近傍事例を用いた対話における感情認識

石渡 太智 [1,2]　美野 秀弥 [1]　後藤 淳 [1]　山田 寛章 [2]　徳永 健伸 [2]
[1]NHK 放送技術研究所　[2] 東京工業大学 情報理工学院
{ishiwatari.t-fa,mino.h-gq,goto.j-fw}@nhk.or.jp　{yamada,take}@c.titech.ac.jp

概要

ソーシャルメディアでの感情分析や感情的かつ共感的な対話システムの構築を目的として対話における各発話の感情認識 (Emotion Recognition in Conversations: ERC) が注目を集めている．ERC では，発話の内容だけでなく，発話間の関係が話者の感情に大きな影響を与えることが知られている．従来手法の多くは，発話間の関係を抽出し，高い認識性能を達成した．このような手法は，単体で高い認識性能を示すことが多いが，性質の異なるモデルを組み合わせることでさらなる性能向上が期待できる．本研究は，単体で高い性能を発揮するモデルが出力する感情ラベルの確率分布と，性質の異なる別のモデルを用いて検索した近傍事例から作成した確率分布とを組み合わせる手法を提案する．評価実験において，提案手法は ERC における 3 つのベンチマークデータセットのうち，2 つのデータセットでベースモデル単体の認識率を上回る性能を達成した．また並べ替え検定において，提案手法はベースモデル単体に対して統計的に有意な結果を示した．

1　はじめに

ソーシャルメディアでの感情分析 [1] や感情的かつ共感的な対話システムの構築 [2] を目的として対話における各発話の感情認識 (Emotion Recognition in Conversations: ERC) が注目を集めている．先行研究として，再帰型ニューラルネットワークを用いて発話の内容を抽出する手法 [3] や，事前学習済み BERT モデルを用いて発話の内容を抽出する手法 [4] が提案されている．ERC では，発話の内容だけでなく発話間の関係が話者の感情に大きな影響を与えることが知られているため [5]，近年では，発話の内容だけでなく発話間の関係も考慮する手法が提案されている [6, 7]．

感情認識の問題設定では，単一のモデルでも高い認識性能を示すことが多いが，異なるモデルの出力を組み合わせることでさらなる性能向上が期待できる．そこで本研究は，異なるモデルを用いて検索した近傍事例から確率分布を作成し，ベースモデルに組み合わせる手法を提案する．近傍事例を活用した手法は，機械翻訳 [8, 9, 10, 11] や固有表現抽出 [12]，文法誤り訂正 [13] などの幅広い問題設定で活用され，有効性が示されている．しかしながら，対話における感情認識では，近傍事例を応用した手法の有効性が示されていない．

そこで本研究は，性質の異なるモデルによって検索した近傍事例から確率分布を作成し，ベースモデルの確率分布に足し合わせる手法を提案する．ベースモデルには，発話の内容だけでなく発話間の関係も考慮することで高い認識性能を達成した DAG-ERC モデル [7] を用いる．評価実験において，ERC における 3 つのベンチマークデータセットのうち，2 つのデータセットでベースモデル単体の認識率を上回る性能を達成した．さらに，1 つのデータセットで，並べ替え検定によって，ベースモデル単体よりも統計的に有意な結果を得ることを確認した．

2　提案手法

はじめに ERC の問題設定を示す．ERC では，対話における各発話 $x_1, x_2, \cdots, x_N$ の，*neutral*, *surprise*, *fear* などの感情ラベル $y_1, y_2, \cdots, y_N$ を認識する．N は 1 つの対話に現れる発話の数を示す．

提案手法は，ベースモデルによる確率分布作成 (2.1)，近傍事例の抽出 (2.2)，近傍事例による確率分布作成 (2.3)，モデルの組み合わせ (2.4) の 4 つで構成される．詳細を次で説明する．提案手法の概要を図 1 に示す．本手法は，事前学習済みのモデルを使用し，推論フェーズのみで動作する．従って，新たな学習パラメータを必要としない．

