画像へのマルチタグとそれらの重要度の自動付与

杉山　裕哉†　　青野　雅樹††

† 豊橋技術科学大学 情報・知能工学課程　〒 441–8580 愛知県豊橋市天伯町字雲雀ヶ丘 1-1
†† 豊橋技術科学大学 情報・知能工学系　〒 441–8580 愛知県豊橋市天伯町字雲雀ヶ丘 1-1
E-mail: †y-sugiyama@kde.cs.tut.ac.jp, ††aono@tut.jp

あらまし　画像には複数のオブジェクト、属性などの豊富な情報が潜在的に存在する．近年、画像に対して複数のタグを自動付与する研究が活発に行われている．従来の自動タグ付与の研究は、画像に対するマルチタグ問題として取り組まれていている．しかし，実際の画像に写るオブジェクトにはその画像を見た人にとって注目度の差がある．すなわち，付与されるタグにも重要度のばらつきがあると考えられる．本研究では、マルチタグを付与するだけではなく、それらの重要度の自動付与までを含む手法を提案する．実験では MS-COCO をデータセットとして用いて行った．ベースラインとしては、マルチタグ問題で得られる確率値をそのまま重要度とするものを採用した．提案手法では、得られたマルチタグのうち、ある閾値を超えたタグに対して、分散表現モデルを適用した結果に、最初に得られた確率値を重みとして乗じ総和の平均を計算する．実験の結果，ベースラインより良い精度を得ることができた．

キーワード　画像, タグ自動付与, 深層学習

1. はじめに

インターネットの普及に伴い, Instagram, Flickr, Facbook など写真を投稿できる SNS サービスが人気を集めている．日々膨大な写真画像がアップロードされており, これらの自動分類, 自動タグ付けの研究がさかんである (例 [6])。画像には複数のオブジェクト, 属性, それらの関係など豊富な情報が存在する. 画像に自動タグ付けづる一般的な手法の一つに, 画像に対して各タグが付与される確率を求め, 閾値を用いてタグの決定を行う手法がある. 通常, 付与されているタグは付与されているか, いないかの 2 値状態である．しかし, タグクラウド [1] や Wordcloud [2] に代表される研究にあるように, 実際の画像に写るオブジェクトの重要度には, 一般的に違いがある.

本研究では, タグをマルチラベル問題として付与するだけではなく, 各画像データに複数の注釈文（アノテーション）が付随したデータセット（例 MS COCO) を前提とし, 注釈文中の単語頻度をもとに重要度を設定し, ニューラルネットワークで訓練を行った後, 与えられた未知の画像に対して, 自動的にタグとその重要度を付与できるシステムの開発を試みた．以降, 次節で関連研究について述べ, 3 節でデータセット, 4 節でベースライン手法, 5 節で提案手法, 6 節で実験結果について報告する．最後にまとめと今後の課題について述べる.

2. 関連研究

マルチタグ予測問題に関する研究は Li ら [4] や, Zhu ら [5] などのような様々な方法で行われている．近年では Wang ら [6] や, Zhang ら [7] などのような深層学習を用いた手法が活発に行われている．また, 画像のみではなく単語や文章から特徴量を抽出し, タグ予測を行うものもある．しかし, これらの研究では最終的な出力結果はタグの有無, またはランキング上位の数個のタグを取り出すといったものである.

図 1　Wordcloud の例

一方, 深層学習の手法の一つに會下ら [8] のような分類問題と回帰問題の両方のタスクを行うマルチタスク手法が知られている．會下らの研究では料理画像から料理名 (分類) とカロリー (回帰) を予測するものである.

本研究では, 會下らの研究に感化されて, 画像内のオブジェクトを表現するタグ推定問題（分類問題）と, 各タグの重要度を予測する回帰問題を同時に解決するマルチタスク問題として取り組んだ.

