画像アノテーション課題からテキスト分類課題へ ～深層学習を用いたモダリティ変換の試み～

鈴木釈規[†1]　池田大輔[†2]

概要：国際ワークショップなどにより様々な研究課題に対するテストコレクションの整備が進み，研究課題毎の性能把握が容易になっている．解決性能が低い課題を高い課題に変換出来れば，それらの差分だけ変換前課題の性能向上が期待できる．画像アノテーションでは入力画像に対して画像を説明するキーワードを付与する．キーワードは，例えば，人が写る画像でも，女性や観光客などシーンに応じて変化する曖昧性がある．最新の画像アノテーション性能は，テキスト分類のそれと比較しても低い．こうした現状から着想を得て，画像アノテーションをテキスト分類課題に変換するモダリティ変換手法を提案する．また現在得られている実験結果を報告する．

キーワード：画像アノテーション，テキスト分類，マルチモーダル，ニューラルネットワーク，ウィキペディア

Converting Image Annotation into Text Classification: A Trial of Changing Modalities of Data with Neural Networks

TOKINORI SUZUKI[†1]　DAISUKE IKEDA[†2]

Abstract: Thanks to test collections for shared tasks arranged by international workshops, it has been easier to grasp performances of research tasks than ever. If we can replace a research task where systems have showed poor performances into another task where systems have showed high performances, we expect the performance gain from the difference of the solution levels among them. Taking two tasks: Image Annotation and Text Classification as a trial, we propose a method of modality converting from images to texts using a neural network. We report the experimental results that we got so far and discuss the research plan.

Keywords: Image Annotation, Text Classification, Multimodal, Neural Network, Wikipedia

1. はじめに

画像認識課題のコンペティション ILSVRC [22]や情報検索分野の国際ワークショップ NTCIR a)，ImageCLEF b)などの共有研究課題の策定・評価用テストコレクションの整備によって，研究者間での様々な研究課題の統一された実験設定の共有が進み，解決性能を大まかに把握することが容易になっている．

画像アノテーション [8]も一連の ImageCLEF 会議で共有課題に設定されてきた研究課題である．画像アノテーション課題では，画像のみを入力として，画像を説明するキーワードを付与する．図 1 上表に，令和元年発表の既存研究 [17]の解決性能を示す．画像アノテーション課題の解決性能は，新しい手法であっても，F 値で 50 を超えてない．

画像アノテーション課題に対する見方を変えて，仮に画像をテキストと考えると，テキストを，カテゴリに相当するキーワードに分類するテキスト分類課題 [19]とみることが出来る．図 1 下表に，平成 27 年発表のテキスト分類の一つのデータセットでの性能を示す．注目する点は，同表

Table 1: Results of different approaches on IAPR

Approaches	Feature	P	R	F1
PLSA-WORDS[21]	HC	12	17	14.1
MBRM[21]	HC	21	14	16.8
JEC[20]	HC	25	16	19.5
...				
SSL-AWF	VGG	**58**	**43**	**49.4**

Model	20News
BoW + LR	92.81
Bigram + LR	93.12
BoW + SVM	92.43
...	...
RCNN	**96.49**

Table 2: Test set results for the datasets. The top, m

図 1　テキスト分類と画像アノテーション論文の実験結果表の転写．上表は画像アノテーション [17]の実験結果．下表はテキスト分類 [15]の分類結果（数値は F 値）．

†1 九州大学　大学院システム情報科学府
Kyushu University, Graduate School of Information Science and Electrical Engineering
suzuki.tokinori.070@s.kyushu-u.ac.jp
http://tokinoris.com/

†2 九州大学　大学院システム情報科学研究院
Kyushu University, Faculty of Information Science and Electrical Engineering
daisuke@inf.kyushu-u.ac.jp

a http://research.nii.ac.jp/ntcir/index-en.html
b https://www.imageclef.org/

A tourist group is squatting at a lookout with a wooded slope and a bald mountain in the background.

A woman in a room with white walls, a bunk bed made of wood on the left, another single bed on the right and a broad window ...

