ラベル間の意味の違いを考慮した Few-shot テキスト分類

大橋 空 [1]　高山 隼矢 [1]　梶原 智之 [2]　荒瀬 由紀 [1]
[1] 大阪大学大学院情報科学研究科　[2] 愛媛大学大学院理工学研究科
[1]{ohashi.sora, takayama.junya, arase}@ist.osaka-u.ac.jp
[2]kajiwara@cs.ehime-u.ac.jp

1 はじめに

テキスト分類タスクにおいて、ニューラルネットワークに基づくモデル [1,2] が多数提案されており、大きな成功を収めている。特に BERT [3] による文の表現生成は多くのテキスト分類タスクにおいて最高性能を達成し、その有効性が示されている。しかしこれらの機械学習モデルでは、少数の事例しか持たないラベルが存在する場合、訓練時に過学習を起こしやすいという問題が知られている [4]。この問題に対処するために、ラベルごとに数個の事例のみが訓練データとして与えられる設定の Few-shot テキスト分類 [5–8] が活発に研究されている。

Few-shot テキスト分類では、距離学習やメタ学習を用いる手法が成功を収めている。例えば、様々な粒度の注意機構を用いて分類が容易なベクトル表現を生成する手法 [6] や、注意機構において単語の共起パターンを考慮する手法 [8] が提案されている。これらの先行研究ではラベル表現同士の類似度を考慮しないため、意味的に近いラベルが混同されやすく、それらの分類が難しいという課題がある。

この問題に対処するため、本研究ではラベル表現同士を比較し、各ラベル表現を互いに分離する学習手法を提案する。提案手法では、ラベル間の意味の違いを把握し、各ラベルに特有の情報のみをラベル表現に織り込んだ表現を生成する。これによりラベル間の違いが明確になり、分類が容易になると期待できる。

提案手法の有効性を検証するために、Huffpost [9] および FewRel [10] のデータセットを用いて Few-shot テキスト分類の実験を行った。実験の結果、両データセットにおいて提案手法がベースラインモデルの性能を有意に改善し、提案手法の有効性を確認できた。さらに提案手法は、分類対象のラベルが増加するにつれて、また各ラベルの事例数が減少するにつれて、その有効性が増すことが明らかとなり、より困難な設定の Few-shot 分類において分類精度の向上に貢献することが示された。

2 背景知識：Few-shot テキスト分類

2.1 問題定義

Few-shot テキスト分類では、サポート集合およびクエリ集合が入力として与えられる。サポート集合とは、テキストとラベルの組の集合 $S = \{(x_i, y_i)\}_{i=1}^{NK}$ である。N はサポート集合に含まれるラベルの種類であり、K はラベルごとのサンプル数である。以降、$N = n, K = k$ である場合を n-way k-shot 分類と呼ぶ。また、ラベル L のみが含まれるサポート集合を $S_L = \{(x_p, y_p)|y_p = L\}$ と表記する。クエリ集合とは、分類対象であるテキストの集合 $Q = \{q_j\}_{j=1}^{M}$ のことを指す。Few-shot テキスト分類モデルは、この各クエリ q_j に対してラベルを推定することを目的とする。なお、サポート集合 S が持つラベルの集合とクエリ集合 Q に対応するラベルの集合は等しい。

Few-shot 分類では、エピソード [11] と呼ばれるデータセットの部分集合を単位として訓練および評価を行う。訓練エピソードは、訓練用データから無作為に選択された N 種類のラベルについて、そのラベルを持つ事例を $K + m$ 個だけ無作為抽出し、サポート集合とクエリ集合に分割して作成する。ただし $m = \frac{M}{N}$ である。評価エピソードも評価用データから同様に作成し、クエリ集合ごとの正解率を平均してモデルの評価とする。

2.2 Few-shot テキスト分類モデル

本節では、Few-shot テキスト分類モデルの一般形について説明する。まず、サポート集合およびクエリ集合に含まれる文をそれぞれベクトル表現に変換する。S_i に属する文を s_i^p $(1 \leq p \leq K)$、Q に属する文を $q_j \in Q$、文符号化器を $E(\cdot)$ とすると、文のベクトル表現 $\boldsymbol{s}_i^p$ および $\boldsymbol{q}_j$ は次のようになる。

