Vol.2021-NL-249 No.8
2021/7/28

事例ベース推論を行うニューラルモデルの説明性とハブ現象の関係

佐藤 俊[1,a)]　大内 啓樹[2,3,b)]　塙 一晃[3,1,c)]　佐々木 翔大[3,1,d)]　乾 健太郎[1,3,e)]

概要：ニューラルネットワークを用いたモデル（ニューラルモデル）によって，画像処理や自然言語処理の諸タスクにおける予測性能は飛躍的に向上した．一方で，「なぜモデルがそのような予測をしたのか」を理解することは，人間にとって極めて困難であることが指摘されている．予測の「説明性」に関する問題点に対して，k 近傍法のように訓練事例との類似度にもとづいて予測を行うモデルが近年注目を集めている．この種のモデルは事例ベースモデルと呼ばれ，予測への貢献度の高い訓練事例を予測根拠として提示することが容易であるという利点を持つ．しかし，k 近傍法においては，同じ訓練事例が複数のテスト事例の近傍事例として過度に重複して出現する「ハブ」と呼ばれる現象が度々観測される．これまでの研究で，ハブ現象が事例ベースニューラルモデルの説明性に与える影響は明らかになっていない．本研究では，画像と言語データを用いた分類問題において，ニューラルモデルの枠組みで k 近傍法を使用する場面を想定し，ハブ現象が予測の説明性に悪影響を与えることを定量的に示し，かつその問題の緩和策について明らかにする．

キーワード：事例ベース推論，ハブ現象，説明性

1. はじめに

— なぜこのモデルはそのような予測をしたのか？

日頃から機械学習モデルを扱っているとこのような疑問が沸くことがある．昨今の人工知能（AI）研究の進展と加速度的な社会実装の推進に伴い，AI および機械学習モデルの予測に対する「説明」が世界的にも重要視されている．日本の総務省の出した「AI 利活用原則案」[*1]や EU の「Ethics guidelines for trustworthy AI」[*2]，米国 DARPA（Defense Advanced Research Projects Agency）の「Explainable Artificial Intelligence (XAI) Project」[*3]などに代表されるように，その注目度は世界規模で日増しに高まっている．

[1] 東北大学
Tohoku University
[2] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[3] 理化学研究所
RIKEN
[a)] shun.sato.p8@dc.tohoku.ac.jp
[b)] hiroki.ouchi@is.naist.jp
[c)] kazuaki.hanawa@riken.jp
[d)] shota.sasaki.yv@riken.jp
[e)] inui@ecei.tohoku.ac.jp

[*1] https://www.soumu.go.jp/main_content/000564147.pdf
[*2] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[*3] https://www.darpa.mil/program/explainable-artificial-intelligence

この背景には，深層学習によって大きく発展したニューラルネットワークおよびそれをベースとしたモデル（以下，ニューラルモデル）の影響がある．ニューラルモデルによって，自然言語処理や画像処理の各タスクにおける予測性能は飛躍的に向上した．一方で，「**モデルがなぜそのような予測をしたのか**」を理解することは，人間にとって極めて困難であることが指摘されている [1]．そのような状況で，ニューラルネットワークに基づく特徴抽出器と**事例ベース推論**を組み合わせたモデル（以下，事例ベースニューラルモデル）が注目を集めている [2], [3], [4], [5]．事例ベース推論とは，k 近傍法に代表されるように，訓練事例との類似度にもとづいてテスト事例に対して予測を行う推論パラダイムである．この種の推論法では予測への貢献度の高い訓練事例を提示することが容易であり，機械学習の専門知識を持たないユーザーにとってもモデルの挙動を直感的に理解可能な場合が少なくない．このように訓練事例を予測根拠として提示することは**事例ベース説明**と呼ばれる [6], [7]．理解の容易性にとどまらず，モデルの予測や挙動に対するユーザーの理解を促進し，より自信を持った意思決定につながるとの報告がある [8], [9], [10], [11], [12]．

本研究では，事例ベースニューラルモデルについて「説明性」の観点から詳細に分析する．より具体的には，**ハブ**と呼ばれる現象が説明性に及ぼす影響を明らかにする．

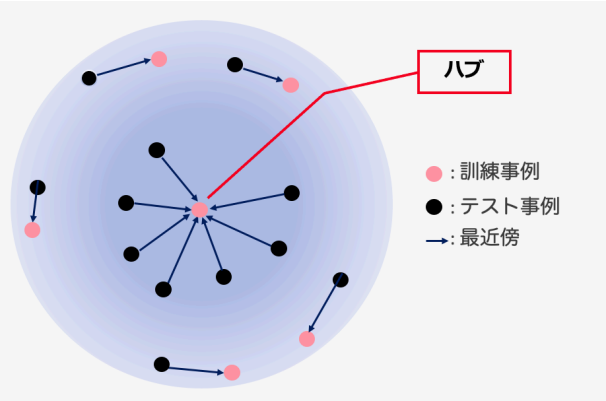

図 1 特徴ベクトル空間上でのハブのイメージ図

図 1 はハブの例を示している．ここで，中央付近に位置する訓練事例が多くのテスト事例の最近傍事例となっている．このように，複数のテスト事例の近傍事例として過度に選出されてしまう訓練事例のことをハブと言う．これまでの研究では，ハブがモデルの説明性にどのような影響を及ぼすかは明らかになっていない．しかしながら図 1 で見たように，極端に多くのテスト事例に対して同一の訓練事例を予測根拠として提示することが，説明という観点で適切なのかは甚だ疑問である．そこで本研究では，言語と画像の両データを用いた詳細な実験を行い，ハブの出現がモデルの予測の説明性（特に説明妥当性）に及ぼす影響を分析する．その結論として，ハブは説明妥当性を大きく損なう一因となることを定量的に示す．また，事例間の類似度を計算する際に，その特徴ベクトルのノルムを正規化することによってハブの発生を抑制でき，説明妥当性を保つことにつながることも示す．