図 2　画像アノテーション課題のデータ例：IAPR-TC12 テストコレクションの画像と画像説明文の組みの二例．背景色の単語は名詞であり，キーワードとされる．

の上位に記載される手法はベースラインの位置付けであるが，F 値で 90 を超えている点である．

これらの数値比較は，課題設定やデータセットが異なるため意味があるものではないが，画像アノテーションからテキスト分類に課題を置き換えることが出来るのであれば，性能の差分だけ解決性能の向上が期待できるという直観を得ることが出来る．実際に評価実験では，性能向上を確認している．

既存手法の解決性能の差に加えて，課題変換の視点には別の利点もある．図 2 は画像アノテーション課題データセットの画像と説明文の例である．画像に付与されるキーワードは，元々は自由記述の画像説明文中に用いられる単語である．図で左右の画像どちらも人物が被写体となるが，左では観光客（tourist），右では女性（woman）がキーワードとなる．キーワードはシーンに応じて変化するため，類似する画像情報（ここでは人物）を異なるキーワードに対応づける点が難しさとなる．

画像からテキストへの課題変換によっては，上記の難しさを緩和することが期待できる．図 2 の画像にキーワードを付与する課題から，画像を，キーワードを説明しているテキストに変換し，変換テキストにキーワードを付与する課題と考える．キーワードを説明するテキストを，百科事典サイト・ウィキペディア c)の該当記事とする．図 3 は両キーワードのウィキペディア記事のワードクラウドであり，両者の記事中に頻出する単語が異なり，特徴の差異が際立つ．同図のテキストに，それぞれ観光客（tourist），女性（woman）のキーワードを付与する課題に変換が出来る．

本研究では，既存研究からの解決性能の差に着目して，課題変換によって，画像アノテーション課題の性能向上を目指す．本報告では，ニューラルネットワークを用いた画像データをテキストデータに置き換えるモダリティ変換手法を提案する．また，現在得られている，モダリティ変換手法と課題置換についての評価実験の結果を報告する．

以降の本論文の構成は次の通りである．第 2 節では，画像アノテーション課題をテキスト分類課題に置き換えるた

c https://en.wikipedia.org/

図 3　観光客（tourist）と女性（woman）のウィキペディア記事のワードクラウド．記事中の出現頻度の高い単語ほど文字が大きい．

めのモダリティ変換手法を提案し，変換データを用いた分類手法を紹介する．第 3 節では，提案手法の評価実験と現在までに得られている結果を報告する．第 4 節では，本研究と既存研究の関係をまとめる．第 5 節では，本報告のまとめと今後の展望を述べる．

2. 画像からテキストへのモダリティ変換手法

本節では，提案手法である画像データからテキストデータへのモダリティ変換，及び変換データを利用した分類器の説明をする．図 4 に本手法の流れを示す．本節では，同図の左から，本手法で用いる畳み込みニューラルネットワーク（CNN）と CNN からの特徴抽出の説明から始め，モダリティ変換，変換テキストデータ分類モデルの順に説明をする．

