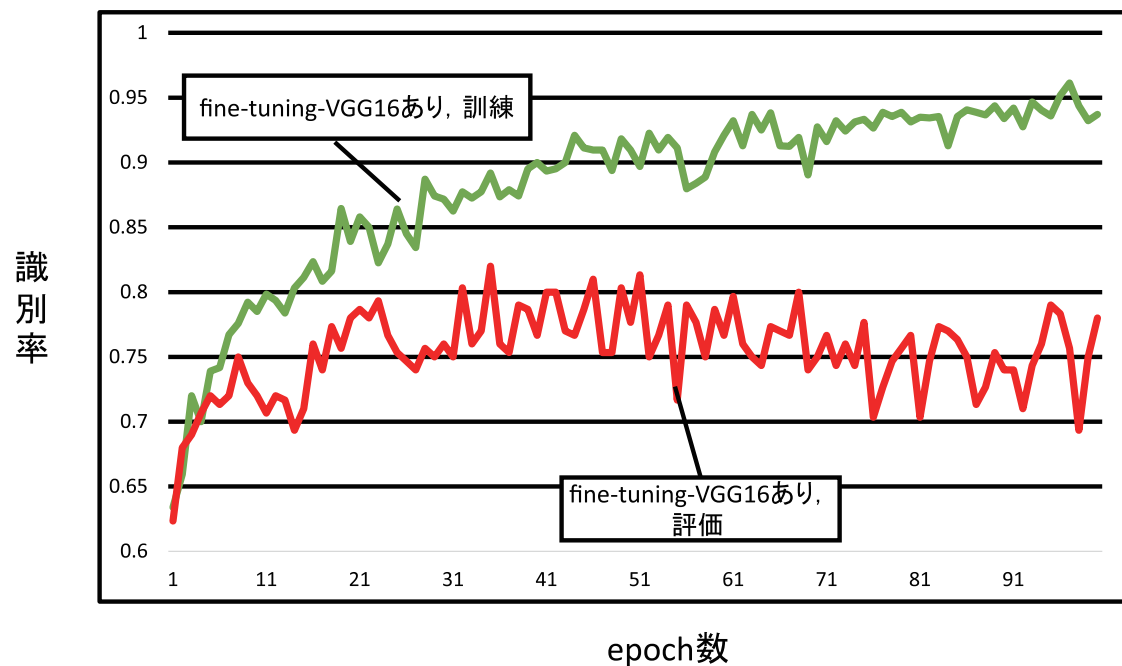
対象学術分野のうち、化学と統計分野に属する学術用語をクエリとして収集した参照用学術用語解説ウェブページ集合を評価事例とし、それ以外の事例を訓練事例として、評価実験を行った。学習曲線を図 5 に示す。「fine-tuning-VGG16 あり」のモデルの方が、少ない epoch 数でより高い性能を達成していることが分かる。また、「レイアウトの見易いウェブページ」、「レイアウトの見易くないウェブページ」の別に、VGG16 モデルの識別の信頼度の下限値を変化させて、下限値を超える事例のみを評価対象として再現率・適合率曲線をプロットした結果を図 6 に示す。この結果から、「fine-tuning-VGG16 ありのモデル」の有効性が確認でき、かつ、「レイアウトの見易くないウェブページ」の方が相対的に識別が容易であることが分かる。また、信頼度の下限値なしの場合では、70%程度の適合率のもとで全評価事例の二値判定が達成できている。

5. 関連研究

分かりやすさの全体評定を行うのに必要な個別因子のうち、見易さ以外の因子の自動評定については、本論文の関連タスクとして、[2] において、HTML 構造上の特徴を利用した見易さ以外の個別因子及び、全体評定の自動評定の手法を提案している。この研究と本稿で述べた手法を組み合わせることによって、個別因子全ての自動評定が可能になり、本研究の目的である用



(a) fine-tuning-VGG16 あり / なし



(b) fine-tuning-VGG16 ありのみ

図 5 学習曲線

語解説ページの体系化を行うことができる。

この学術用語解説ウェブページの分かり易さ自動評定タスクに関連して、コミュニティ型質問応答の分野においては、回答者による回答の良質さを自動評定する手法についての研究が行われている [1, 5]。また、特に、本論文における学術用語解説ウェブページの見易さの自動評定に関する関連タスクとして、[3] においては、本論文とほぼ同様の深層学習手法によって、プレゼンテーションスライドの画像情報に対する印象を予