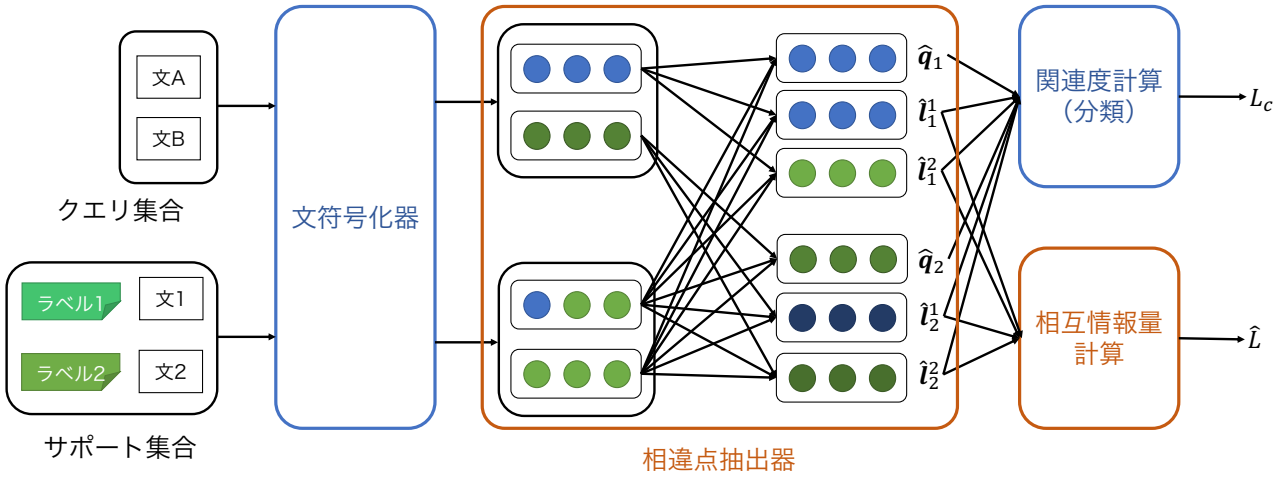

図 1　提案する Few-shot テキスト分類モデルの全体図（赤色の部分が提案手法）

$$\boldsymbol{s}_i^p = E\left(s_i^p\right) \tag{1}$$

$$\boldsymbol{q}_j = E(q_j) \tag{2}$$

文の符号化には、再帰的ニューラルネットワーク [2] や畳み込みニューラルネットワーク [1]、BERT [3] など、任意の手法を適用する。

次に、同じラベルを持つ K 文のベクトル表現からラベル表現を生成する。ラベル表現を生成する関数を $C(\cdot)$ とすると、ラベル表現 $\boldsymbol{l}_i$ は次のようになる。

$$\boldsymbol{l}_i = C\left(\boldsymbol{s}_i^1, \boldsymbol{s}_i^2, \cdots, \boldsymbol{s}_i^K\right) \tag{3}$$

$C(\cdot)$ として、平均プーリングや最大プーリング等が用いられる。

最後に、$\boldsymbol{l}_i$ と $\boldsymbol{q}_j$ の関連度を計算し、クエリ表現と最も関連度の高いラベル表現を選択することによってラベルを推定する。関連度を計算する関数を $R(\cdot)$ とすると、ラベルの確率分布は以下の式で計算できる。

$$p(i|\boldsymbol{l}_1, \cdots, \boldsymbol{l}_N, \boldsymbol{q}_j) = \frac{e^{R(\boldsymbol{l}_i, \boldsymbol{q}_j)}}{\sum_p e^{R(\boldsymbol{l}_p, \boldsymbol{q}_j)}} \tag{4}$$

$R(\cdot)$ にはコサイン類似度等、任意の手法を適用する。

分類の損失関数 L_c には、負の対数尤度を用いる。

$$L_c = -\frac{1}{M}\sum_{i=1}^{M} \log p(y_j) \tag{5}$$

ここで、y_j は q_j に対応する真のラベルを表す。

3　提案手法

図 1 に提案手法の概要を示す。本手法は、2.2 節で述べた構成の Few-shot テキスト分類モデルに、3.1 節の相違点抽出器と 3.2 節の損失関数を導入するマルチタスク学習のアプローチである。相違点抽出器を用いてより分類が容易となるラベルの特徴量を抽出することで、意味的に近いラベルの分類を容易にし、分類精度を向上させる。