2.1 畳み込みニューラルネットワークからの特徴抽出

初めに，提案手法では画像の特徴量を抽出するために，CNN を用いたキーワード分類器を訓練する．入力画像が，キーワード分類器に与えられた時の，ニューラルネットワ

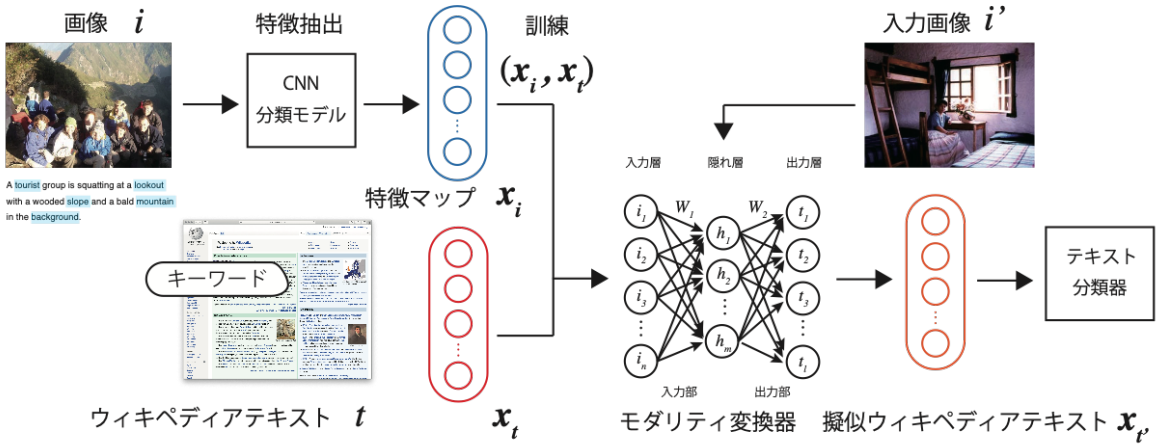

図 4　提案手法の流れ

ーク中の活性化した特定の層（特徴マップ）を画像の特徴量として利用する．

本手法では CNN アーキテクチャである *ResNet* [12]の特徴マップを用いる．ResNet は，画像認識課題コンペティション ILSVRC 2015 [22]で，分類性能で 1 位となった CNN である．ネットワークは，224×224 ピクセルの画像の入力層として，畳み込み層，活性化，バッチ正規化層から構成される複数の層をブロックとして，ブロックを複数回繰り返して，プーリング層と全結合層というネットワーク構成をとる．各ブロックは，その前の層の出力と共に，前のブロックの出力（ショートカット）の 2 つを入力にとる．前ブロックの入力を明示的に入力とすることで，ショートカット構造は，各ブロックに対応する潜在的な写像を学習する，残差関数として機能するように設計がされている．

深層学習フレームワーク Keras の 50 層からなる ResNet の実装 d)を用いた．ImageNet [6] e)の画像から収集された 1,000 クラス，100 万枚以上の画像についての事前学習済みモデルをネットワークの重みの初期値として用いた．ResNet の最後の層に，評価用データセットのキーワード数（第 3 節）のサイズの全結合層を付け加えて，分類器の学習を行った．最適化には，ミニバッチサイズ 32 で，確率的勾配降下法 [3]を用いた．初期学習率を 0.1，重み減衰値を 0.0001，モーメンタムを 0.9 とした．

上述の様に学習された ResNet ネットワークでは，入力画像 i は 224×224 ピクセルの画像に変換されて，ニューラルネットを順方向に通過して行く．入力画像 i が各層を通過していき，最終層である全結合層の特徴マップ $\boldsymbol{x}_i \in \mathbb{R}^d$ を i の特徴ベクトルとする．特徴マップには，局所的な特徴を表す中間の畳み込み層 [20, 27]の利用も考えられるが，今回は検証の容易さから，画像全体を表す特徴量として最終層の全結合層 [2, 21]を用いた．

2.2 モダリティ変換対象のテキストデータ

本手法では，画像の説明文中のキーワードについての記述を持つテキストデータに変換を目標とする．変換の対象にするテキストデータとして，匿名・協調型編集の百科事典サイトである英語版ウィキペディアの記事を用いる．ウィキペディアは，匿名・協調型という特色のため，内容の正確性など記事の品質に関する議論はあるが，科学記事のそれは，専任の編集者などが携わるブリタニカ百科事典 f)に類似するという報告もある [9]．本研究では，品質及び記事取得の容易さから同サイトの記事を用いる．

評価実験（第 3 節）で用いるデータセットの画像キーワードとウィキペディア記事の対応関係を調査では，実験で用いる全 291 キーワードに対応する見出し語があった．キーワードに対して，複数の見出し語がある（ウィキペディア中で，曖昧なページ）場合は，人手で対応関係を判断した．画像内容とキーワードの対応関係判定の自動化 [24]を検討する必要がある．291 件のウィキペディア記事で，タイトルと本文のテキストを用いて，一記事あたりの平均単語数 2,682 単語，異なり語数 37,230 語であった．

2.3 画像からテキストデータへのモダリティ変換器

次元削減などの教師なし課題に用いられるニューラルネットワーク・自己符号化器はデータ中の有用な特徴を学習する [10]．自己符号化器に着想を得て，本手法のモダリティ変換器は，それと類似する構成とすることで，画像中の特徴をテキストに対応させることを目指す．

画像特徴ベクトルをテキストデータに変換する，モダリティ変換器を図 4 中央右に示す．ここでは複雑なネットワーク構成の利用も考えられるが，課題変換の有効性の評価に重点を置き，単純な入力層，隠れ層，出力層からなる 3 層のニューラルネットワークを用いる．変換器では，画像

d https://keras.io/applications/#resnet
e http://www.image-net.org/
f https://www.britannica.com/