3. ベースライン手法

ベースライン手法は図 2 に示すように CNN と MLP の二つの要素でマルチタグ予測を行う分類問題で、重要度は分類問題

で得られた確率値をそのまま利用するものとする．本節ではこの手法について述べる．

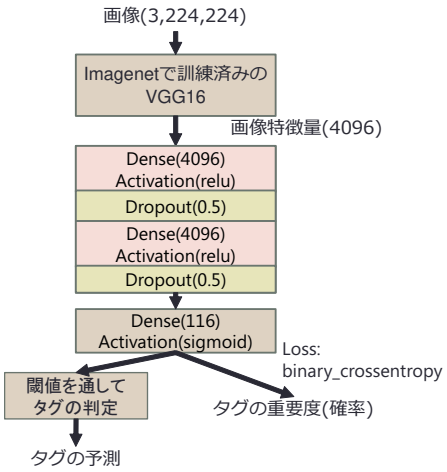


図 2　ベースライン手法

3.1　CNN

画像特徴量を抽出するために，畳込みニューラルネットワーク (CNN) を使用する．CNN には ImageNet [9] で訓練済みの VGG16 [10] を使用する．VGG16 の構造は図 3 に示すような構造で，本研究で抽出する特徴量は fc7 の出力を用いる．そのため本研究で抽出される特徴量は 4096 次元である．

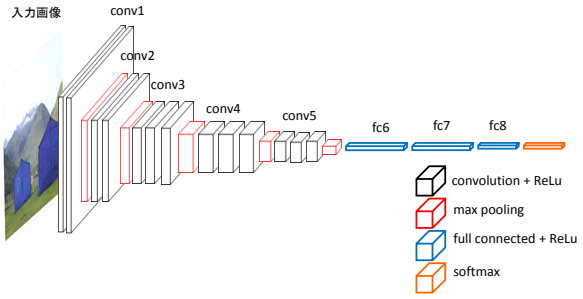


図 3　VGG 16 の概要図と出力に使う全結合層

3.2　MLP

画像特徴量から MLP を用いてマルチタグ予測を行う．入力は画像特徴量を与える．中間層の次元は 4096 次元としている．出力層の次元は，我々が事前に抽出し頻度の多かった名詞から N 個選んでいるため，N 次元である．実装では，N=116 とした．タグの選出方法については 5.1.1 で述べる．活性化関数には中間層に relu 関数，出力層に sigmoid 関数を用いる．dropout 率は 0.5 とする．

3.3　損 失 関 数

正解タグのインデックスに対応するモデルの出力値の誤差は，binary cross entropy で計算し，これを損失関数としている．この値は正解ラベルに対応する出力がどの程度高いかを表す値である．すなわち，正解ラベルに対応する出力値が高く，正解ラベル以外の出力値が小さいほど誤差が小さいということになる．目標とする値を p，実際の出力を q として，以下の式で定義される．ベースライン手法ではこの値が最小化するように訓練を行う．

$$E = -p \log q - (1-p)\log(1-q)$$

4.　提 案 手 法

以下の図 4 は本研究で提案した手法の概要である．大きく二つのモデルを作成し，タグと重要度の学習，予測を行う．比較手法として，マルチタグ予測のみを MLP で行う手法を用意し，これをベースラインとして提案手法の有効性を示す．

