単語画像の輪郭強調と複雑背景の合成による単語認識の高精度化

阿知破千浩† 山下隆義† 中澤満‡ 益子宗‡ 山内悠嗣† 藤吉弘亘†
†中部大学 ‡楽天技術研究所
E-mail: hf@cs.chubu.ac.jp

Abstract

文字認識において，学習画像を生成して，学習サンプルの収集コストを抑える手法 [1] が提案されている．この手法は，単純背景下に文字がある場合を想定している．一方，インターネット上の商品画像や広告等の背景は非常に複雑であり，従来の生成手法では文字認識が困難である．そこで提案手法では，文字の輪郭を強調した文字画像と背景を合成して文字画像を生成する方法を提案する．生成した文字画像を学習に用いることにより広告上に印字された文字に対しての識別精度を向上させることが可能となる．また，本生成手法を単語の生成にも応用する．生成した単語画像を学習に用いることにより，生成した単語画像に対しての識別精度を向上させることが．広告上に印字された単語に対しても識別精度を向上させることが可能となる．さらに，誤認識した斜体文字やブラー含む文字を学習に加えることで誤認識の抑制も可能となる．

1 はじめに

Deep Convolutional Neural Network(DCNN) による画像中の文字認識の研究は，手書き文字や情景中の看板などの文字認識に利用されている．文字認識には，オンライン文字認識とオフライン文字認識がある．オンライン文字認識は，タブレットや Personal Digital Assistant(PDA) から入力されるテキストをオンラインで認識する方法である．一方，オフライン文字認識は，紙に書かれた文書をスキャンし，その文字を自動的にコンピュータで処理可能なテキストデータに変換して認識する方法である．これらは背景色が均一な環境を対象としていたが，近年は複雑な背景下に文字がある情景画像を対象とした単語認識が注目されている．情景画像の文字認識は，画像中の看板やポスターなどの文字位置を特定し，その単語を認識する．この単語認識を応用することで，インターネット通販等の商品画像から商品の情報やその付加情報を収集することができる．

しかし，DCNN を用いて文字認識を行うには，大量の文字画像を必要とする．一般背景下で文字認識を行う場合，様々な広告や看板などの文字画像を大量に収集し，各画像にラベルを付与して学習サンプルを作成する必要がある．一方で，公開されている文字認識のデータセットも多数存在しているが，実用化の際に目的としたフォントで構築されたデータセットは少ない．これらの問題を解決するために，フォントデータと背景画像を用いて学習画像を生成し，学習サンプルの収集コストを削減する手法 [1] が提案されている．このような生成手法は，様々なフォントデータを用いることで必要なフォントの文字を自由に生成できるため，学習サンプルの収集コストを抑えることができる．この手法は，英単語が単純背景下にある場合を想定している．しかし，英単語や数字に比べて日本語は，漢字やひらがな，カタカナ等の種類や形状が酷似した文字が多く，認識が困難である．また，手書き文字画像や情景画像中の文字の背景は単色背景である場合が多い．一方，インターネット上の商品画像や広告等の背景は複雑であり，この生成手法では文字認識が困難である．そのため，広告等に出現する文字及び単語を認識するには，輪郭付与や背景合成などの加工が必要である．

そこで本研究では，文字及び単語の輪郭を強調させることで広告等に出現する文字及び単語の認識精度を向上させる．また，文字及び単語画像を合成する際に複雑な背景画像を合成することで，認識が難しい背景下の文字及び単語に対する認識精度の向上も期待できる．

2 従来手法

文字認識には，MNIST Dataset[4] を用いた手書き文字認識やビームサーチ手法 [5] を用いた情景画像中の文字認識 [6] 等がある．先述した文字認識において DCNN を用いて学習するには，実画像の学習サンプルを用意する必要がある．しかし，手書き文書や文字を含む情景画像等の実画像の文字認識のデータセットを作成するには，大量に収集し，各画像にラベルを付与して学習サンプルを大量に作成する必要がある．そのため，目的に応じたデータセットの構築は必要不可欠である．文献 [1] では，フォントデータと背景画像を用いて学習画像を生成している．画像を生成することにより，実画

像を大量に用意する必要がない．また，様々なフォントデータを用いることで必要なフォントの文字を自由に生成できる．文献 [1] の手法では，生成した文字画像を DCNN で認識させている．