特徴ベクトル $\boldsymbol{x_i}$ を擬似ウィキペディアテキストベクトル $\boldsymbol{x'}_t \in \mathbb{R}^l$ への変換を学習する．

各層のニューロン数は，入力層のサイズは画像特徴の次元数 n ，隠れ層のサイズ m ，出力層のサイズ lとなる．入力層と隠れ層かなる入力部では，次の関数 f によって，$\boldsymbol{x_i}$ を低次元の潜在表現ベクトル $\boldsymbol{h} \in \mathbb{R}^m$ に変換する．

$$\boldsymbol{h} = f(\boldsymbol{i}) = ReLU(\boldsymbol{W_1} \cdot \boldsymbol{x}_i + \boldsymbol{b_1}),$$

ここで，$ReLU(\cdot)$ はランプ関数を表し，$ReLU(x) = \max(0, x)$となる．$\boldsymbol{W_1}$，$\boldsymbol{b_1}$は入力部のパラメータであり，それぞれ，重み行列 $\boldsymbol{W_1} \in \mathbb{R}^{m \times n}$，バイアス項 $\boldsymbol{b_1} \in \mathbb{R}^m$ である．

出力部では，潜在表現ベクトル $\boldsymbol{h}$ を関数 g によりテキストベクトル $\boldsymbol{x'}_t$ に写像する．

$$\boldsymbol{x'}_t = g(\boldsymbol{h}) = ReLU(\boldsymbol{W_2} \cdot \boldsymbol{h} + \boldsymbol{b_2}),$$

$\boldsymbol{W_2} \in \mathbb{R}^{n \times m}$ は重み行列，$\boldsymbol{b_2} \in \mathbb{R}^{\boldsymbol{n}}$ はバイアス項であり，出力部のパラメータとなる．モダリティ変換器は，入力 $\boldsymbol{x_i}$ と出力 $\boldsymbol{x'}_t$ の変換誤差を最小化するように学習がされる．変換誤差は，二乗誤差を変換誤差として，損失関数 $\mathcal{L}$ を次式とする．

$$\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{x}_{t\prime}) = \|\boldsymbol{x}_i - \boldsymbol{x'}_t\|^2$$

損失関数は，確率的勾配降下法 [3]を用いて最適化を行う．

2.4 分類手法

第 2.3 節で変換した擬似テキストを特徴量として，キーワード分類器を訓練する．入力画像擬 i' から変換された擬似テキストウィキペディアテキスト t' が画像像説明文中のキーワード $k \in K$ が割り当てられる確率 $P(k|t')$ としてモデル化する．一つの入力に対して，複数のキーワードが付与されるため，マルチラベル分類問題として扱う．出力層に，$Softmax$ 関数 [10]を用いる三層からなるニューラルネットワークを用いた．

$$P(k|t') = Softmax(\boldsymbol{w}_k \cdot \boldsymbol{h}_{t'} + b_k),$$

$$\boldsymbol{h}_{\boldsymbol{t'}} = \sigma(H_{x_{t\prime}} + \boldsymbol{b_n})$$

$\boldsymbol{x}_{t\prime} \in \mathbb{R}^d$ は擬似ウィキペディアテキスト t' の特徴ベクトルであり，行列$H_{x_{t\prime}} \in \mathbb{R}^{d \times h}$，$\boldsymbol{w}_k \in \mathbb{R}^d$，バイアス項 $\boldsymbol{b_n} \in \mathbb{R}^h$，$b_k \in \mathbb{R}$ はモデルパラメータである．今回は，隠れ層のサイズとして，$h = 1000$として設定をした．

次に，上述の擬似テキスト特徴量を含め，キーワード分類に用いることが出来る，特徴量を以下にまとめる．

- **画像特徴量**　第 2.1 節で紹介した ResNet [12]の最終層の全結合層の特徴マップ．（本評価実験では 1,000 次元のベクトルを設定した．）

- **テキスト特徴量**　画像のキーワードに対応するウィキペディア記事のテキストデータ（第 2.2 節）の特徴量．本研究では，この特徴量を目指す上限として位置付ける．そのため，タイトルと本文中の単語の出現頻度（ベースラインとされる [15]）と簡便なものとした．異なり語数に対応する．（評価実験で用いるウィキペディアの異なり語数，37,230 次元のベクトル．）