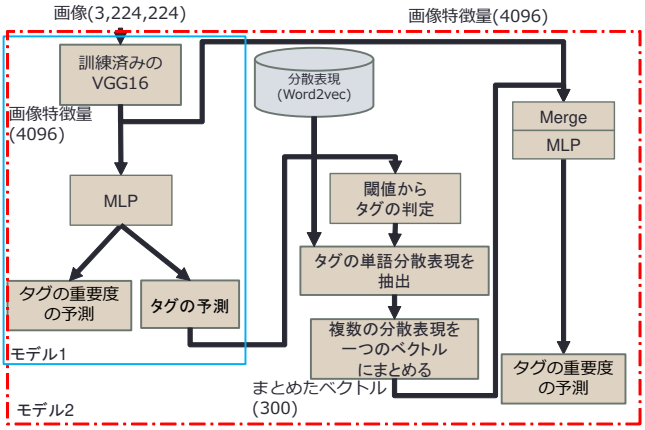


図 4　提案手法の概要

一つ目のモデルは画像特徴量を入力に MLP を用いてタグとその重要度を同時に学習するモデルである．二つ目は，一つ目のモデルに対して画像特徴量と単語の分散表現を入力として,MLP を用いてタグの重要度を学習するモデルを追加したモデルである．以降，前者をモデル 1，後者をモデル 2 とする．本節ではこれらのモデルについて述べる．

4.1　モ デ ル 1

4.1.1　モデル概要

モデル 1 を図 5 に示す．画像を入力として，ベースライン手法と同様に,ImageNet データセットで訓練済みの VGG16 を使用して画像特徴量を抽出する．抽出する特徴量は fc7 の出力を用いて 4096 次元の特徴ベクトルである．この特徴量を用いて MLP にてタグと重要度を学習する．学習時には,1 枚の画像に対して、タグのリスト (0,1 の 2 値状態でタグの有無を表すもの) と重要度のリスト (0〜1.0 までの実数値) を答えとして与え，学習する．本モデルでは MLP の出力層がタグの予測と重要度の予測のそれぞれに必要であるため，出力層が 2 つ存在する．中間層の次元数を 4096 次元，出力層の次元数は上述のように 116 次元とする.dropout 率は 0.5 とする．また，活性化関数は中間層に relu 関数，出力層ではタグの予測と重要度の予測どちらに対しても sigmoid 関数を用いる．

4.1.2　損 失 関 数

損失関数はタグの学習には binary cross entropy を，タグの重要度の学習には Mean Squared Error（MSE) を用いる．binary cross entropy はベースライン手法と同じものを用いる．MSE