3.1　相違点抽出器

相違点抽出器では、(3) 式で得たラベル表現 $\boldsymbol{l}_i$ を互いに比較し、各ラベルに固有の情報のみを持つラベル表現 $\hat{\boldsymbol{l}}_i$ に修正する。ここで、ラベル i に固有の情報を持つ理想的なベクトル表現 $\hat{\boldsymbol{l}}_i$ は、$\hat{\boldsymbol{l}}_j$ $(i \neq j)$ との相互情報量 $I(\hat{\boldsymbol{l}}_j; \hat{\boldsymbol{l}}_i)$ が 0 となる、すなわち各ラベル表現が独立となる、と仮定する。ただし、クエリ q_i に関して無関係な情報を抽出するのを防ぐため、ラベル表現とクエリ表現を同時に考慮する。具体的には、ラベル表現 $\boldsymbol{l}_1, \cdots, \boldsymbol{l}_N$ とクエリ表現 $\boldsymbol{q}_j$ を以下の式を用いて変換する。

$$\boldsymbol{H}_k = \mathrm{MultiHeadSelfAttention}(\boldsymbol{l}_1, \cdots, \boldsymbol{l}_N, \boldsymbol{q}_j) \tag{6}$$

$$\hat{\boldsymbol{l}}_i^j = \mathrm{GELU}(\boldsymbol{W}_1 \boldsymbol{H}_{k,l_i} + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2 \tag{7}$$

$$\hat{\boldsymbol{q}}_j = \mathrm{GELU}(\boldsymbol{W}_1 \boldsymbol{H}_{j,q_j} + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2 \tag{8}$$

ここで、MultiHeadSelfAttention$(\cdot)$ は 1-layer 8-head の自己注意機構 [12] の出力であり、GELU$(\cdot)$ は活性化関数 [13] である。$\boldsymbol{H}_{j,l_i} \in \mathbb{R}^{(N+1)\times d}$ は自己注意機構

表 1　実験結果（太字は ProtoNet と比べて、*は ProtoNet + 相違点抽出器と比べて、有意（$p < 0.05$）な性能改善を示す）

	Huffpost				FewRel			
	5-Way		10-Way		5-Way		10-Way	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
ProtoNet	51.03	68.36	37.42	55.81	78.61	88.92	65.97	80.38
ProtoNet + 相違点抽出器	**51.76**	**69.07**	**38.08**	**56.85**	77.35	88.85	64.96	80.44
ProtoNet + 相違点抽出器 + $\hat{L}$	**52.34***	**69.66***	**38.83***	**57.26***	**79.52***	**89.28***	**68.08***	**82.51***
Bao らの手法 [8]	42.12	62.97	-	-	70.08	88.07	-	-

の出力のうち、$\boldsymbol{I}_i$ に対応するベクトルを表す。d は入力ベクトルの次元数である。

3.2　損失関数の設計

本研究では L_c に加え、相違点抽出器が各ラベルに固有の情報のみを含むような制約を付加する損失関数 $\hat{L}$ を新たに定義する。

$$\hat{L} = \sum_{1 \le i,j \le N, i \ne j} I(\hat{\boldsymbol{I}}_i, \hat{\boldsymbol{I}}_j) \tag{9}$$

最終的な損失関数は次のようになる。

$$L = L_c + \beta \hat{L} \tag{10}$$

ここで、$\beta > 0$ は $\hat{L}$ の重みである。

相互情報量 $I(\hat{\boldsymbol{I}}_i; \hat{\boldsymbol{I}}_j)$ を直接計算するのは困難であるため、相互情報量の上界を最小化することで間接的に相互情報量を最小化する。具体的には、Cheng ら [14] に従い、以下を最小化する。

$$\hat{I}(\hat{\boldsymbol{I}}_i; \hat{\boldsymbol{I}}_j) = \sum_a R_a \tag{11}$$

$$R_a = \left[\log p_\theta(\hat{\boldsymbol{I}}_i^a | \hat{\boldsymbol{I}}_j^a) - \frac{1}{|Q|} \sum_b \log p_\theta(\hat{\boldsymbol{I}}_i^a | \hat{\boldsymbol{I}}_j^b) \right]$$

ここで、$p_\theta(\cdot)$ は、確率 $p(\hat{\boldsymbol{I}}_i^a | \hat{\boldsymbol{I}}_j^a)$ を、パラメータ θ のニューラルネットワークで近似した関数である。