- **擬似テキスト特徴量**　第 2.3 節のモダリティ変換器を用いて画像特徴量を変換した擬似的なテキストデータの特徴量である．（評価実験で用いるウィキペディアの異なり語数，37,230 次元のベクトル．）

3. 評価実験

評価実験では，「画像から変換したテキストによって，画像検索課題を解く時に，既存の画像を用いる手法よりも性能向上を達成できるか」を明らかにするために，実験を設定及び評価を行った．以降，本実験の設定，得られた結果の順に述べる．

3.1 実験設定

本実験では，画像アノテーション課題評価用データセット IAPR TC-12 [11]を用いた．IAPR TC-12 は，図 2 に例示する様な人々，動物，風景，街やスポーツなど様々な場面を収めた 20,000 画像からなり，各画像は，英独西の三言語の自由記述による画像説明文（同図の画像下）が付与されている．

データセットの使用方法は，画像アノテーション課題の既存研究 [18]と同等の設定を用いた．英文の画像説明文を対象として，Stanford POS tagger [25]を用いて品詞解析を行い，名詞の単語をキーワード候補とした．白黒の画像に付与されているものや低頻度のものを，キーワードから除外した．表 1 にまとめる通り，頻度が高いキーワードから上位 291 件，それらが付与された 19,008 画像を評価に用いた．無作為抽出により 9 対 1 に分割し，17,106 件を訓練用に，1,902 件をテスト用とした．

表 1　実験設定での IAPR TC-12 データセットの統計量

キーワード数	画像数	平均キーワード数
291	19,008	5.8

提案手法に対して，次の二手法を比較対象とした．一つ目の手法は，第 2.4 節中の画像特徴量を用いて，キーワードに分類する手法（画像分類手法）である．二つ目は，画像情報は用いないで，キーワードに対応するタイトルのウィキペディア記事を特徴量として分類する手法（テキスト分類手法）である．この比較手法は，画像キーワードとウィキペディアの対応関係が与えられている理想的な設定で，上限の位置づけである．用いる特徴量は，テキスト特徴量（第 2.4 節）である．分類モデルには SVM を用いた．分類

器の実装には，SVM light [14]を用いた．

画像検索課題の評価尺度として，各画像につき手法の出力する分類スコアの高い上位 5 件のキーワードを出力とした精度，再現率が用いられる [4, 7, 18]．各尺度の計算は次式となる．

$$\text{精度} = N_C/N_A$$
$$\text{再現率} = N_C/N_H$$

ここでN_Aは一つのキーワードに対して，手法が出力する画像数で，N_Cは手法の分類が正解した画像数，N_Hはデータセット中の画像数である．F 値は精度と再現率の調和平均である．

3.2 実験結果

実験結果を表 2 にまとめる. テキスト分類手法が, 精度, 再現率，F 値の全尺度で最も高い結果を表した．F 値は，0.826 であり，この結果は, テキスト分類課題の他データセットでの結果(0.9 前後) [15]よりも若干低い．次に，提案手法の擬似テキスト分類手法，画像分類手法と続く．それぞれ F 値は，0.447，0.367 である．提案手法の擬似テキスト分類手法と画像分類手法との比較では，F 値で 0.8 向上している．正解しているキーワード総数は，画像分類手法で 3,331，提案手法で 4,836 であり，総数でも高い数値となっている．

表 2　実験結果（全キーワードの平均値）

手法	精度	再現率	F 値
画像分類手法 [12]	0.476	0.299	0.367
擬似テキスト分類手法	0.537	0.383	0.447
テキスト分類手法 (上限)	0.781	0.868	0.826

3.3 結果分析

キーワードをタイプ別に分類して，提案手法と画像分類手法の結果の比較分析を行った．キーワードのタイプとして，それが指し表す事象によって，6 つのタイプ（物や動物，場所，人，建物などの構造物，その他）に大別した．

提案手法が，上手くいっているタイプは，人を表すキーワードと建物などの構造物のキーワードであった．人を表すキーワードは，女性（woman），観光客（tourist），サイクリスト（cyclist）など，14 件中 12 件で提案手法の F 値が上回っていた．建物のキーワードは，ビル（building），支柱（column）などで，12 件中 11 件で同様に高い数値を示した．人のキーワードに関しては，提案手法では，シーンに応じてキーワードの付与が比較的に上手くいった．