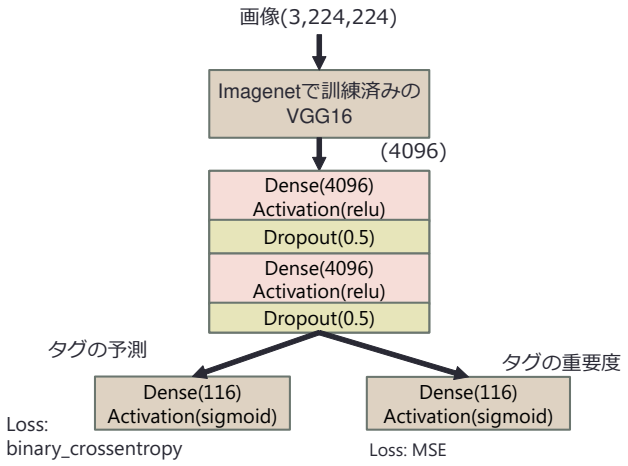


図 5　モデル 1 の概要図

に関してはデータ総数 N, 正解データの重要度 Wt, 予測値 Wp として, 以下の式で定義される．どちらの損失関数も値が最小化するように訓練を行う．

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Wt_i - Wp_i)^2$$

4.2　モデル 2

4.2.1　モデル概要

モデル 2 を図 6 に示す．このモデルは画像とタグ情報から, 画像特徴量と単語の分散表現をそれぞれ抽出し,MLP を用いてタグの重要度を学習するモデルをモデル 1 に追加し, 拡張するモデルである．画像特徴量はモデル 1 と同様, 訓練済みの VGG16 の fc7 の出力を用いて 4096 次元の特徴ベクトルを抽出する．単語の分散表現はタグ情報を入力とし, GoogleNews で訓練済みの Word2vec を用いて抽出する．このとき複数の分散表現が抽出される．これらの分散表現を, 予測されたタグ数を N, 単語分散表現を v, タグ予測時の出力 (確率) を p として, 以下に定義する式を用いて加重平均化を行い,1 つの 300 次元のベクトルにする．次に, 画像特徴量とベクトルを連結して 1 つの特徴ベクトルにする．これは keras の Merge レイヤーの 1 つ,Concatenate レイヤーを用いて連結処理を行い,4396 次元のデータを出力する．このときそれぞれの特徴量は連結する前に正規化の処理をしておく．正規化には L2 正規化の手法を用いる．学習時には正解タグデータから分散表現を抽出し, 処理を行った 300 次元のベクトルと画像特徴量を入力として, 重要度のリストを答えとして与え学習を行う．MLP は中間層の次元が 4096 次元, 出力層の次元が 116 次元となっており, 活性化関数は中間層に relu 関数, 出力層に sigmoid 関数を用いる．dropout 率は 0.5 とする．

$$WA = \frac{\sum_{i=1}^{N} vi \times pi}{N}$$

4.2.2　損失関数

損失関数はモデル 1 のタグの重要度の学習と同様に MSE を用いる．訓練時にはこの値が最小化するように訓練を行う．

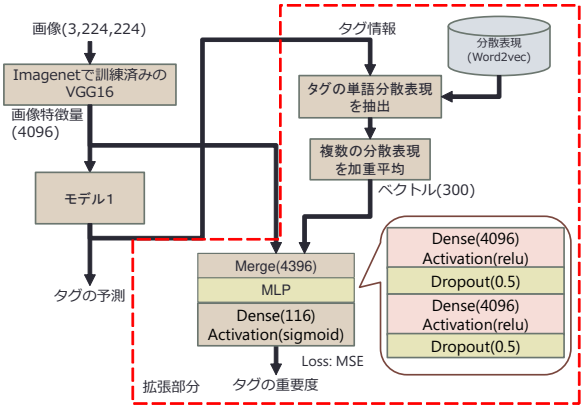


図 6　モデル 2 の概要図

5.　実　　験

5.1　データセット

本研究ではデータセットに MS-COCO [11] を用いて提案手法の訓練と評価を行った．このデータセットは図 7 に示すように, 各画像に対して Amazon Mechanical Turk によって与えられた 5 つの注釈文が付与されている．我々はこのデータセットを使ってタグに付与される重要度の設定を行った．本実験では訓練に約 32000 枚, 評価に約 8000 枚を使用した．

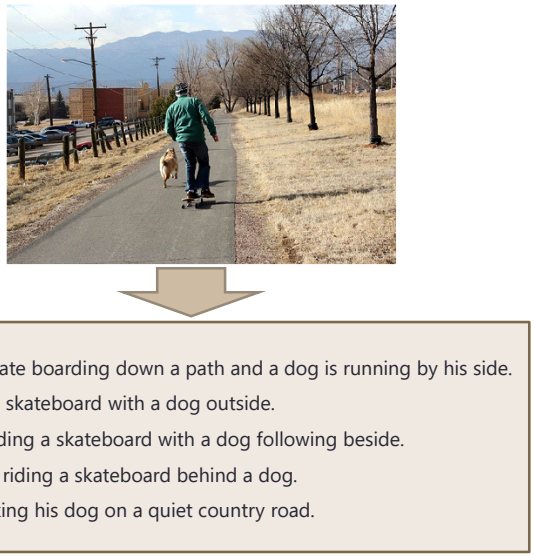


図 7　MS-COCO の画像例と 5 つの注釈例

5.1.1　タグの選出

本研究の実験に用いるタグの選出について述べる．タグの選出は後述するタグの重要度がある程度設定可能なタグデータにする必要があるため, 使用する画像の注釈文全体から一定回数出現している単語をタグの候補としてあげる．ここでは各タグのデータ数が学習に十分が値になるように出現回数の基準を 1000 に設定し, 注釈文中の名詞をタグとして選出とした．その結果 116 個のタグを選出した．