測する手法を提案している。

6. おわりに

本論文では、学術用語解説ウェブページの分かり易さの全体評定を行うために必要な個別因子のうち、「レイアウトの見易さ」について、深層学習により自動評定を行う手法を提案した。学術用語解説ウェブページのトップ画面を画像化して評価実験を行い、高い性能を達成することができた。今後の課題としては、[2] における全体評定の自動評定手法において本論文の出

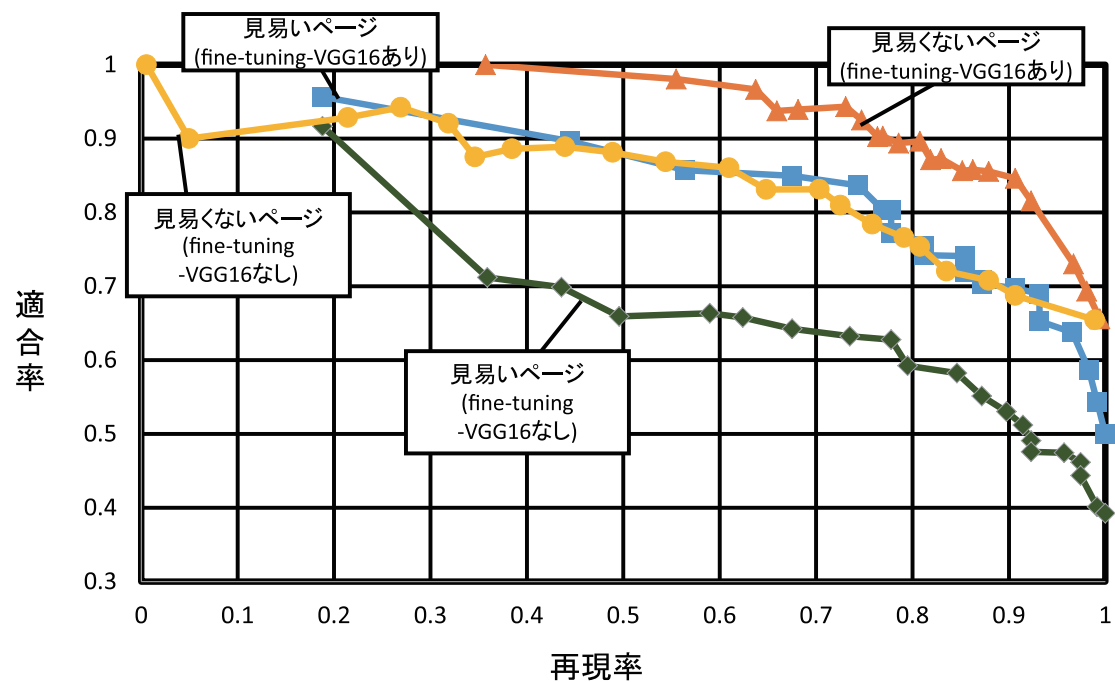


図 6 評価結果

力を組み込むことによって、全体評定の性能向上を実現することが挙げられる。

文 献

- [1] 石川大介, 酒井哲也, 関洋平, 栗山和子, 神門典子. コミュニティ QA における良質回答の自動予測. 情報知識学会誌, Vol. 21, No. 3, pp. 362–382, 2011.
- [2] 春日孝秀, 塩川隼人, 韓炳材, 宇津呂武仁, 河田容英. HTML 構造上の特徴を利用した学術用語解説ウェブページの分かり易さの自動評定. 第 10 回 DEIM フォーラム論文集, 2018.
- [3] 大山真司, 山崎俊彦, 相澤清晴. プレゼンテーションスライドの客観評価と印象予測. 第 16 回 FIT 講演論文集, 第 3 巻, pp. 45–52, 2017.
- [4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F.-F. Li. ImageNet large scale visual recognition challenge. *CoRR*, Vol. abs/1409.0575, , 2014.
- [5] T. Sakai, D. Ishikawa, N. Kando, Y. Seki, K. Kuriyama, and C.-Y. Lin. Using graded-relevance metrics for evaluating community QA answer selection. In *Proc. 4th WSDM*, pp. 187–196, 2011.
- [6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. 3rd ICLR*, 2015.