3　提案手法

文字と単語の生成と加工は，図 1 の画像の生成，余白の追加と輪郭の強調，複雑背景の合成の 3 つのステップからなる．画像の生成と加工の手順を以下に述べる．

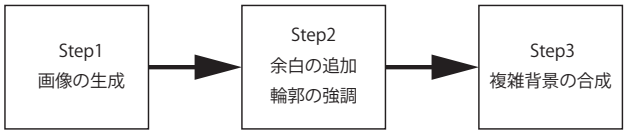

図 1　生成と加工の 3 ステップ

3.1　画像の生成

文字画像は，フォントデータと背景画像の合成で生成する．はじめに，生成する文字のリストを用意する．そして，フォントデータをもとにリストの文字を生成する．生成した文字と背景画像を合成して文字画像とする．フォントデータは一般的なフォントである MS ゴシックや MS 明朝などと合わせてインターネット通販で用いられる源柔ゴシックなど合計 22 個を用いる．図 2 に文字画像生成の流れを示す．

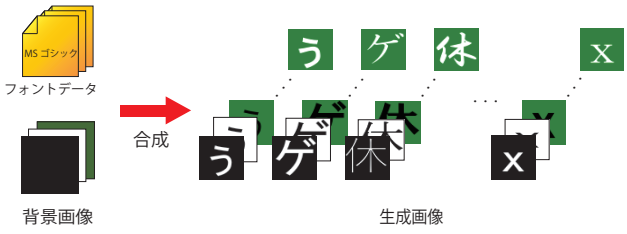

図 2　文字画像生成の流れ

3.2　余白の追加と輪郭の強調

インターネット広告の複雑背景上に印字されている文字は，縁取りなどの装飾が施されている．このような文字の認識精度の向上のため，文字の輪郭強調を画像生成に導入する．まず，生成した画像に対して文字が画像の中心になるように余白を追加する．この際，引き伸ばしてリサイズすると，文字のアスペクト比が変化してしまう．そのため，生成した文字の長辺を取得し，矩形の一辺が取得した長辺のサイズになるように余白を追加する．次に，画像中の文字を強調するために輪郭を追加する．文字色と異なる色で文字を膨張させ，元の文字と組み合わせることで文字の輪郭を追加する．輪郭の色は背景と文字色との中間色をとる．そ

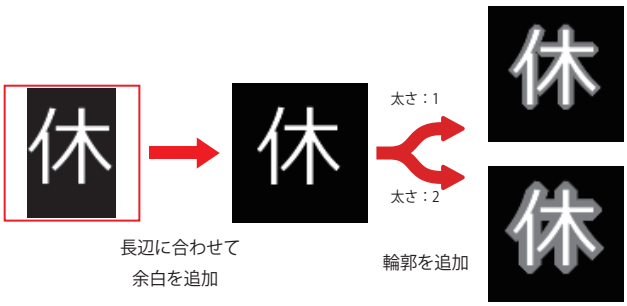

図 3　文字画像の余白の追加と輪郭強調の流れ

図 4　文字画像の背景色の置換の流れ

の際，輪郭の太さが 2 種類の画像を生成する．図 3 に文字画像の余白の追加と輪郭強調の流れを示す．

3.3　複雑背景の合成

商品画像上に印字されている文字の認識精度の向上のため，生成した文字画像の単色背景を情景画像のような複雑背景に置換し，背景を含む合成画像を生成する．ここでは図 4 に示すように，単色背景の色を緑色とし，文字及び輪郭の色は別の色とする．輪郭が中間色の場合，ノイズが発生して正しく合成できないため，文字画像の文字色と輪郭色を 2 色に統一する．合成する背景は，あらかじめ用意した広告画像の一部をランダムに切り出した画像である．

3.4　単語への応用

本研究では生成した文字画像を用いて DCNN を学習する．また，文字認識だけでなく単語認識にも有効であるか検証するために提案する生成方法を単語生成にも応用する．単語認識の対象をする単語リストをもとに，生成した文字画像を合成して単語画像を生成する．そして，文字と同様に複雑背景の合成を行う．

3.5　DCNN の構造

生成した画像を DCNN に入力して，学習する．図 5 に DCNN のネットワーク構造を示す．各層のパラメータを表 2 に示す．ネットワークは，畳み込み 3 層，全結合 1 層の全 4 層である．各層のフィルターサイズは，畳み込み 1 層目が 5×5，2 層目が 5×5，3 層目が 5×5 である．プーリング層にはマックスプーリングを用い