4　実験

本実験では、相違点抽出器および損失関数 $\hat{L}$ の有効性を検証する。

4.1　実験設定

Bao ら [8] によって公開[1] されている以下の 2 つのデータセットを用いて実験する。

Huffpost　英語版ハフポストのタイトルから記事のカテゴリを推定するタスク。訓練用データ、検証用データ、評価用データには、それぞれ 20 種類、5 種類、16 種類のラベルが含まれており、事例数はラベルごとに 900 である。

FewRel　エンティティ間の関係を推定するタスク。訓練用データ、検証用データ、評価用データには、それぞれ 65 種類、5 種類、10 種類のラベルが含まれており、事例数はラベルごとに 700 である。

ベースラインとして、ProtoNet [4] を用いる。ただし、(1) 式および (2) 式の文の符号化には BERT [3] を使用し、(3) 式のラベル表現の生成には平均プーリングを用いる。提案手法として、ProtoNet + 相違点抽出器および ProtoNet + 相違点抽出器 + $\hat{L}$ を比較する。両者は共通のモデル構造だが、前者は 3.2 節の損失関数を用いず (5) 式の損失関数のみで訓練するものである。また、本実験設定において最高性能を達成している Bao ら [8] の手法とも比較する。

本実験では、全てのモデルを 5-Way 1-Shot の設定で訓練した。最適化アルゴリズムには Adam [15] を用いた。学習率は、1e-5, 3e-5, 5e-5 の中から、検証用データにおける正解率が最高の値を選択した。提案手法における (10) 式の重み β は、1e-6, 1e-4, 1e-2, 1 の中から、同様に検証用データを用いて選択した。

4.2　実験結果

実験結果を表 1 に示す。[2] 両データセットにおいて、提案手法はベースラインの性能を常に有意に改善した。また、$\hat{L}$ を用いずに訓練した場合、常に性能が悪化することも確認した。以上の結果から、ラベルに固有の情報を抽出することは Few-shot テキスト分類において有用であると言える。

1）https://github.com/YujiaBao/Distributional-Signatures

2）[8] では ProtoNet より高い性能が報告されているが、BERT の fine-tuning 設定を調整した結果、ProtoNet が Bao らの手法を上回った。

表 2　1-Shot 設定での Way 数による性能変化

Way	2	4	6	8	10
ProtoNet	92.13	82.14	75.24	70.33	65.97
提案手法	92.38	82.86	76.34	71.30	68.08
差分	0.25	0.72	1.10	0.97	2.11

表 3　5-Shot 設定での Way 数による性能変化

Way	2	4	6	8	10
ProtoNet	96.51	91.25	86.99	83.48	80.38
提案手法	96.68	91.51	87.43	84.09	82.51
差分	0.17	0.26	0.44	0.61	2.13

4.3　性能上昇幅に関する分析

本節では、Way 数および Shot 数を変化させたときの性能の変化について分析する。表 1 において、Huffpost では 5-Way 1-Shot 設定で 1.3 ポイント、10-Way 1-Shot 設定で 1.4 ポイントの改善が見られた。また、FewRel データセットにおいては 5-Way 1-Shot 設定で 0.4 ポイント、10-Way 1-Shot 設定で 2.1 ポイントの改善が見られた。これらの結果から、提案手法は Way 数が増加するにつれて有効性が増すと予想できる。これを検証するため、2-Way、4-Way、6-Way、8-Way における ProtoNet および提案手法の性能の変化を調査した。表 2 および表 3 に、FewRel データセットにおける実験結果を示す。なお、評価は 1-Shot および 5-Shot の設定で行った。Way 数が増加するにつれて、提案手法による性能の上昇幅が増加する傾向があることが確認できる。つまり、提案手法は分類対象が多くなるにつれ有効性が高まると言える。これは、分類対象となるラベルが増加するにつれて、意味的に類似したラベルが出現する確率が高くなるためであると推測される。

また、 1-Shot 設定と 5-Shot 設定を比較した結果、10-Way 設定を除いて、提案手法による性能の上昇幅は 1-Shot 設定の方が大きい。事例数が少なくなるほど、文符号化器によって分類に有効なラベル表現を生成することが困難になる。提案手法はラベル間の意味的な違いを抽出するため、その利得は事例数が少ない場合においより顕著になると考える。