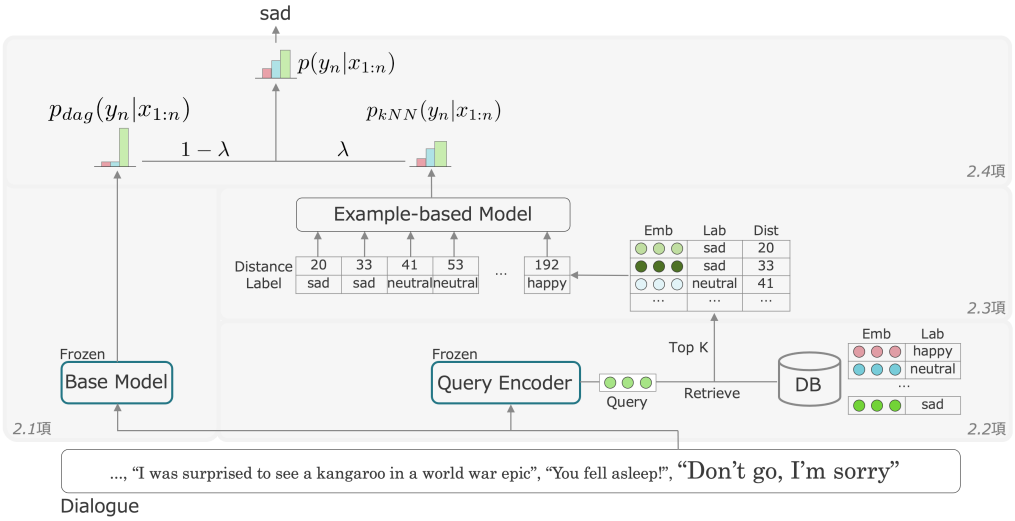

図 1　提案手法の全体図．はじめにベースモデルによる感情ラベルの確率分布 p_{dag} を作成する．次に，Query Encoder が出力した特徴量ベクトルを用いて，データベースから近傍事例を検索する．得られた上位 K 個の近傍事例に付与された感情ラベルの頻度に基づき確率分布 p_{kNN} を作成する．ベースモデルの確率分布 p_{dag} と近傍事例の確率分布 p_{kNN} を重み係数 λ を用いて足し合わせ，確率分布 p を得る．最後に，最も確率の高い感情ラベルを出力する．Frozen はモデルのパラメータを固定した状態を示す．

2.1　ベースモデルによる確率分布作成

発話の内容だけでなく発話間の関係も考慮した DAG-ERC モデル [7] を用いて，確率分布を作成する．DAG-ERC モデルは，RoBERTa-large モデル [14] を用いて対話における各発話の内容を考慮した特徴量ベクトルを抽出する．さらに，グラフニューラルネットワーク [15, 16] を用いて自身の発話からの影響と他者の発話からの影響を考慮した特徴量ベクトルを抽出する．発話の内容を示す特徴量ベクトルと発話間の影響を示す特徴量ベクトルを結合し，順伝播型ニューラルネットワークと Softmax 関数を用いて感情ラベルの確率分布を作成する．得られた確率分布を $p_{\text{dag}}(y_n|x_{1:n})$ とする．本手法では，事前学習済みの DAG-ERC モデルを利用し，全てのパラメータを固定する．なお，本論文ではベースモデルとして DAG-ERC を用いるが，ERC タスクで事前に学習した他のモデルも利用可能である．

2.2　近傍事例の抽出

ベースと異なるモデルを用いて，近傍事例を検索する方法について述べる．まず，訓練データセットの各サンプルに対して，Query encoder モデルを用いて特徴量ベクトルを抽出する．訓練データに付与された感情ラベルも取り出し，特徴量ベクトルとラベルのペアを取得する．作成した特徴量ベクトルを Key，感情ラベルを Value として，Key-Value のペアからなるデータベースを作成する．Key の特徴量を $\mathbf{h}$，Value のラベルを v とする．本稿では Query encoder として，ERC の問題設定で事前にファインチューニングを行った RoBERTa-large モデル[1)] を用いる．

続いて，評価対象の対話を入力し，先ほどの Query encoder モデルを用いて，特徴量ベクトルを取得する．Query encoder モデルを用いて特徴量ベクトルを導く関数を $f(x_{1:n})$ とする．取得した特徴量ベクトルを Query として，データベースに登録された特徴量ベクトルとの類似度を計算し，類似度の高い上位 K 個の事例を，近傍事例として取得する．抽出した近傍事例は，特徴量ベクトル (Emb)，感情ラベル (Lab)，Query との距離 (Dist) によって構成され，その集合を $R^n = \{(\mathbf{h}_i, v_i, d(\mathbf{h}_i, f(x_{1:n}))), i \in \{1, 2, \cdots, K\}\}$ とする．$d(\cdot,\cdot)$ はユークリッド距離を用いる．