本研究の貢献は以下の 3 点である．

- ハブ現象が事例ベースニューラルモデルの説明性に与える影響を定量的に調査した世界初の研究である．
- 言語と画像の両データにおいて，事例間の類似度計算として特徴ベクトル間の内積計算を行うことがハブの発生の一因となり，ハブが説明妥当性（Plausibility）を大きく損なうことを明らかにした．
- 各事例の特徴ベクトルのノルムの正規化がハブの抑制効果を持ち，説明妥当性を保つことにつながることを明らかにした．

2. 問題設定

2.1 事例ベース推論

本研究では多クラス分類問題を扱う．ニューラルモデルによって各事例を特徴ベクトル空間に写像し，その空間上で事例ベース推論の一種である k 近傍法によって推論する．k 近傍法では各テスト事例に対して，類似度の高い上位 k 個の訓練事例を選出し，それらの訓練事例に紐付くラベルの多数決によってテスト事例のラベルを予測する．

以下に具体的に記す．n 個の事例からなる訓練事例の集合を $X_{\text{train}} := \{x_1, x_2, \ldots, x_n\}$ と表す．各訓練事例 $x_i \in X_{\text{train}}$ をニューラルモデル（エンコーダ）f に入力し，特徴ベクトル $\mathbf{h}_{x_i}$ に変換する: $\mathbf{h}_{x_i} = f(x_i)$．この処理を訓練事例集合 X_{train} の全事例に対して施し，特徴ベクトル表現の集合 $\mathcal{H}_{\text{train}} := \{\mathbf{h}_{x_1}, \mathbf{h}_{x_2}, ..., \mathbf{h}_{x_n}\}$ を得る．各テスト事例 x_{test} も同様に特徴ベクトル $\mathbf{h}_{x_{\text{test}}}$ に変換する．これらの特徴ベクトル表現には，学習済みニューラルモデルの最終層のベクトル表現を用いる．この時テスト事例 x_{test} に対する予測ラベル $\hat{y}_{\text{test}}$ は $\mathbf{h}_{x_{\text{test}}}$ からみた $\mathcal{H}_{\text{train}}$ における k 個の近傍訓練事例のラベルの多数決によって予測される．

2.2 予測の説明が満たすべき性質とその評価方法

機械学習モデルの予測に対する説明はどのようなものであるべきだろうか？説明が満たすべき性質や要件は現在も多角的に検討され，いくつもの提案がなされている．その中でも Faithfulness [13], [14], [15] と Plausibility [16], [17], [18] が特に注目を集めている．

Faithfulness（忠実性）： 予測に対する説明が，モデルの実際の予測過程をどの程度反映しているかを表す概念

Plausibility（妥当性）： 予測に対する説明が，ユーザーにとってどの程度納得のいくものであるかを表す概念

2.2.1 事例ベースモデルの Faithfulness

ここで，前述したような事例ベース推論の Faithfulness について考える．この種の推論法では，訓練事例が予測ラベルの決定に直接的に貢献する．したがって，予測に顕著に貢献した訓練事例を根拠とする説明は，実際の予測過程を忠実に反映しており，ゆえに Faithful な説明であるといえる．例えば「テスト事例 x_{test} の最近傍訓練事例 $\hat{x}_i$ のラベルが y_i だったので，そのテスト事例にも同様にラベル y_i を付与した」場合を考える．この予測の根拠として最近傍訓練事例 $\hat{x}_i$ を引き合いに出し，「この訓練事例 $\hat{x}_i$ はテスト事例 x_{test} の最近傍訓練事例であり，この訓練事例に紐づくラベル y_i をテスト事例にも付与した」と説明したならば，それは実際の予測過程の忠実な説明となっている．このように，k 近傍法のような事例ベース推論において「予測」とその「説明」は表裏一体の関係にあるため，事例に基づく説明は Faithful な説明とみなせる．残すは，そのような事例に基づく説明がもう一つの観点である Plausibility を満たすかどうかが争点となる．

一方で，予測を行うモデル（メインモデル）とは別に，予測に対する説明を与えるモデル（説明のための補助モデル）を用いるアプローチも存在する [19], [20]．このアプローチでは，メインモデルの予測を入力として，補助モデルが予測に対する説明を出力する．この種のアプローチでは，メインモデルが実際に辿った予測過程を説明に反映し

ている保証はない．この特質は上述した事例ベースモデルとは対照的である．事例ベースモデルはそのモデルデザイン自体が Faithful な説明を生成する機構となっている．

2.2.2 Plausibility の評価方法

Plausibility の評価方法はオープンクエスチョンであり，未だ確立した方法はない．そのような現状の中で，Plausibility 測定法の有望なものとして **Identical Subclass Test** [7] がある．

まず前提として，Identical Subclass Test では各事例について「メインクラス」と「サブクラス」という 2 種類のクラスラベルを持つことを想定する．メインクラスは複数のサブクラスから構成され，サブクラスの上位クラスと位置付けられる．例として,「猫」「蛙」「飛行機」「自動車」の 4 つのサブクラスがある場合を考える．この時，「猫」と「蛙」の 2 つのサブクラスは「動物」というより上位のメインクラスに属し，「飛行機」「自動車」の 2 つのサブクラスは「乗り物」というメインクラスに属する，といったように定義できる．このようなメインクラスとサブクラスの存在を Identical Subclass Test では想定する．

次に，これらのクラスラベルを用いて説明妥当性を測定する．具体的には，メインクラスの情報のみ（上述した例では「動物」と「乗り物」のメインクラス）を用いてモデルを訓練する．そしてテスト時は，各テスト事例に対してその最近傍の訓練事例を選出する．この時，テスト事例とその最近傍訓練事例が同一のサブクラスに属していれば正解とする．注意点として，サブクラスの情報は教師信号として訓練時に用いない．すなわち，メインクラスの教師情報のみから，モデルが適切なサブクラス（テスト事例と同一のサブクラスに属する訓練事例）をどの程度予測できるかを問うている．