5　関連研究

Few-shot 分類の先行研究は、距離学習を用いる手法とメタ学習を用いる手法の 2 つに大別できる。

距離学習を用いる Few-shot 分類モデルとしては、ユークリッド距離で分類を行う手法 [4]、注意機構を用いてラベル表現を生成しコサイン類似度を基に分類を行う手法 [11]、ニューラルネットワークを用いて関連度を計算する手法 [16]、グラフニューラルネットワークを用いた手法 [17] などが存在する。

メタ学習を用いる手法では、数回のパラメータ更新で良好な汎化性能が得られるよう分類器のパラメータを生成する手法 [18]、パラメータ更新方法や学習率を学習する手法 [19,20]、勾配からパラメータ更新方法をニューラルネットワークを用いて決定する手法 [21–23] が存在する。

自然言語処理における Few-shot 分類でも、距離学習やメタ学習を用いる手法が提案されている。例えば、ベクトル表現の計算時にクエリとサポートを相互に参照する手法 [5] や、様々な粒度の注意機構を適用し、より分類が容易なベクトル表現の獲得を目指した手法 [6,7]、単語の共起パターンから注意スコアを計算する手法 [8] などが存在する。

これらの既存研究では、ラベル間の意味の違いを考慮しておらず、意味的に近いラベルを誤分類する恐れがある。本研究は、ラベル同士を比較して意味的な違いを考慮することで、この問題の解決を目指すものである。

6　おわりに

本研究では、Few-shot テキスト分類モデルの性能改善に取り組んだ。既存手法ではラベル同士を明示的に比較しておらず、ラベル間の意味的な違いを陽に考慮することができなかった。このため、意味的に類似したラベルが出現すると意味の違いを把握できずに誤分類してしまう。これを解決するため、ラベル表現を比較し互いの相互情報量が低くなるようラベル表現を修正する手法を提案した。実験結果より、本手法はベースラインである ProtoNet を全ての設定で有意に上回ることが確認された。また、本手法は同時に分類するラベルの種類数が多くなるにつれてその有効性が高まることが確認された。

今後は、任意の Few-shot テキスト分類モデルに関する提案手法の有効性を調査する予定である。

謝辞

本研究は JST（AIP-PRISM, 課題番号: JPMJCR18Y1）の支援を受けたものです。

参考文献

[1] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1746–1751, 2014.

[2] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, 2016.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[4] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, Vol. 30, pp. 4077–4087, 2017.

[5] Zhi-Xiu Ye and Zhen-Hua Ling. Multi-Level Matching and Aggregation Network for Few-Shot Relation Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2872–2881, 2019.

[6] Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. Hierarchical Attention Prototypical Networks for Few-Shot Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 476–485, 2019.

[7] Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. Hybrid Attention-Based Prototypical Networks for Noisy Few-Shot Relation Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6407–6414, 2019.

[8] Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. Few-shot Text Classification with Distributional Signatures. In *Proceedings of the Eighth International Conference on Learning Representations*, 2020.

[9] Rishabh Misra. News Category Dataset, 2018.

[10] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4803–4809, 2018.

[11] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems*, Vol. 29, pp. 3630–3638, 2016.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998–6008, 2017.

[13] Dan Hendrycks and Kevin Gimpel. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units. *CoRR*, 2016.

[14] Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. Improving Disentangled Text Representation Learning with Information-Theoretic Guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7530–7541, 2020.

[15] Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, pp. 1–15, 2015.

[16] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to Compare: Relation Network for Few-Shot Learning. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

[17] Victor Garcia Satorras and Joan Bruna Estrach. Few-Shot Learning with Graph Neural Networks. In *Proceedings of the Sixth International Conference on Learning Representations*, 2018.

[18] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70, pp. 1126–1135, 2017.

[19] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-SGD: Learning to Learn Quickly for Few Shot Learning. *CoRR*, 2017.

[20] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your MAML. In *Proceedings of the Seventh International Conference on Learning Representations*, 2019.

[21] Marcin Andrychowicz, Misha Denil, Sergio Gómez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, Vol. 29, pp. 3981–3989, 2016.

[22] Sachin Ravi and Hugo Larochelle. Optimization as a Model for Few-Shot Learning. In *Proceedings of the Fifth International Conference on Learning Representations*, 2017.

[23] Ke Li and Jitendra Malik. Learning to Optimize. In *Proceedings of the Fourth International Conference on Learning Representations*, 2016.