2.3　近傍事例による確率分布作成

検索した近傍事例を用いて確率分布を作成する．2.2 項で抽出した K 個の近傍事例の，感情ラベル v_i と Query との距離 $d(\mathbf{h}_i, f(x_{1:n}))$ を用いる．Khandelwal らの手法 [8] を参考に，ラベルと距離を用いて確率分布を算出する式を下式で示す．

$$p_{\text{kNN}}(y_n|x_{1:n}) \propto \sum_{(\mathbf{h}_i, v_i)} \mathbb{1}_{y_n=v_i} \exp\left(\frac{-d(\mathbf{h}_i, f(x_{1:n}))}{T}\right) \quad (1)$$

1)　https://huggingface.co/roberta-large

$\mathbb{1}_{y_n=v_i}$ は，近傍 v_i が感情ラベル y_n と同一である場合に 1 を返す指示関数である．ハイパーパラメータ T は温度を示す．

2.4 モデルの組み合わせ

2.1 項で作成したベースモデルによる確率分布 p_{dag} と，2.3 項で作成した近傍事例による確率分布 p_{kNN} を組み合わせる．係数 λ を用いて重み平均を取り，ベースモデルの出力と近傍事例を組み合わせた確率分布を作成する．式を下式に示す．

$$p(y_n|x_{1:n}) = \lambda p_{\text{kNN}}(y_n|x_{1:n}) + (1-\lambda)p_{\text{dag}}(y_n|x_{1:n}) \quad (2)$$

重み係数 λ は，ハイパーパラメータである．最後に，作成した確率分布の中で，最も確率の高い感情ラベルを，推論結果として出力する．

3 実験

3.1 データセット

ERC における 3 つのベンチマークセットを用いて，提案手法の有効性を検証する．訓練データ，検証データ，テストデータの割合と評価方法を表 1 に示す．また各セットにおける対話数と発話数，クラス数を示す．

MELD [17] は，複数の俳優が登場する *Friends* という TV ドラマの，一部シーンを切り取った映像と音声の書き起こしからなるデータセットである．また，1 つの対話に複数の話者が登場し，各発話には (*neutral*, *happiness*, *surprise*, *sadness*, *anger*, *disgust*, or *fear*) のうち 1 つが付与される．

IEMOCAP [18] は 2 人の話者が，1 対 1 の会話を行う様子を収録した映像と音声の書き起こしからなるデータセットである．各発話には，(*happy*, *sad*, *neutral*, *angry*, *excited*, or *frustrated*) のうち 1 つが付与される．

EmoryNLP [19] は TV ドラマ *Friends* から，一部のシーンを切り取り収集したデータセットである．MELD と比較してデータサイズとラベルの種類が異なり，各発話には (*neural*, *sad*, *mad*, *scared*, *powerful*, *peaceful*, or *joyful*) のうち 1 つが付与される．

3.2 評価指標

Shen らの手法 [7] で用いられた評価指標と同じ，Weighted-F1 値を全てのデータセットの評価に用いる．また，ノンパラメトリック検定の一つである並べ替え検定を用いて，有意差を検定する．

3.3 その他の実験設定

その他の実験設定を示す．DAG-ERC は [7] で報告された学習パラメータを用いて，事前に学習した．また，Query encoder として用いる Finetuned-RoBERTa は，学習率を $5e$-6 に設定し，損失関数に Cross Entropy Loss を，最適化に RAdam optimizer [20] を用いて事前に学習した．

また，ベースモデルと Query encoder の特徴量ベクトルの次元数を 1024 とし，近傍事例の数 K を 32 に，温度 T は 1000 に設定した．ハイパーパラメータである係数 λ は $(0, 0.25, 0.5, 0.75, 1)$ の中から，検証データで最も Weighted-F1 値が高くなるものを選択した．近傍事例の検索は，faiss[21] を用いた．全ての実験結果は 5 回行い平均値を用いた．$512GB$ メモリの AMD EPYC 7F52 CPU と，$40GB$ メモリの NVIDIA A100 の GPU を用いて実験を行った．