なぜサブクラスの一致率を Plausibility として評価するのかを**図 2** を用いて説明する．右図と左図ではどちらもテスト事例に対してメインクラス「動物」を正しく予測できている．しかしながら，その予測に対する根拠として提示されている最近傍訓練事例は異なる．正例（左図）では，テスト事例と同じく「猫」のサブクラスに属する訓練事例を根拠として提示している．一方負例（右図）では，テスト事例と異なる「蛙」のサブクラスに属する訓練事例を根拠として提示している．この場合ユーザーにとって，なぜ「猫」のテスト事例に対する予測根拠が「蛙」の事例になるのか疑念が生じるだろう．説明として直感的に受け入れ難く，すなわち Plausibility（妥当性・納得度）が低い説明になっている．以上のように，テスト事例と同一のサブクラスに属する事例の方が，そうでない事例よりも予測根拠として Plausibility の観点で適切であると言える．Identical Subclass Test はこの直感に基づいて，サブクラスの一致率を Plausibility を表す指標として評価に用いている．

図 2 Identical Subclass Test の正例（左）と負例（右）

2.3 ハブ現象

ハブ現象は近傍検索において複数のテスト事例の近傍事例として特定の訓練事例が過度に重複して出現する現象をさし，高次元空間において広く観測されている現象である [21], [22]. ハブが発生する原因については，不明な部分も多いが，大まかな傾向としてデータ集合の中心に近い事例がハブになりやすいことが報告されている [21], [23].

先行研究 [23], [24], [25], [26] では zero-shot 学習や分類問題など様々なタスクにおいてこのハブ現象の解消による予測性能の改善が報告されている．ハブが予測性能に与える影響については多くの研究がなされている一方で，ハブが事例ベース説明に与える影響は明らかになっていない．本研究では Identical Subclass Test を用いて，事例ベース推論におけるハブ現象が事例ベース説明に与える影響を定量的に分析する．

3. 分析の軸とする項目

本研究では，ハブが事例ベース推論の説明性に対してどのような影響を与えるかを分析する．一般にニューラルモデルを用いた場合の k 近傍法において，ハブが発生する条件はわかっていない．そこで，種々の条件（説明変数）を変化させながら，「どのような条件下でハブが発生するのか」「ハブの発生は事例ベース推論の説明性にどのような影響を及ぼすか」を調査する．

3.1 説明変数

ハブの発生や事例ベース推論は事例間の類似度計算と密接に関わる．類似度計算は主に，(1) 各事例の特徴ベクトル表現と (2) それらのベクトル間の類似度を表す指標から構成される．1 つ目の要素に関しては，ニューラルモデルを用いて各事例を特徴ベクトルに変換するため，そのモデルの訓練時に用いる損失関数によって，最終的に得られる特徴ベクトルの質が大きく変わる．2 つ目の要素に関しては，得られた特徴ベクトル間にどのような類似尺度を定義するかによって類似度の順位が代わり，選出される最近傍事例も変わる．以上のことを鑑み，本研究ではこれら 2 つの要素の取りうる値を変化させながら分析を行う．

(1) 損失関数: {Cross Entropy Loss，Triplet Loss}
(2) 類似尺度: {L2 距離，コサイン類似度，内積値 }

損失関数として，事例ベースニューラルモデルの訓練に用いられる標準的な 2 種類の損失関数を取り上げる．類似尺度としては，上記の代表的な 3 種類を取り上げる．それらの組み合わせ，つまり合計で 6 種類のモデルにおいて，ハブとその説明性への影響を分析する．

3.2 損失関数の定式化

言語処理分野でも典型的に用いられる Cross Entropy Loss と，事例間の類似度を直接的に取り込んだ損失関数である Triplet Loss [27] について説明する．

Cross Entropy Loss は以下のように定式化される．

$$L_{\text{CrossEntropy}}(x) = -\mathbf{t}\log(\text{Softmax}(\mathbf{W}\mathbf{h}_x + \mathbf{b})) \quad (1)$$

ここで，2 節で述べたように，ニューラルモデル f を用いてある訓練事例 x を h 次元の特徴ベクトル $\mathbf{h}_x \in \mathbb{R}^h$ に変換している．つまり $\mathbf{h}_x = f(x)$ である．また，$\mathbf{W} \in \mathbb{R}^{M\times d}$ は重み行列，$\mathbf{b} \in \mathbb{R}^{M\times 1}$ はバイアス項，$\mathbf{t} \in \mathbb{R}^{1\times M}$ は x の正解ラベルの箇所に 1 がたつ 1-hot ベクトル，M はクラス数を表す．

Triplet Loss を用いた訓練では，ある訓練事例 x に対して，それと同じクラス（正例）に属する訓練事例 x_+ との類似度が高くなるように，逆に異なるクラス（負例）に属する訓練事例 x_- との類似度は低くなるように訓練することが目的となる．

$$L_{\text{Triplet}}(x) = \max(0, d(\mathbf{h}_x, \mathbf{h}_{x_+}) - d(\mathbf{h}_x, \mathbf{h}_{x_-}) + m) \quad (2)$$

ここで，$\mathbf{h}_x = f(x), \mathbf{h}_{x_\text{p}} = f(x_+), \mathbf{h}_{x_\text{n}} = f(x_-)$ とし，$d(\cdot)$ は L2 距離，コサイン距離，内積値のいずれかを表す[*4]．$m \in \mathbb{R}$ はマージンを表す．

以上を踏まえて本研究では，各データセットにおいてデータ間の類似尺度（L2 距離/コサイン距離/内積値）とモデルの訓練に用いる損失関数（Cross Entropy Loss/Triplet Loss）をそれぞれ変化させた場合の「ハブの出現度合い」と「Plausibility（説明妥当性）」の挙動の分析を行う．

4. 実験

4.1 データセット

実験では文書分類タスクおよび画像分類タスクを用いる．文書分類タスクでは **20 Newsgroups** データセット [28] を用いる．このデータセットは「政治」や「科学」などの 20 種類のニュース記事から構成される 20 クラス文書分類データセットである．このデータのうち明らかにデータ内の単語数が少ないデータを除去するため，データ内の総単語数が 30 単語を超えたデータのみを用いた．訓練事例の数は 9937 個，テスト事例の数は 1105 個である．