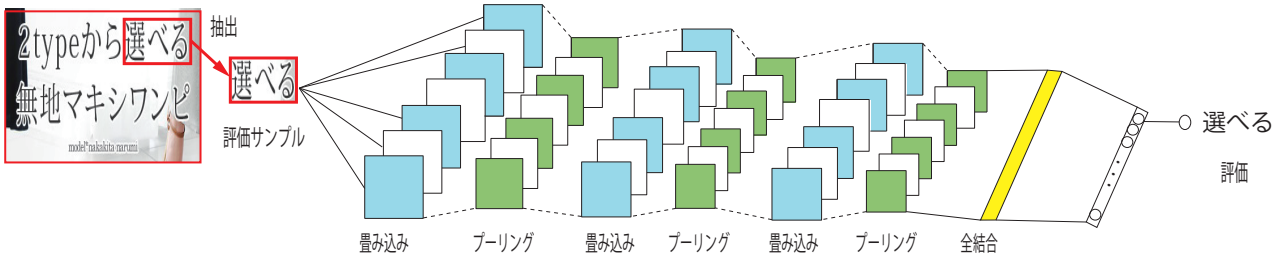

図 5　DCNN による単語認識の流れ

る．各層のプーリングサイズは 1 層目が 3×3，2 層目が 3×3，3 層目が 3×3 である．全結合のユニット数は 1 層目が 1,344，2 層目は 4096 である．活性化関数には ReLU[7] を用いる．また，全結合層に Dropout[8] を使用する．出力は文字認識の場合 1253 クラス，単語認識の場合，241 クラスである．入力サイズは文字認識の場合 32×32，単語認識の場合 96×96 である．最適化方法には AdaGrad[9] を用いる．ミニバッチサイズは 32，エポック数は 50 である．

表 1　クラスのカテゴリとクラス数

クラス名	クラス数
ひらがな	86
カタカナ	86
数字	10
英語 (大・小)	52
漢字	1006
記号	13

表 2　学習のネットワーク構成

入力	文字認識	単語認識
畳み込み	5 × 5 × 32(ReLU)	5 × 5 × 96(ReLU)
MaxPooling	3 × 3	3 × 3
畳み込み	5 × 5 × 32(ReLU)	5 × 5 × 96(ReLU)
MaxPooling	3 × 3	3 × 3
畳み込み	5 × 5 × 64(ReLU)	5 × 5 × 192(ReLU)
MaxPooling	3 × 3	3 × 3
dropout	0.5	0.5
全結合	4096(softmax)	4096(softmax)
出力	1253	241

DCNN の学習には誤差逆伝播法を用いる．誤差逆伝播法 [10] は，DCNN の出力と教師データとの誤差の勾配を出力層から入力層にかけて逆伝播させ，ネットワークの各パラメータを更新する教師付き学習アルゴリズムである．出力層と中間層の誤差勾配 $\bigtriangledown E_{kl}$ は，式 (1) のように表す．ここで，E は誤差関数，w_{kl} は DCNN のパラメータ，δ_l は出力層における教師データとの誤差，O_l は出力ユニットの出力，U_k は中間層 2 の出力である．

$$\begin{aligned} \bigtriangledown E_{kl} &= \frac{\partial E}{\partial w_{kl}} \\ &= \delta_l \cdot O_l(1-O_l) \cdot U_k \end{aligned} \tag{1}$$

また，中間層 1 と中間層 2 との間における誤差勾配 $\bigtriangledown E_{jk}$ は式 (2) のように表す．ここで，w_{jk} は中間層 1 と中間層 2 との間の結合重み，δ_k は中間層 2 に逆伝播された誤差，U_k は中間層 1 の出力である．

$$\begin{aligned} \bigtriangledown E_{jk} &= \frac{\partial E}{\partial w_{jk}} \\ &= \delta k \cdot (1-U_k) \cdot U_j \end{aligned} \tag{2}$$

また，中間層 1 と入力層との間における誤差勾配 $\bigtriangledown E_{ij}$ は式 (3) のように表す．ここで，w_{ij} は入力層と中間層との間の結合重み，δ_k は中間層 1 に逆伝播された誤差，S_i は入力層ユニットの出力である．

$$\begin{aligned} \bigtriangledown E_{ij} &= \frac{\partial E}{\partial w_{ij}} \\ &= \sum_j (\delta j) \cdot (1-U_j) \cdot S_i \end{aligned} \tag{3}$$