5.1.2　タグの重要度

本研究で用いるタグの重要度に対して, 次のように設定する．

上述のように 1 枚の画像に対して 5 つの注釈文が付与されている．それらの注釈文での出現回数の多い単語は, 重要度が高いと仮定し, 出現回数を重要度として設定する．回数の数え方としては,1 つの注釈文にその単語が 1 度でも出現していればカウントする．すなわち,5 つの注釈文のうち 2 つの文に出現していれば出現回数は「2」,5 つの文すべてに出現していれば,「5」となる．これらに対して, 正規化を行い, 範囲を 0 から 1.0 にしたものを各タグの重要度として設定する．重要度の設定の一例を図 8 に示す．この場合タグとして設定された「dog」が 5 つ中 5 つ,「man」が 4 つ,「skateboard」が 3 つであるため, それぞれの重要度は 1.0,0.8,0.6 となる．

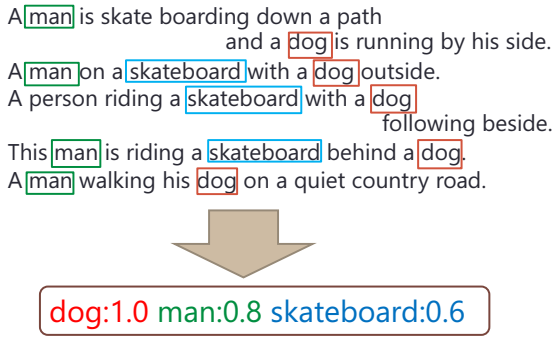


図 8　タグの重要度設定の一例

5.2　訓　　練

訓練時について述べる. エポック数は 50 として, ミニバッチ学習を行い, バッチサイズはベースライン手法も提案手法も 32 とした. 入力時の画像のサイズは 224 × 224 とした.

5.3　テ ス ト 時

テスト時の流れについて述べる. モデル 1 とモデル 2 のテスト時の流れ図を図 9,10 に示す. テスト時には未知の画像のみを入力とする. モデル 1 の場合図 9 に示すように, 未知画像から抽出された特徴量を入力として, タグの予測と重要度の予測を行う. タグの予測の出力は確率であるため, 閾値を用いて, タグの有効判定を行う. モデル 2 の場合は図 10 に示すように, まずモデル 1 に画像特徴量入力し, タグの予測を行う. そして閾値によって有効とされたタグのデータを入力として追加, 拡張したモデルを用いて単語の分散表現を抽出する. 抽出した複数の分散表現を加重平均化を用いて 1 つのベクトルにする際, 重みはモデル 1 で出力されたタグの確率を用いる. この処理によって得られたベクトルと画像特徴量を入力として MLP によって重要度の予測を行う.

5.4　評 価 方 法

5.4.1　タグ予測（マルチラベル問題）の評価指標

タグ予測の評価指標として mean Average Precision(mAP) と Jaccard 係数を用いる. それぞれについて述べる.

a）　mean Average Precision(mAP)

mean Average Precision(mAP) は検索結果における各再現率レベルでの適合率（Precision) の平均である平均適合率 (Average Precision:AP) の平均を求めたものである. 値は 0 から 1