[*4] L2 距離では距離が小さいほど類似度が高いとみなされるが，コサイン類似度と内積値では値が大きい方が類似度が高いとみなされる．そこで，コサイン類似度と内積値にはマイナス − をかけ，L2 と同じく値が小さい方が類似度が高いとみなすようにする．

画像分類のデータセットとしては，**CIFAR10** [29]，**MNIST** [30]，**Fashion MNIST** [31] の 3 つのデータセットを用いる．CIFAR10 は乗り物と動物の画像，MNIST は 0 から 9 までの数字の画像，Fashion MNIST は衣服に関する画像から構成され，いずれも 10 クラス画像分類データセットである．CIFAR10 の訓練事例の数は 45000 個，テスト事例の数は 5000 個である．MNIST および Fashion MNIST の訓練事例の数は 54000 個，テスト事例の数は 6000 個である．

今回は全ての実験において，提供されている訓練データのうち 9 割を訓練事例，残りの 1 割をテスト事例として計測を行っている．

4.2 モデル設定

文書分類モデルには畳み込みニューラルネットワーク（CNN） [32] を用いた．単語埋め込み層においては、事前学習された 200 次元の GloVe ベクトル [33] を用い，訓練中に単語ベクトルの値の更新も行った．CNN にはカーネルサイズを 3, フィルタのサイズを 3,4,5 とした Max Pooling を行った．損失の最適化には Adam [34] を用い，学習率の初期値は $\rho = 0.001$ とした．

画像分類モデルは ImageNet [35] で事前学習を行った ResNet18 [36] を用いた．事前学習済みのモデルパラメータについても訓練中に値の更新を行った．また画像分類モデルの損失の最適化には SGD を用い，学習率の初期値は $\rho = 0.01$ とした．

訓練時における損失関数には Cross Entropy Loss と Triplet Loss のいずれかを用いた．テスト時における分類予測は訓練後の各モデルの最終層の表現ベクトルを用いて k=10 の k 近傍法を用いて行った．また全ての実験はシード値の異なる 5 つのモデルを用いて行い，それらの性能の平均値を最終的な実験値とした．

4.3 評価指標

4.3.1 ハブの出現度合いの計測

先行研究 [21] に従い，ハブの出現度合いの計算は式（3）のように，各テスト事例の近傍 k 事例の中に各訓練事例が何回含まれるかという分布 N_k の歪度 S_{N_k} を用いて行う．

$$S_{N_k} = \frac{\sum_{i=1}^{l}(N_k(i) - E[N_k])^3/l}{Var[N_k]^{3/2}} \quad (3)$$

l は訓練事例の数であり，S_{N_k} は任意の実数値をとる値である．S_{N_k} がどのくらいの値の時に「ハブが発生している」という明確な基準は存在しないが，経験的な数値としてこの値が 20 を超えると明らかに人の目から見て「ハブ現象」が出現していると考えることができる．

4.3.2 説明妥当性の計測

説明妥当性の計測には 2.2 節で述べた Identical Subclass Test を用いる．先行研究 [7] に従い，既存のクラスからラ

ンダムにクラスを抽出してそれらを同一のクラスとみなすことで、2 つのメインクラスを定義する．また、元々データについていたラベルをサブクラスとして扱う．つまり、文書分類については 1 つのメインクラスが 10 個のサブクラス，画像分類については 1 つのメインクラスが 5 個のサブクラスからなる．メインクラスを対象とした 2 クラス分類の予測が正解しているテスト事例の中で，根拠として提示される近傍事例が，テスト事例と同じサブクラスである割合が Identical Subclass Test における性能値となる．

4.4 実験結果

実験結果を**表 1** に示す．Fashion MNIST の実験において Cross Entropy Loss/内積を用いた時に $S_{N_1} = 163.92$，Triplet Loss/内積を用いた時に $S_{N_1} = 164.34$ となり，いずれのデータにおいても内積を類似尺度とした場合に顕著にハブが発生していることがわかる．またそれらの条件下での Identical Subclass Test の値が L2 距離やコサイン距離を類似尺度とした場合の値と比較して低くなっていることがわかる．この結果から，類似尺度として内積を用いた場合にハブ現象が事例ベース説明における「説明妥当性」を著しく低下させる現象であることがわかった．

CIFAR10 の実験において Cross Entropy Loss/L2 を用いた時に $S_{N_1} = 64.02$，Fahion MNIST の実験において Cross Entropy Loss/L2 を用いた時に $S_{N_1} = 137.80$ となっており，L2 距離を類似尺度とした場合においてもいくつかの条件でハブが発生していることがわかる．内積の場合のハブとは異なり L2 距離のハブでは，CIFAR10 で他のハブが発生してない条件に比べて Identical Subclass Test の値が低下しているものの，平均的にはハブが発生していない場合の条件と比べて大きな差は見られなかった．

5. 分析

5.1 ハブ事例の観察

今回の実験では，類似尺度を L2 距離および内積にした場合の二つの条件でハブ現象が観測された．

5.1.1 内積におけるハブ事例

まず類似尺度を内積にした場合におけるハブについて述べる．**図 3** に CIFAR10 においてハブになった事例とそれを最近傍の事例として選択したテスト事例 10 個をランダムにサンプリングしたものを記載した．この実験では CIFAR10 の 10 個のラベルのうち「飛行機」,「船」,「トラック」,「猫」,「犬」の 5 つサブクラスを 1 つのメインクラスとしてマージしており，記載している 11 個の画像はいずれも同じメインクラスに属する．右の事例はサブクラスとして「猫」のラベルがついており，5000 個あるテスト事例のうち約半数にあたる 2526 事例がこの画像を最近傍として選択しているためハブになっていた．図 3 においてこの事例を最近傍に選んでいる 10 個のテスト事例はいずれもサブクラスが「猫」以外の事例であり，予測の根拠として提示されている右の事例とは明らかに類似していない画像となっていることがわかる．

またこのハブの事例はマージされたメインクラスの中で**最もノルムが大きい訓練事例**であった．内積におけるハブは他のデータセットにおいてもノルムが大きい傾向が見られた．