本研究ではネットワークの更新にミニバッチ学習法を用いる．ミニバッチ学習法は，1 回の学習に複数のサンプルを使用して各パラメータの更新量を算出する手法であり，DCNN の学習において一般的に用いられる．ミニバッチ学習法は，パラメータの更新回数を削減できる．また，1 回の更新で全てのサンプルを用いて学習を行うバッチ学習と比較して，計算量を削減できる．1 回の更新に用いるサンプル数をバッチサイズと呼ぶ．バッチサイズを M とすると，誤差関数 E は式 (4) となる．

$$E = \frac{1}{2} \sum_{m=1}^{M} \sum_{k=1}^{c} (T_k - o_k)^2 \tag{4}$$

AdaGrad では，g によって過去の勾配の二乗和を記憶しておき，その平方根で η 割ったものを式 (5) に示すように学習率として，DCNN の学習パラメータ w を更新する．

$$\begin{aligned} g^{t+1} &= g^t + \frac{\partial E}{\partial w^t}^2 \\ w^{t+1} &= w^t - \frac{\eta}{\sqrt{g^{t+1}}} \frac{\partial E}{\partial w^t} \end{aligned} \tag{5}$$

4　評価実験

本研究で提案する生成手法の有効性を確認するために 4 つの実験を行う．

図 6　認識対象の例

4.1　実験 1：生成画像の評価

実験 1 では，輪郭強調と複雑背景合成の有効性を単語の生成画像にて評価する．比較方法は同一の評価画像を使用した際の認識精度とする．評価方法に Top5 accuracy を用いる．Top5 accuracy は教師信号と同じ推定クラスの確率が上位 5 位以内であれば認識成功とする判定基準である．本実験では，Top1 accuracy から Top5 accuracy までを評価する．図 6 の認識対象は，実際にインターネット通販に用いられる単語の上位 241 クラスである．

学習用サンプルの輪郭強調なし複雑背景なしの単語画像を図 7(a)，輪郭強調あり複雑背景ありの単語画像を図 7(b) に，輪郭強調あり複雑背景なしの単語画像を図 7(c)，輪郭強調あり複雑背景ありの単語画像を図 7(d) に示す．学習用に使用する生成画像は，輪郭強調なし複雑背景なしの文字画像 1,725,264 枚で 1 文字あたり 66 枚，輪郭強調なし複雑背景合成ありの文字画像 111,806,648 枚で 1 文字あたり 3,432 枚，輪郭強調あり複雑背景なしの文字画像 5,332,859 枚で 1 文字あたり 198 枚，輪郭強調あり複雑背景合成ありの文字画像 19,288,464 枚で 1 文字あたり 6,930 枚である．評価用サンプルには，学習に用いていない輪郭強調なし複雑背景なしの文字画像 558,876 枚，輪郭強調なし複雑背景合成ありの文字画像 37,648,788 枚，輪郭強調あり複雑背景なしの文字画像 1,810,383 枚，輪郭強調あり複雑背景合成ありの文字画像 65,273,244 枚を用いる．

(a) 輪郭強調なし複雑背景なし　(b) 輪郭強調あり複雑背景なし

(c) 輪郭強調なし複雑背景あり　(d) 輪郭強調あり複雑背景あり

図 7　生成した単語画像の例

実験結果を表 3 に示す．表 3 より，生成画像の単語認識がどのデータにおいても成功していることがわかる．

表 3　生成画像の識別精度の比較 [%]

輪郭	背景	top1	top2	top3	top4	top5
なし	なし	99.4	99.8	99.8	99.8	99.9
あり	なし	96.0	97.7	98.2	98.6	98.8
なし	あり	99.1	99.6	99.8	99.9	99.9
あり	あり	99.6	99.9	99.9	99.9	99.9

4.2　実験 2：単語認識精度の評価

実験 2 では輪郭強調の有無と複雑背景の有無の有効性を実画像にて評価する．文字の生成画像と単語の生成画像の 2 つで評価する．比較方法は実験 1 と同様に Top1 accuracy から Top5 accuracy を用いる．