上記の二タイプと比較すると，場所と物や動物の二つのタイプであった．場所のタイプでは，空港（airport），ビーチ（beach），廊下（corridor）など場所を表すキーワード 60 件中 41 件であり，物や動物のキーワードでは，全キーワード 52 件中 38 件で提案手法が，高い数値を示した．その割合は，7 割前後と上記と比較すると低い数値となった．

4. 関連研究

本研究では，画像アノテーション課題の画像データをテキストデータに変換することで，テキスト分類課題への置き換えを目指す．本節では，画像アノテーションとテキスト分類の既存研究と本研究の関係を簡単にまとめる．

画像アノテーション課題は，ImageCLEF 2006 会議 [5]で共有タスクに設定されるなど, 2000 年代に活発に研究が行われている [13, 16, 17, 18, 29]. 本節では, 比較的近年の 2010 年代以降の研究 [18, 29]を取り上げる. 近傍画像からキーワードを特定するアプローチ [18, 29]では, 疎な画像特徴量をクラスタリングによってグループ化をした特徴量 [29]や画像の色やテクスチャの特徴量を距離とする手法 [18]が提案されている．半教師有り手法では，異なる画像の特徴を用いる, CNN 分類器と画像の低レベル特徴量を用いた分類器（SVM），二つの分類器を共訓練によって学習するが提案されている [17]．これらの手法は画像の特徴量やキーワードを特徴量として利用する．一方で，本研究では画像の特徴量を，テキストに置き換えを行う点で異なる．

テキスト分類課題 [19]で提案されている手法は本研究の変換後のデータへの適用が考えられる．テキスト分類には映画レビュー文 [23]などの感情分析値分類 [26]なども含まれる．本研究の画像説明文のキーワードを分類は，ネットニュース g)や新聞記事 [1]などの文章の主題分類が, 画像のシーンに応じたキーワードを付与する点で, 本研究に近い. 文章の主題分類データセットを評価実験に用いている研究 [15, 28]を取り上げて紹介する. 再帰型ニューラルネットワーク（RNN）に続いて，最大プーリング層を加えることにより，文章の後ろだけでなく全体の特徴を考慮しニューラルネットワーク構成をした分類モデル [15]や，テキスト分類の複数のデータセットから文章と分類ラベルのスコアを学習することにより，共通する特徴を利用するマルチタスク学習に基づいた手法が提案されている [28].

5. おわりに

本報告では，画像アノテーション課題をテキスト分類課題に変換するためのモダリティ変換手法の提案及び，変換データでの分類実験・これまで得られている結果について報告をした．

実験結果では，完全に画像をウィキペディアテキストに置き換えられれば，画像アノテーション課題の解決性能を

g http://qwone.com/~jason/20Newsgroups/]

大きく向上できるという知見を得た．ただ提案手法とテキスト分類手法では，実験結果に数値的に開きがある．変換手法の改善及び，テキスト特徴量への変換精度の検証は今後の取り組みとなる．また検証データを増やした実験による，さらなる検証も今後の課題となる．

参考文献

[1] Apté, C., Damerau, F. and Weiss, S. M.. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*. 1994, vol. 12, no. 3, 233‑251.

[2] Babenko, A., Slesarev, A., Chigorin, A. and Lempitsky, V.. Neural Codes for Image Retrieval. *Proceedings of the 13th European Conference on Computer Vision (ECCV '14),* 2014, pp. 584‑599.

[3] Bertsekas, D. P. and Tsitsiklis, J. N.. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization (SIOPT)*. 2000, vol. 10, no. 3, 627‑642.

[4] Carneiro, G., Chan, A. and Moreno, P.. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2010, vol. 29, no. 3, 394–410.

[5] Clough, P., Grubinger, M. Deselaers, T. Hanbury, A. and Müller, H.. Overview of the ImageCLEF 2006 Photographic Retrieval and Object Annotation Task. *Working Notes for CLEF 2006 Workshop, 2006*, pp. 579‑594.

[6] Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.. ImageNet: A large-scale hierarchical image database. *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, 2009, pp. 248‑255.

[7] Feng, S. L., Manmatha, R. and Lavrenko, V.. Multiple Bernoulli Relevance Models for Image and Video Annotation. *Proceedings of the 2004 IEEE Computer Society on Computer Vision and Pattern Recognition (CVPR '04)*. 2004, pp. 1002–1009.