5.1.2 L2 距離におけるハブ事例

次に L2 距離を類似尺度にした場合におけるハブについて述べる．**図 4** に内積の場合と同様に CIFAR10 においてハブになっている事例とそれを最近傍の事例として選択し

表 1 各データセットにおけるマージしたクラスでの 2 値分類の予測性能 (Acc), ハブの発生度合い (S_{N_1}),Identical Subclass Test(IST)

損失関数	Cross Entropy			Triplet		
類似尺度	L2	cos	内積	L2	cos	内積
CIFAR10						
Acc	0.90	0.92	0.91	0.89	0.88	0.88
S_{N_1}	137.80	8.77	149.44	3.64	4.41	150.35
IST	0.64	0.69	0.20	0.68	0.69	0.19
MNIST						
Acc	1.00	1.00	1.00	1.00	1.00	1.00
S_{N_1}	4.18	3.38	157.48	3.18	3.16	164.30
IST	0.98	0.99	0.28	0.97	0.96	0.20
FashionMNIST						
Acc	0.96	0.96	0.96	0.96	0.96	0.96
S_{N_1}	64.02	10.07	163.92	9.01	4.04	164.34
IST	0.89	0.90	0.24	0.85	0.87	0.19
20news						
Acc	0.93	0.94	0.92	0.94	0.93	0.91
S_{N_1}	5.15	9.91	63.30	16.72	11.77	61.60
IST	0.67	0.67	0.15	0.50	0.59	0.13

図 3 CIFAR10 における内積のハブ事例とハブを最近傍（予測根拠）に選んだテスト事例

図 4 CIFAR10 における L2 距離のハブ事例とハブを最近傍（予測根拠）に選んだテスト事例

表 2 訓練事例の数を 1 割まで削減した場合の FahionMNIST におけるマージしたクラスでの 2 値分類の予測性能 (Acc), ハブの発生度合い (S_{N_1}),Identical Subclass Test(IST)

損失関数	Cross Entropy			Triplet		
類似尺度	L2	cos	内積	L2	cos	内積
Acc	0.96	0.96	0.96	0.96	0.96	0.96
S_{N_1}	27.63	3.91	53.65	16.31	4.23	54.48
IST	0.87	0.88	0.23	0.77	0.84	0.19

ているテスト事例 10 個をランダムにサンプリングしたものを記載した．この実験では CIFAR10 の 10 個のラベルのうち「飛行機」,「船」,「トラック」,「猫」,「犬」の 5 つサブクラスを 1 つのメインクラスとしてマージしており，記載している 11 個の画像はいずれも同じメインクラスに属する．右の事例は 263 個のテスト事例によって選択されておりハブになっていた．このハブの事例はサブクラスとして「船」が割り当てられた画像であった．この画像を最近傍として選択しているテスト事例を見てみると，サブクラスが「飛行機」や「猫」である画像も含まれており，このハブの画像が予測の根拠として説明妥当性にかけていることがわかる．

この条件下での全テスト事例での Identical Subclass Test の結果は 0.642 であったが，このハブを選択したテスト事例のみで Identical Subclass Test を行った場合の値は 0.13 であり，定量的にも L2 距離におけるハブ現象が事例ベース説明に悪影響を与えていることがわかる．

またこのハブの事例はマージされたメインクラスの中で**最もノルムが小さい訓練事例**であった．他のデータセットにおける L2 距離でのハブの事例にはいずれも訓練事例集合の中でノルムが小さい事例がなりやすい傾向が見られた．

5.2 訓練事例数を減らした場合の分析

k 近傍法を用いた推論においてはテスト事例の数と訓練事例の数の積の計算量が近傍検索に必要になってくるため，実応用を考えると近傍検索に用いる訓練事例の数をサンプリングして削減することは容易に考えられる．そこで我々もこうした実応用を想定し，訓練事例の数をサンプリングした場合のハブの傾向についても分析を行った．

表 2 に訓練事例の数を約 1 割まで削減した場合の結果を示す．表 1 の結果とは検索対象の訓練事例の数が異なるため，単純比較することはできないが，全ての訓練事例を用いた場合には Triplet loss と L2 距離を用いた時のハブの発生度合いが 16.31 と大きくなっていることがわかる．訓練事例の数を減らした際に Triplet Loss と L2 距離を用いた場合のハブの発生度合いが大きくなる現象は, マージする前の Fashion MNIST や CIFAR10 においても観測されている．

図 5 に訓練事例を 1 割まで削減した場合の Fashion MNIST の Triplet/L2 におけるモデルのランダムシードごとの最近傍におけるハブの出現度合い（S_{N_1}）と予測性能（Acc）を記載した．この結果をみると Triplet/L2 においてハブの出現度合いがモデルのシードによって大きく揺れていることがわかる．一方でハブの出現度合いの大小にかかわらず，モデルの予測性能はシードによってほとんど変化していないことがわかる．これは「ハブの出現」がモデルの予測性能の優劣だけでは判断しかねる事象である可能性を示唆している．すなわちモデルの予測性能が開発データ上で優れたモデルであってもハブが発生し，k 近傍法の予測に対する事例ベース説明の説明妥当性に欠けるモデルである可能性がある．このように L2 距離におけるハブ現象には興味深い現象が観測されており，今後より詳細に調査をしていく予定である．

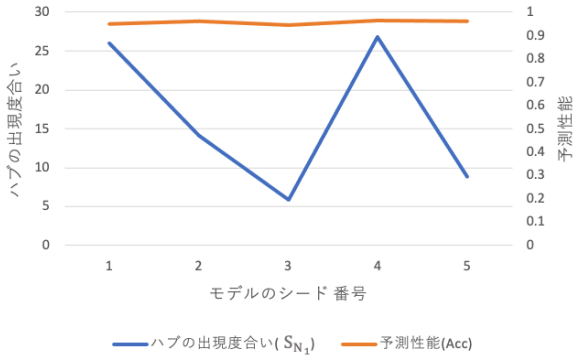

図 5 訓練事例を 1 割まで削減した場合の Fashion MNIST の Triplet/L2 におけるモデルのランダムシードごとのハブの出現度合い（S_{N_1}）と予測性能（Acc）