4.2.1　実画像の文字認識

評価用サンプルを図 8，複雑背景に用いたサンプルを図 9 を用いる．学習用に生成した輪郭強調なし複雑背景なしの文字画像を図 10(a)，輪郭強調あり複雑背景なしの文字画像を図 10(b)，輪郭強調なし複雑背景合成ありの文字画像を図 10(b)，輪郭強調あり複雑背景合成ありの文字画像を図 10(d) に示す．学習用に使用する生成画像は，輪郭強調なし複雑背景なしの文字画像 181,683 枚で 1 文字あたり 145 枚，輪郭強調あり複雑背景なしの文字画像 452,330 枚で 1 文字あたり 361 枚，輪郭強調なし複雑背景合成ありの文字画像 5,278,736 枚で 1 文字あたり 4,212 枚，輪郭強調あり複雑背景合成ありの文字画像 9,195,126 枚で 1 文字あたり 7,345 枚である．評価用に使用するサンプルは 277 枚である．文字は 1253 種類である．

図 8　評価用サンプルの例

図 9　合成に使用する背景画像の例

(a) 輪郭強調なし複雑背景なし　(b) 輪郭強調あり複雑背景なし

(c) 輪郭強調なし複雑背景あり　(d) 輪郭強調あり複雑背景あり

図 10　生成した文字画像の例

実験結果を表 4 に示す．表 4 より，輪郭を強調することで輪郭強調あり複雑背景なしの文字画像は，輪郭強調なし複雑背景なしの文字画像と比べて認識精度が Top1 accuracy で約 12%，Top5 accuracy で約 7%向上したことがわかる．また，複雑背景合成することで輪郭強調あり複雑背景なしの文字画像は，輪郭強調あり複雑背景なしの文字画像に比べて認識精度が Top1 accuracy において約 1.6%，Top5 accuracy において約 2.7%向上したことがわかる．

表 4　文字画像における輪郭有無の比較 [%]

輪郭	背景	top1	top2	top3	top4	top5
なし	なし	50.3	62.0	66.2	69.6	71.5
あり	なし	62.7	72.5	75.8	77.5	78.7
なし	あり	64.1	73.1	76.9	79.4	80.9
あり	あり	64.3	74.4	78.2	80.2	81.4

4.2.2　実画像の単語認識

評価用サンプルには実験 1 で用いた評価サンプルの 50 枚を用いる．学習用に使用する生成画像は，生成画像の評価に使用した生成画像と同様の輪郭強調なし複雑背景なしの文字画像 21,406 枚で 1 文字あたり 89 枚，輪郭強調あり複雑背景なしの文字画像 63,657 枚で 1 文字あたり 265 枚，輪郭強調なし複雑背景合成ありの文字画像 1,100,528 枚で 1 文字あたり 4,566 枚，輪郭強調あり複雑背景合成ありの文字画像 2,220,130 枚で 1 文字あたり 9,241 枚である．

実験結果を表 5 に示す．表 5 より，輪郭を強調することで輪郭強調あり複雑背景なしの文字画像は，輪郭強調なし複雑背景なしの文字画像と比べて認識精度が Top1 accuracy で約 20.4%，Top5 accuracy で約 19.8%向上したことがわかる．また，複雑背景合成することで輪郭強調あり複雑背景なしの文字画像は，輪郭強調あり複雑背景なしの文字画像に比べて認識精度が Top1 accuracy において約 14.8%，Top5 accuracy において約 14.3%向上したことがわかる．

表 5　実画像による精度の比較 [%]

輪郭	背景	top1	top2	top3	top4	top5
なし	なし	27.6	34.1	36.8	40.2	42.1
あり	なし	48.0	56.0	58.8	60.7	61.9
なし	あり	52.0	58.5	60.7	63.8	64.4
あり	あり	62.8	69.3	70.9	73.4	76.2

4.3　実験 3：単語画像の枚数の統一

前述の実験では手法ごとに学習枚数が異なっていた．そこで，実験 3 では学習枚数を統一することで学習枚数に量による認識精度への影響を除去する．これにより，学習枚数に依存しない提案手法の認識精度をする．評価比較方法は実験 1 と同様に Top1 accuracy から Top5 accuracy を用いる．

学習用サンプルの単語画像は実験 2 で用いた単語画像を，評価用サンプルは実験 2 で用いた評価サンプルを用いる．学習用に使用する生成画像は，各 2,402,417 枚である．

実験結果を表 6 に示す．表 6 より，学習枚数を統一しても認識精度が Top1 accuracy において約 14.8%，Top5 accuracy において約 14.3%認識精度が向上したことがわかる．