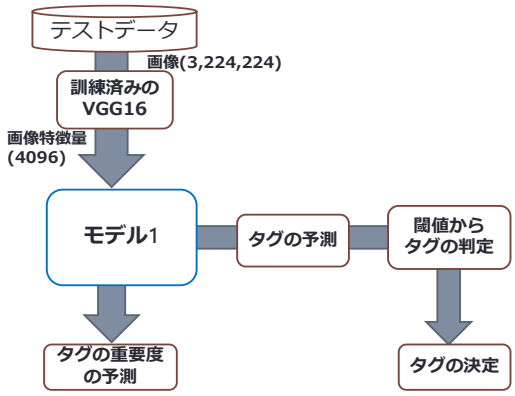


図 9　モデル 1 のテスト時の流れ

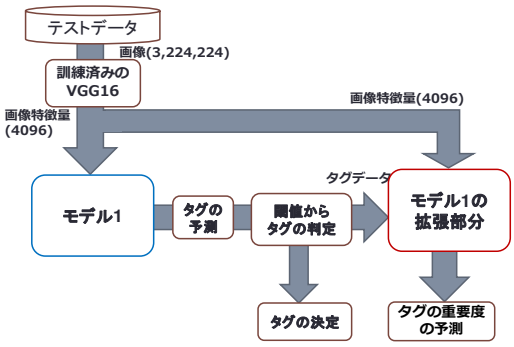


図 10　モデル 2 のテスト時の流れ

の値をとり, 値が大きいほど精度が良いということになる. 正解タグが得られた時点での AP を求め, それらの平均値を求めることで得られる. データ総数 N, 適合率 P として,AP と mAP の定義式を以下に示す.

$$\mathrm{AP} = \frac{1}{N}\sum_{i=1}^{N} P_i$$

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}_i$$

b）　Jaccard 係数

Jaccard 係数は二つの集合の類似度を表現するものである. 定義式は以下のようになる. 二つの集合の積集合を和集合で割ったものであり, 同じ要素がどの程度含まれているかを表す. 二つの集合は本実験においては正解タグデータ Y_t と予測タグデータ Y_p である. そのため, この値によってあらわされるものはタグの誤予測及び, 予測できなかったタグがどの程度含まれていないかを表す. したがって値が大きいほど, 精度が良いということになる.

$$\mathrm{Jaccard} = \frac{Y_t \cap Y_p}{Y_t \cup Y_p}$$

5.4.2　タグ重要度（回帰問題）の評価指標

タグ重要度の評価指標として MSE と Mean Absolute Error(MAE) を用いる. それぞれについて述べる.

a）　Mean Squared Error(MSE)

Mean Squared Error(MSE) は二乗平均誤差であり, 予測値

が正解からどの程度乖離しているかを示す. 今回は正解タグの重要度の誤差のみを取り出し, 評価する. この値が小さいほど, 予測された重要度は正解データに近づいているということになる. 正解タグ数を N, 重要度の正解データを Wt, 予測結果を Wp として以下の式で定義する.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(Wt_i - Wp_i)^2$$

b) Mean Absolute Error(MAE)

Mean Absolute Error(MAE) は平均絶対誤差であり,MSE と同様, 予測値がどの程度乖離しているかを示す. 今回は正解タグの重要度の誤差のみを取り出し, 評価する. この値が小さいほど, 予測された重要度は正解データに近づいているということになる. 正解タグ数を N, 重要度の正解データを Wt, 予測結果を Wp として以下の式で定義する.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|Wt_i - Wp_i|$$

5.5 結　　果

実験結果を以下の表に示す. ベースライン手法と比較して, モデル 1 によってタグ予測の精度が向上していることが確認できる. しかし, ベースライン手法と比較して, タグの重要度の精度は低下してしまっていることが確認される. また, モデル 2 のタグの重要度の予測の結果を見ると, ベースライン手法よりも精度が向上していることがわかる.

表 1　実 験 結 果

手法　評価	重要度		タグ予測	
	MAE	MSE	Jaccard	mAP
ベースライン	0.2778	0.1284	0.2478	0.5086
モデル 1	0.3073	0.1490	0.3211	0.5660
モデル 2	0.2143	0.0783	0.3221	0.5660

5.6 考　　察

成功例を図 11,12 に示す. 図 11 は主にタグ予測が成功した例である. ベースラインでは正解タグを一つも予測できていないのに対して, 提案手法ではある程度の正解タグをつけることに成功している. これは重要度を同時に学習することによって, タグ予測に影響が表れていると考えられる. 主に「baseball」に関しては重要度が高いため, 影響を受けていると考えられる. 図 12 は主に重要度の予測の成功例である. この画像はどちらの手法も正解タグを予測することに成功しているが, 重要度においてベースラインよりも, 提案手法の方が重要度の値が近づいていることがわかる. これは単語の分散表現を入力に加えたことによる情報量の増加によって精度が向上したと考えられる.