6. 関連研究

6.1 事例ベース推論に関する研究

「なぜそのような予測をしたのか」という機械学習モデルの予測根拠の提示は，ユーザと機械学習モデルの協業においてに極めて重要である [37]．予測根拠を提示する手法は多岐に渡るが，その中でも，テスト事例の予測に寄与した訓練事例を見せることでユーザにとって直感的な説明を提供する事例ベース説明が注目を集めている．

6.1.1 近傍事例を用いた事例ベース手法

Papernot ら [2] は画像認識のために，k 近傍法と分類器による予測を組み合わせた手法を提案している．ノイズに対する頑健性を保ちつつ，ユーザーに対して直感的な予測の根拠や予測の確信度を提示できることを示した．この手法は文書分類においても有効であることが，後続の研究によって報告されている [38].

大内ら [5] は言語の系列ラベリングタスクのための事例ベースモデルを提案している．このモデルでは，特徴ベクトル空間上であるテスト事例と各訓練事例の類似度を計算し，その類似度に基づいて予測ラベルを決定する．固有表

現抽出をはじめとする 3 つの系列ラベリングタスクの実験を通して，標準的なニューラルモデルと同等の性能を保ちながら，予測の根拠として解釈性の高い事例ベース説明を提示できることを示した．

Khandelwal ら [4] は，標準的なニューラル機械翻訳に k 近傍法を組み合わせることで翻訳性能が向上するとともに高い解釈性も実現できることを示した．

これらの手法はモデルの特徴量に k 近傍法を直接適用して予測を行っているわけではないが，予測に寄与した訓練事例を予測の根拠として提示できる点で本研究で用いた k 近傍法による事例ベース推論と共通している．

6.1.2 訓練事例の予測に対する影響度の計測

Koh ら [39] は各訓練事例がモデルの予測結果に与える影響の大きさを計測する**影響関数**という手法とともに，その値が大きい事例を予測に対する説明として提示すること提案した．

Pruthi ら [40] は訓練途中のモデルのパラメータを保存し，それらを用いて訓練事例がモデルの損失に与える影響を計測することで，先行研究と比較して計算量を抑えつつ影響度の計測ができることを示した．

Elnaz ら [41] は予測の根拠としてより直感的な説明となりうる訓練事例を特定する影響関数を改良した RelatIF という手法を提案し，定性的に優れた事例ベース説明ができることを示した．

Zhang ら [42] は言語処理分野において訓練事例の予測への影響度の計算を訓練事例について文全体ではなく，スパンの単位で行うことで影響度の計算性能が向上することを報告した．

こうした影響関数を用いた事例ベース説明の手法は盛んに研究が行われているが，予測の根拠として提示される訓練事例は，推論過程を直接反映しているわけではない．そのため，今回我々が用いた k 近傍法による事例ベース推論と比較して提示される根拠事例の Faissfulness の観点において議論の余地がある．

6.1.3 事例ベース説明の評価

事例ベース説明に関する研究の隆盛に伴い，「予測に対する説明をどのように評価すべきか」という未解決問題の重要性も高まっている．Hooker ら [43] は特定の訓練事例を取り除いた再訓練を行った前後のモデルの予測性能の比較によって訓練事例の予測に対する影響度を計測する手法を提案した．

塙ら [7] は，Plausibility の評価指標としてテスト事例と訓練事例のラベルの一致度を用いた指標を提案し，複数の説明性手法について網羅的に評価を行った．本研究においては，ハブ現象が予測の説明の妥当性に与える影響の評価指標として，塙らが提案している Identical Subclass Test を用いて計測を行った．

6.2 ハブ現象に関する研究

ハブ現象の原因の解明（6.2.1 節）や，ハブ現象の解消に関する研究（6.2.2 節）には多くの先行研究が存在している．

6.2.1 ハブの発生原因

近傍検索におけるハブ現象は，データのドメインを問わず多くのデータセット，タスクにおいて観測されている [21], [22]. ハブが出現する原因についてはいくつかの研究がある．Radovanović ら [21] は高次元空間で L2 距離を用いた場合にハブが出現する理論的背景を示している．高次元空間においてはデータ集合の中心（データ平均）により多くのデータが集中しやすい **spatial centrality** と言う現象が顕著に発生し，データ集合の中心に近い訓練事例がハブになりやすいと報告されている．この観察結果は，データの中心として原点を仮定した場合には，L2 距離においてノルムが小さいデータがハブになりやすいという我々の観察結果とも一致する．

鈴木ら [23] は類似尺度が内積の場合においても同様に，データの中心との内積が大きい事例がハブになりやすいことを示した．我々の観察では，内積におけるハブがノルムが大きい事例がなりやすくなっており，先行研究による観察結果と異なる．

6.2.2 ハブの軽減に関する研究

6.2.2.1 データや類似尺度の後処理によるハブ現象の軽減

Schnitzer ら [44] は，ハブを選択している事例の多くは，ハブから見たときの近傍事例にはなりにくいと言う近傍関係の非対称性に着目した．データ間の距離関係を，そうした近傍関係の非対称性がある場合に低い確率が割り当てられる Mutual Proximity という値に変換し，それを近傍検索に用いることでハブの軽減および k 近傍法の分類性能が向上することを示した．

鈴木ら [23] は spatial centrality の性質を利用し，データ全体を中心化することによって内積におけるハブを軽減できることを示した．

6.2.2.2 Shrinkage を利用したハブ現象の軽減

重藤ら [24] は zero-shot 学習の文脈で，L2 距離において異なる空間へデータを写像した際に，写像先の空間でのデータがよりデータの中心の近くに写像される **Shrinkage** と言う現象に注目し，通常行われる事例空間（X）からラベル空間（Y）の写像（$X \to Y$）とは逆向きの写像（$Y \to X$）を行うことで，検索対象のデータ集合 X がデータ中心付近に近づくこと防ぎ，ハブの抑制に成功している．この Shrinkage を利用したハブの抑制手法の有効性はニューラルネットを用いた zero-shot 学習 [25] や通常の分類問題 [26] においても報告されている．