[8] Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, E., Gaizauskas, R., Villegas M. and Mikolajczyk, K.. Overview of the ImageCLEF 2016 Scalable Web Image Annotation Task. *Working Notes of CLEF 2016 – Conference and Labs of Evaluation Forum*. 2016, pp. 254–278.

[9] Giles, J.. Internet encyclopaedias go head to head. *Nature*. 2005, 438, 900–911.

[10] Goodfellow, I., Bengio, Y. and Courville, A.. *Deep Learning*. 2016, Cambridge, MA, USA.

[11] Grubinger, M., Clough, P., Müller, H. and Deselaers, T.. The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems. *Proceedings of the International Workshop OntoImage 2006 Language Resources for Content-Based Image Retrieval in conjunction with the fifth edition of the International Conference on Language Ressources and Evaluation (LREC '06)*. 2006, pp. 13–23.

[12] He., K., Zhang, X., Ren, S. and Sun, J.. Deep Residual Learning for Image Recognition. *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*. 2016, pp. 770–778.

[13] Jeon, J., Lavrenko, V. and Manmatha, R.. Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03)*. 2003, pp. 119–126.

[14] Joachims, T.. Making Large-Scale SVM Learning Practical. *Advances in kernel methods: support vector learning*. Schölkopf, B., Burges, C. and Smola, A. (Eds.), 1999, MIT Press, Cambridge, MA, USA, pp. 169–184.

[15] Lai, S., Xu, L., Liu, K. and Zhao, J.. Recurrent Convolutional Neural Networks for Text Classification. *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI-15),* 2015, pp. 2267‑2273.

[16] Lavrenko, V., Manmatha, R. and Jeon, J.. A Model for Learning the Semantics of Pictures. *Proceedings of the 17th International Conference on Neural Information Processing Systems (NIPS '04)*. 2004, pp. 553–560.

[17] Li, Z., Lin, L., Zhang, C., Ma, H. and Zhao, W.. Collaborating CNN and SVM for Automatic Image Annotation. *Proceedings of the 2019 ACM International Conference on Multimedia Retrieval (ICMR '19)*. 2019, pp. 63–67.

[18] Makadia, A., Pavlovic, V. and Kumar, S.. Baselines for image annotation. *International Journal of Computer Vision*. 2010, vol. 90, no. 1, 88–105.

[19] Manning, C. D., Raghavan, P. and Schütze, H.. Text classification and Naive Bayes. *Introduction to Information Retrieval*. 2008, Cambridge University Press, Cambridge, UK.

[20] Ng, J. Y.-H., Yang, F. and Davis, L. S.. Exploiting Local Features from Deep Networks for Image Retrieval. *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*. 2015, pp. 53–61.

[21] Razavian, A. S., Azizpour, H., Sullivan, J. and Carlsson, S.. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition, *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '14)*, 2014, pp. 512–519.

[22] Russakovsky, O., Deng, J., Su, H., Krause J., Satheesh, S., Ma., S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, Li.. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision.* 2015, vol. 115, no. 3, 211–252.

[23] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. and Potts, C.. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP '13)*, 2013, pp. 1631–1642.

[24] Suzuki, T., Ikeda, D., Galuščáková, P. and Oard, D.. Towards Automatic Cataloging of Image and Textual Collections with Wikipedia. *Proceedings of the 21st International Conference on Asia-Pacific Digital Libraries (ICADL '19)*, 2019, pp. 167–180.

[25] Toutanova, K., Klein, D., Manning, C. D. and Singer, Y.. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL '03)*, 2003, pp. 252‑259.

[26] Wang, S. and Manning, C.. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL '12)*, 2012, pp. 90‑94.

[27] Wang, Y., Chen, C., Wang, J. and Zhu, Y.. Learning Discriminative Features for Image Retrieval. *Proceedings of the 2019 ACM International Conference on Multimedia Retrieval (ICMR '19)*. 2019, pp. 96–104.

[28] Zhang, H., Xiao, L., Chen, W., Wang, Y. and Jin, Y.. Multi-Task Label Embedding for Text Classification. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18)*, 2018, pp. 4545–4553.

[29] Zhang, S., Huang, J., Huang, Y., Yu, Y., Li, H. and Metaxas, D. N.. Automatic Image Annotation Using Group Sparsity. *Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR '10)*. 2010, pp. 3312–3319.