4 実験結果

4.1 比較実験

ベースモデル単体との比較結果を表 2 に示す．ベースモデル単体の手法として，DAG-ERC[7] と Finetuned-RoBERTa を用いた．提案手法の Query encoder として，DAG-ERC, Vanilla-RoBERTa, Finetuned-RoBERTa の 3 種類を比較した．表 2 の avg，std，p-score は，それぞれ平均値，標準偏差，DAG-ERC を基準にした p 値を示す．

表 2 の結果より，3 つのベンチマークデータセットの内，2 つのデータセットでベースモデル単体を上回る認識性能を達成した．また，MELD データセットにおいて，Finetuned-RoBERTa を Query encoder として用いた手法 (#4) は，ベースモデルである DAG-ERC(#0) に対して p-score が 5% を下回った．以上より，提案手法が DAG-ERC に対して統計的に有意な差を示すことが確認できた．

次に，Query encoder の 3 種類を比較する．MELD データセットと IEMOCAP データセットにおいて，Finetuned-RoBERTa を用いた結果 (#4) が DAG-ERC を用いた結果 (#2) を上回った．以上より，ベースモデルと性質の異なるモデルを組み合わせる方法の有効性が確認できる．

一方で，EmoryNLP データセットでは，提案手法 (#2, 3, 4) の Weighted-F1 値が，DAG-ERC 単体を用い

データセット	対話数			発話数			クラス数	評価方法
	訓練データ	検証データ	テストデータ	訓練データ	検証データ	テストデータ		
MELD	1038	114	280	9989	1109	2610	7	Weighted-F1
IEMOCAP	108	12	31	5320	490	1630	6	Weighted-F1
EmoryNLP	713	99	85	9934	1344	1328	7	Weighted-F1

表 1 MELD，IEMOCAP，EmoryNLP ベンチマークデータセットの割合と評価方法．訓練データ，検証データ，テストデータにおける対話数と発話数，クラス数を示す．

#	Models	Query Encoder	MELD			IEMOCAP			EmoryNLP		
			avg	std	p-score	avg	std	p-score	avg	std	p-score
0	DAG-ERC	-	63.17	0.08898	-	66.57	0.5727	-	**38.12**	0.1609	-
1	Finetuned-RoBERTa	-	62.58	0.4870	3.175	55.37	0.8856	0.00	36.79	0.6305	0.00
2	Ours	DAG-ERC	63.16	0.08302	86.51	66.60	0.5372	89.68	**38.12**	0.1609	95.24
3		Vanilla-RoBERTa	63.17	0.08898	90.48	66.65	0.4787	80.16	38.09	0.1855	69.05
4		Finetuned-RoBERTa	**63.69**	0.2990	0.7937	**66.72**	0.4506	65.87	37.82	0.5449	38.10

表 2 MELD, IEMOCAP, EmoryNLP ベンチマークデータセットにおけるベースモデル単体との比較実験．avg, std, p-score は，それぞれ平均値，標準偏差，DAG-ERC を基準にした p 値を示す．ボールド体は最も性能が高い値を示す．各値は 5 回の実験による Weighted-F1 値の平均値を示す．

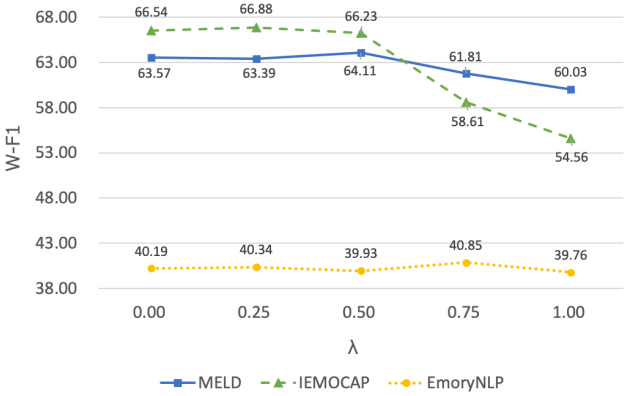

図 2 各ベンチマークデータセットにおける係数と Weighted-F1 値の関係．W-F1 は Weighted-F1 値を示す．

た結果 (#0) に比べて，低い値を示した．これは係数 λ として，0 や 0.75 が選ばれたことが原因と考えられる．係数 $\lambda = 0$ の場合，近傍事例による確率分布は利用されず，ベースモデル単体の出力がそのまま利用される．そのため，性能に変化が生じなかった．また，λ が大きい場合，近傍事例による確率分布に重きが置かれる．関連する事例がデータベースに存在しない場合も，近傍事例による確率分布に重きが置かれてしまう．そのため，ロバスト性が失われ，性能が落ちてしまった．今後は，サンプル全体で一つの係数 λ を設定するのではなく，サンプルごとに動的に係数 λ を変更することを検討する．