6.2.2.3 訓練時の工夫によるハブ現象の軽減

Lample ら [45], Smith ら [46], Joulin ら [47] は単語ベクトルの構築時の際に，ハブを選択した場合の損失関数や予

測確率に罰則がつくように改良をすることで単語ベクトルを用いた言語タスクでの性能向上を示した．Liu ら [48] も同様に，画像と言語のマッチングタスクについて，訓練時の損失の計算時にハブの出現に対して罰則を加えることでハブの軽減に成功した．

6.2.2.4 ハブの軽減に関する研究と本研究との関連

6.2.2.1-6.2.2.3 節で言及した研究では，ハブの軽減によって k 近傍法による予測性能がどのように変化したかについて主に議論している．こうした研究の中にもハブが予測の説明に悪影響を及ぼすことを経験的，および定性的に述べた論文はいくつか存在している．しかし，k 近傍法を用いた事例ベースニューラルモデルにおいて，予測性能だけでは測れない，「説明妥当性」に関してハブ現象が与える影響を定量的に測定した研究は我々の知る限り，本研究以外には存在しない．

7. おわりに

本研究では k 近傍法を用いた事例ベースニューラルモデルについて，k 近傍法におけるハブ現象が事例ベース説明に与える影響を画像と言語のデータを用いて定量的に分析した．内積を類似尺度とした場合に実験を行った全条件においてハブの発生が観測され，Identical Subclass Test を用いた計測によって事例ベース説明の解釈性を著しく損なうことが明らかになった．

L2 距離を類似尺度とした場合にも複数の条件において，ハブの発生が観測された．L2 距離におけるハブ現象は内積の場合と比較して軽微であることが多いものの，ハブを最近傍に選んだテスト事例における Identical Subclass Test の値は平均的に低くなっており，L2 距離においてもハブ現象が事例ベース説明に悪影響を与えることがわかった．また L2 距離におけるハブ現象はモデルのランダムシードによって大きくその発生の度合いが変化する現象であった．それに加えて，ハブの発生度合いと予測性能の間には明確な相関が関係は見られず，予測性能の高い事例ベースモデルの解釈性が必ずしも高いわけではない可能性が示唆された．

これらの現象を含め,k 近傍法を用いた事例ベースモデルにおけるハブ現象にはまだ解明されていない部分が多いため，ハブ現象が事例ベース説明に与える影響を理論的側面も含めて明らかにすることを今後の研究課題としたい．

謝辞

本研究は JSPS 科研費 JP19K20351 と JST CREST JPMJCR20D2 の助成を受けたものです

参考文献

[1] Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B. and Sen, P.: A Survey of the State of Explainable AI for Natural Language Processing, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (2020).

[2] Papernot, N. and McDaniel, P. D.: Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning, *CoRR*, Vol. abs/1803.04765 (online), available from ⟨http://arxiv.org/abs/1803.04765⟩ (2018).

[3] Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L. and Lewis, M.: Generalization through Memorization: Nearest Neighbor Language Models, *Proceedings of ICLR* (2019).

[4] Khandelwal, U., Fan, A., Jurafsky, D., Zettlemoyer, L. and Lewis, M.: Nearest Neighbor Machine Translation, *International Conference on Learning Representations*, (online), available from ⟨https://openreview.net/forum?id=7wCBOfJ8hJM⟩ (2021).

[5] Ouchi, H., Suzuki, J., Kobayashi, S., Yokoi, S., Kuribayashi, T., Konno, R. and Inui, K.: Instance-Based Learning of Span Representations: A Case Study through Named Entity Recognition, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020).

[6] Caruana, R., Kangarloo, H., Dionisio, J. D., Sinha, U. and John son, D.: Case-based explanation of non-case-based learning methods., *Proceedings of the AMIA Symposium*, American Medical Informatics Association, p. 212 (1999).

[7] Hanawa, K., Yokoi, S., Hara, S. and Inui, K.: Evaluation of Similarity-based Explanations, *International Conference on Learning Representations*, (online), available from ⟨https://openreview.net/forum?id=9uvhpyQwzM⟩ (2021).

[8] Kolodneer, J. L.: Improving human decision making through case-based decision aiding, *AI magazine*, Vol. 12, No. 2, pp. 52–52 (1991).

[9] Cunningham, P., Doyle, D. and Loughrey, J.: An evaluation of the usefulness of case-based explanation, *International Conference on Case-Based Reasoning*, Springer, pp. 122–130 (2003).

[10] Ribeiro, M. T., Singh, S. and Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier, *Proceedings of KDD*, pp. 1135–1144 (2016).

[11] Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, *Proceedings of NIPS*, pp. 4765–4774 (2017).

[12] Molnar, C.: *Interpretable Machine Learning* (2019). https://christophm.github.io/interpretable-ml-book/.

[13] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. and Kim, B.: Sanity Checks for Saliency Maps, *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, p. 9525–9536 (2018).

[14] Lakkaraju, H., Kamar, E., Caruana, R. and Leskovec, J.: Faithful and Customizable Explanations of Black Box Models, New York, NY, USA, Association for Computing Machinery, p. 131–138 (online), DOI: 10.1145/3306618.3314229 (2019).

[15] Jacovi, A. and Goldberg, Y.: Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Association for Computational Linguis-

tics, pp. 4198–4205 (online), DOI: 10.18653/v1/2020.acl-main.386 (2020).

[16] Lei, T., Barzilay, R. and Jaakkola, T.: Rationalizing Neural Predictions, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117 (online), DOI: 10.18653/v1/D16-1011 (2016).

[17] Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. and Doshi-Velez, F.: An evaluation of the human-interpretability of explanation, *arXiv preprint arXiv:1902.00006* (2019).

[18] Strout, J., Zhang, Y. and Mooney, R.: Do Human Rationales Improve Machine Explanations?, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 56–62 (online), DOI: 10.18653/v1/W19-4807 (2019).