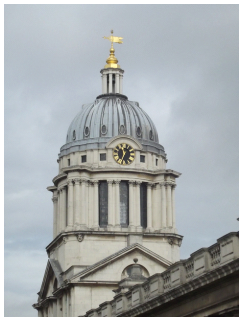
6. まとめと今後の課題

本研究では画像へのマルチタグとそれらの重要度の付与のための手法として, タグと重要度を同時に学習するモデルと画像特徴量と単語の分散表現を用いて重要度の学習を行うモデルを提案した. 実験の結果, ベースライン手法より, タグの予測, 重要度の予測どちらに対しても精度の向上が確認できた. 今後の予定としては, 今回設定した重要度のデータの精度の向上やモデルの改良による精度の向上があげられる.

正解タグ	予測タグ (ベースライン)	予測タグ (提案手法)
bat,player,ball, baseball,plate	picture,man	bat,player,baseball,

	bat	player	ball	baseball	plate
正解	0.2	0.8	0.4	1.0	0.2
ベースライン	0.139	0.231	0.129	0.225	0.05
提案手法	0.249	0.546	0.755	0.889	0.316

図 11　成 功 例 1



正解タグ	予測タグ (ベースライン)	予測タグ (提案手法)
tower,sky ,clock, building	tower,sky,clock ,building,view, background	tower,sky, clock,building, background

	tower	sky	clock	building
正解	0.4	0.2	0.6	0.6
ベースライン	0.987	0.321	0.993	0.915
提案手法	0.344	0.327	0.502	0.429

図 12　成 功 例 2

謝　　辞

本研究の一部は，科研費基盤（B）（課題番号 17H01746）の支援を受けて遂行した．

文　　献

[1] Owen Kaser and Daniel Lemire, “Tag-cloud Drawing: Algorithm for Cloud Visualization,” WWW 2007, 10 pp., 8-12 May, Banff, Canada, 2007.

[2] W. Cui, Y. Wu, S. Liu, F. We, and M. X. Zhou, “Context Preserving Dynamic Word Cloud Visualization,” IEEE Pacific Visualization Symposium 2010, pp.121-128,2-5 March, Taipei, Taiwan, 2010.

[3] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, “Cnn:Single-label to multi-label,” arXiv preprint arXiv:1406.5726, 2014.

[4] X. Li, F. Zhao, and Y. Guo, “Multi-label image classification with a probabilistic label enhancement model,” Proc. Uncertainty in Artificial Intell, 2014.

[5] F. Zhu, H. Li, W. Ouyang, N. Yu, X. Wang, Learning Spatial Regularization with Image-level Supervisions for Multi-label Image Classification

[6] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu.Cnn-rnn: A unified framework for multi-label image classification.CVPR, 2016.

[7] J. Zhang et al.Multi-label Image Classification with Regional Latent Semantic Dependencies, arXiv preprint arXiv:1612.01082, 2016

[8] Takumi Ege and Keiji Yanai: Simultaneous Estimation of Food Categories and Calories with Multi-task CNN, Proc. of IAPR International Conference on Machine Vision Applications

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009.

[10] K.Simonyan,A.Zesserman, Very deep convolutional networks for large-scale image recogni-tion,arXive:1409.1556,2014

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan,P. Doll´ ar, and C. L. Zitnick. Microsoft coco: Common objects in context. In Computer VisionECCV 2014,pages 740755. Springer, 2014.