4.2 係数の分析

続いて，ハイパーパラメータである係数 λ と認識性能の関係を検証する．各ベンチマークデータセットの検証データにおいて，係数 λ を $(0, 0.25, 0.5, 0.75, 1)$ と変化させた時に，Query encoder として Finetuned-RoBERTa を用いた提案手法の認識率がどのように変化するかを分析する．結果を図 2 に示す．

結果から，全てのデータセットでベースモデル単体 ($\lambda = 0$) が示す認識率を，上回る λ が存在することを確認できる．特に λ が 0.25 や 0.5 のときに，高い性能を発揮することがわかった．一方で，λ が大きい場合は，認識性能が極端に劣化することがわかった．

5 おわりに

本稿は，対話における各発話の感情認識において，単体で高い性能を発揮するモデルをベースに，性質の異なるモデルを用いて検索した近傍事例から，感情ラベルの確率分布を作成し，ベースモデルに組み合わせる手法を提案した．3 つのベンチマークデータセットを用いて手法の有効性を確認したところ，2 つのデータセットでベースモデルの認識率を上回る性能を達成した．

今後の展望として，温度 T および係数 λ を学習することを検討している．本稿では温度 T は固定し，係数 λ はハイパーパラメータとして選択した．最適な温度と係数はデータセットのサンプルによって異なると考えられる．今後は，学習によって動的に変更する手法を検討する．また，近傍事例の数 K も同様に変更できる手法を検討する．

謝辞

貴重なコメントや議論を頂いた NHK 放送技術研究所の山田一郎シニアリード，宮崎太郎研究員，安田有希研究員，奥田あずみ研究員に感謝の意を表す．

参考文献

[1] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In **Proceedings of the 13th international workshop on semantic evaluation**, pp. 39–48, 2019.

[2] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Mime: Mimicking emotions for empathetic response generation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, p. 8968–8979, 2020.

[3] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. In **Proceedings of the Twenty-Fifth International Joint Conferences on Artificial Intelligence Organization (IJCAI-16)**, 2016.

[4] Linkai Luo and Yue Wang. Emotionx-hsu: Adopting pre-trained bert for emotion classification, 2019. https://arxiv.org/abs/1907.09669.

[5] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. **IEEE Access**, Vol. 7, No. 1, pp. 100943–100953, 2019.

[6] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, p. 154–164, 2019.

[7] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. Directed acyclic graph network for conversational emotion recognition. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, p. 1551–1560, 2021.

[8] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. In **International Conference on Learning Representations**, 2021.

[9] Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. Adaptive nearest neighbor machine translation. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, p. 368–374, 2021.

[10] Qingnan Jiang, Mingxuan Wang, Jun Cao, Shanbo Cheng, Shujian Huang, and Lei Li. Learning kernel-smoothed machine translation with retrieved examples. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, p. 7280–7290, 2021.

[11] Dexin Wang, Kai Fan, Boxing Chen, and Deyi Xiong. Efficient cluster-based k-nearest-neighbor machine translation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, p. 2175–2187, 2022.

[12] Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. Improving named entity recognition by external context retrieving and cooperative learning. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, p. 1800–1812, 2021.

[13] Masahiro Kaneko, Sho Takase, Ayana Niwa, and Naoaki Okazaki. Interpretability for language learners using example-based grammatical error correction. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7176–7187, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. https://arxiv.org/abs/1907.11692.

[15] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In **Proceedings of the International Conference on Learning Representations**, 2017.

[16] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In **European Semantic Web Conference Springer**, pp. 593–607, 2018.

[17] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, p. 527–536, 2018.

[18] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. **Language resources and evaluation**, Vol. 42, No. 4, p. 335, 2008.

[19] Sayyed M Zahiri and Jinho D Choi. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In **Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence**, 2018.

[20] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In **Proceedings of the International Conference on Learning Representations**, 2019.

[21] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. **IEEE Transactions on Big Data**, Vol. 7, No. 3, pp. 535–547, 2019.