表 6　学習枚数統一後の比較 [%]

輪郭	背景	top1	top2	top3	top4	top5
なし	なし	27.6	34.1	36.8	40.2	42.1
あり	なし	45.2	49.8	52.3	54.2	57.3
なし	あり	55.1	59.1	61.9	63.8	65.6
あり	あり	52.6	63.2	67.5	70.0	72.4

図 11 に認識結果を示す．誤認識の傾向として斜体文字やブラー等の文字が多い．これは，斜体等の文字が学習サンプルに含まれていないことが原因だと考えられる．

4.4　実験 4：斜体とブラーへの対応

実験 2 では．斜体文字やぼやけを含む単語画像は多く誤認識していた．そこで実験 4 では，ブラーと回転を単語画像に加えて，斜体文字とブラーを含む生成画像で学習を行う．評価比較方法は実験 1 と同様に Top1 accuracy から Top5 accuracy を用いる．

学習用サンプルの単語画像は，実験 2 で用いた輪郭あり複雑背景ありの単語画像にランダムで 15 ×15 ガウスフィルターを施したもの，15 度から 180 度のランダ

(a) 認識成功例

(b) 認識失敗例

図 11　認識結果の例

ムで回転を施したもの，加工なしの 3 種類である．評価用サンプルは実験 2 で用いた評価サンプルを用いる．学習用に使用する生成画像は，実験 2 と学習枚数を同じにするため各 65,273,244 枚である．

実験結果を表 7 に示す．表 7 より，斜体文字やブラーを加えて文字を学習することで認識精度が Top1 accuracy において約 14.6%，Top5 accuracy において約 8.3%認識精度が向上したことがわかる．

表 7　加工有無の比較 [%]

加工	top1	top2	top3	top4	top5
なし	62.8	69.3	70.9	73.4	76.2
あり	77.4	81.1	83.3	83.9	84.5

5　おわりに

本稿では，単語画像の輪郭強調と複雑背景の合成による文字および単語認識の高精度化を提案した．提案手法では，生成した文字及び単語画像に輪郭の協調と複雑背景の合成を行う，加工した生成画像を使用し，DCNN によって学習を行い，広告等に出現する文字及び単語の認識精度の向上を実現した．また，誤認識した斜体文字やブラー含む文字を学習に加えることで誤認識の抑制を実現した．今後の課題は，文字と単語の両方の学習による単語認識の高精度化を検討する．

参考文献

[1] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman,“ Synthetic data and artificial neural networks for natural scene text recognition ”，arXive 2014, NIPS Deap Learning Workshop, 2014

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “ Gradient-Based Learning Applied to Document Recognition ”, Proceedings of the IEEE, pp. 2278-2324, 1998.

[3] T. Kobayashi , M. Nakagawa , “ A Pattern Classification Method of Linear-Time Learning and Constant-Time Classification ”,IEICE, 89(11):981-992, November 2006 .

[4] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, “ Gradient-Based Learning Applied to Document Recognition ”, Proceedings of the IEEE, 86(11):2278-2324, November 1998.

[5] C.-L. Liu, M. Koga, and H. Fujisawa. , “ Lexicon-driven segmentation and recognition of handwritten character strings for japanese address reading ”, IEEE Trans. Pattern Anal. Mach. Intell, 24(11),1425-1437, Nov. 2002.

[6] T. Wang, D. J. Wu, A. Coates, A. Y. Ng, “ End-to-End Text Recognition with Convolutional Neural Networks ”, ICPR, 2012.

[7] V. Nair and G. E. Hinton, “ Rectified Linear Units Improve Restricted Boltzmann Machines ” , International Conference on Machine Learning, pp.807-814, 2010.

[8] G. E. Hinton, N.Srivastava, A.Krizhevsky, I.Sutskever, and R.R.Salakhutdinov, “ Improvingneural networks by preventing co-adaptation of feature detectors ”, Clinical Orthopaedics and Related Research, vol.abs/1207.0850, 2012.

[9] Duchi, John, Elad Hazan, and Yoram Singer.,“Adaptive subgradient methods for online learning and stochastic optimization. ” Journal of Machine Learning Research 12.Jul (2011): 2121-2159.

[10] D. Rumelhart, G. E. Hinton, and R.Williams,“ Learning representations by backpropagationgerrors ”, Nature, pp.533-536, 1986.