[19] Rajani, N. F., McCann, B., Xiong, C. and Socher, R.: Explain Yourself! Leveraging Language Models for Commonsense Reasoning, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4932–4942 (online), DOI: 10.18653/v1/P19-1487 (2019).

[20] Strout, J., Zhang, Y. and Mooney, R. J.: Do human rationales improve machine explanations?, *arXiv preprint arXiv:1905.13714* (2019).

[21] Radovanovi;, M., Nanopoulos, A. and Ivanovi;, M.: Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data, *Journal of Machine Learning Research* (2010).

[22] Dinu, G. and Baroni, M.: Improving zero-shot learning by mitigating the hubness problem, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings* (Bengio, Y. and LeCun, Y., eds.), (online), available from ⟨http://arxiv.org/abs/1412.6568⟩ (2015).

[23] Suzuki, I., Hara, K., Shimbo, M., Saerens, M. and Fukumizu, K.: Centering Similarity Measures to Reduce Hubs, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, Association for Computational Linguistics, pp. 613–623 (online), available from ⟨https://www.aclweb.org/anthology/D13-1058⟩ (2013).

[24] Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M. and Matsumoto, Y.: Ridge Regression, Hubness, and Zero-Shot Learning, *ECML/PKDD* (2015).

[25] Zhang, L., Xiang, T. and Gong, S.: Learning a Deep Embedding Model for Zero-Shot Learning, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

[26] Shigeto, Y., Shimbo, M. and Matsumoto, Y.: A Fast and Easy Regression Technique for k-NN Classification Without Using Negative Pairs, *Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part I* (Kim, J., Shim, K., Cao, L., Lee, J., Lin, X. and Moon, Y., eds.), Lecture Notes in Computer Science, Vol. 10234, pp. 17–29 (online), DOI: 10.1007/978-3-319-57454-7_2 (2017).

[27] Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B. and Wu, Y.: Learning Fine-grained Image Similarity with Deep Ranking, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014).

[28] Lang, K.: NewsWeeder: Learning to Filter Netnews, *Proceedings of the 12th International Machine Learning Conference* (1995).

[29] Krizhevsky, A.: Learning multiple layers of features from tiny images, Technical report (2009).

[30] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324 (online), DOI: 10.1109/5.726791 (1998).

[31] Xiao, H., Rasul, K. and Vollgraf, R.: Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms (2017). cite arxiv:1708.07747Comment: Dataset is freely available at https://github.com/zalandoresearch/fashion-mnist Benchmark is available at http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/.

[32] Yoon, K.: Convolutional Neural Networks for Sentence Classification, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (2014).

[33] Pennington, J., Socher, R. and Manning, C.: GloVe: Global Vectors for Word Representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (2014).

[34] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *Proceedings of the 3rd International Conference on Learning Representations* (2015).

[35] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L.: ImageNet: A large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255 (online), DOI: 10.1109/CVPR.2009.5206848 (2009).

[36] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (online), DOI: 10.1109/CVPR.2016.90 (2016).

[37] Doshi-Velez, F. and Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning (2017).

[38] Wallace, E., Feng, S. and Boyd-Graber, J.: Interpreting Neural Networks with Nearest Neighbors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium, Association for Computational Linguistics, pp. 136–144 (online), DOI: 10.18653/v1/W18-5416 (2018).

[39] Koh, P. W. and Liang, P.: Understanding Black-box Predictions via Influence Functions, *Proceedings of the 34th International Conference on Machine Learning* (Precup, D. and Teh, Y. W., eds.), Proceedings of Machine Learning Research, Vol. 70, PMLR, pp. 1885–1894 (online), available from ⟨http://proceedings.mlr.press/v70/koh17a.html⟩ (2017).

[40] Pruthi, G., Liu, F., Kale, S. and Sundararajan, M.: Estimating Training Data Influence by Tracing Gradient Descent, *Advances in Neural Information Processing Systems* (Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. and Lin, H., eds.), Curran Associates, Inc., pp. 19920–19930.

[41] Barshan, E., Brunet, M. and Dziugaite, G. K.: RelatIF: Identifying Explanatory Training Samples via Relative Influence, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]* (Chiappa, S. and Calandra, R., eds.), Proceedings of Machine Learning Research, Vol. 108, PMLR, pp. 1899–1909 (online), available from ⟨http://proceedings.mlr.press/v108/barshan20a.html⟩

(2020).

[42] Zhang, W., Huang, Z., Zhu, Y., Ye, G., Cui, X. and Zhang, F.: On Sample Based Explanation Methods for NLP:Efficiency, Faithfulness, and Semantic Evaluation (2021).

[43] Hooker, S., Erhan, D., Kindermans, P.-J. and Kim, B.: A Benchmark for Interpretability Methods in Deep Neural Networks, *Advances in Neural Information Processing Systems* (Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R., eds.), Curran Associates, Inc.

[44] Schnitzer, D., Flexer, A., Schedl, M. and Widmer, G.: Using Mutual Proximity to Improve Content-based Audio Similarity, *Proceedings of the 12th International Society for Music Information Retrieval Conference*, Miami (Florida), USA, pp. 79–84 (2011). http://ismir2011.ismir.net/papers/PS1-7.pdf.

[45] Lample, G., Conneau, A., Ranzato, M., Denoyer, L. and Jégou, H.: Word translation without parallel data, *International Conference on Learning Representations*, (online), available from ⟨https://openreview.net/forum?id=H196sainb⟩ (2018).

[46] Smith, S. L., Turban, D. H. P., Hamblin, S. and Hammerla, N. Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax, *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, (online), available from ⟨https://openreview.net/forum?id=r1Aab85gg⟩ (2017).

[47] Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H. and Grave, E.: Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Association for Computational Linguistics, pp. 2979–2984 (online), DOI: 10.18653/v1/D18-1330 (2018).

[48] Liu, F., Ye, R., Wang, X. and Li, S.: HAL: Improved Text-Image Matching by Mitigating Visual Semantic Hubs, *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, AAAI Press, pp. 11563–11571 (online), available from ⟨https://aaai.org/ojs/index.php/AAAI/article/view/6823⟩ (2020).