Mixture-of-Experts の悲観的な統合による頑健な自然言語理解

本多右京 [1]　岡達志 [2]　張培楠 [1]　三田雅人 [1]
[1] 株式会社サイバーエージェント　[2] 慶應義塾大学
{honda_ukyo,zhang_peinan,mita_masato}@cyberagent.co.jp　tatsushi.oka@keio.jp

概要

自然言語理解タスクのデータでは，ショートカットと呼ばれる，ラベルと擬似相関をもつ単純な特徴量が存在することがある．擬似相関はデータ分布の変動に対して頑健でないため，ショートカットへの依存は分布外データでの性能低下につながる．先行研究ではショートカットに依存しないモデルの学習が目標とされてきたが，この学習には実用上大きな困難が伴う．本研究ではこの学習を直接の目標とせず，それぞれ異なる潜在特徴量に基づいて予測する mixture-of-experts モデルをルールにより悲観的に統合することで，頑健に予測する手法を提案する．実験により，実用的な設定において先行研究を上回る分布外性能を示すことを確認した．

1　はじめに

含意関係認識，言い換え同定，事実検証などの自然言語理解タスクのデータセットでは，ラベルと相関する単純な特徴量が存在することが知られている [1, 2, 3, 4]．これらはデータ作成者が用いた経験則や選好，あるいは自然言語の構成的な性質などによって意図せず生じるとされ，様々なデータセットで確認されている [1, 5, 6]．しかし，例えばデータ作成者やデータ作成時の指示が変われば，これらの特徴量とラベルとの相関は容易に変化しうる．図 1 のように，学習データでは含意ラベルの事例で単語一致率が高かったとしても，テストデータの作成者が単語の入れ替えを好んで使った場合には，この単語一致率とラベルとの相関は変化してしまう．このように，学習データにおいてはラベルに対して予測力を持つが，テスト時のデータ分布の変動によって容易にその予測力を失うような単純な特徴量のことを，**ショートカット**という [7, 8]．言い換えれば，ラベルと擬似相関する単純な特徴量を指す．

ショートカットは単純な特徴量でありながらラベルを予測可能であるため，モデルにとって捉えやすく，依存しやすい．しかし，データ分布の変動に対して頑健でないため，これに依存したモデルは分布外データにおいて性能が低下する可能性が高い．そこで，先行研究ではショートカットへの依存を抑制する手法が提案されてきた [9, 10, 11, 12, 13, *i.a.*]．

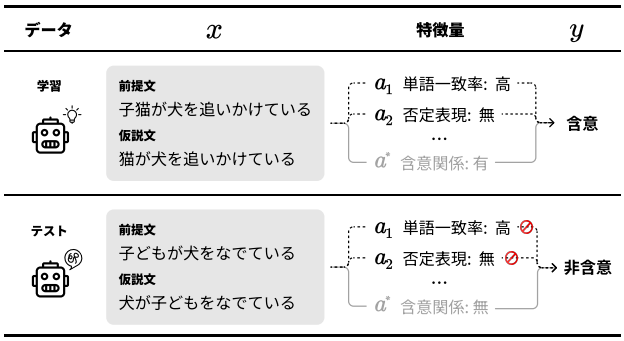

図 1　含意関係認識タスクにおけるショートカットの例．学習データにおいては単語一致率や否定表現の有無がラベルと相関するが，分布外のテストデータではこの相関が変化してしまう．a^* は理想的な，どの分布でも正しくラベルを予測できる特徴量だが，この学習は困難である．

これらの手法では，学習時にショートカットへの依存を抑制することで，どの分布でも正しくラベルを予測可能なモデルを学習することを目標にしている．しかし，学習データとその分布内データでは正しくラベルを予測できる特徴量の利用をあえて抑制することになるため，実際には分布内外のデータにおいて性能のトレードオフが生じてしまう．このトレードオフは分布外データでのハイパーパラメータ調整の必要性など，実用上深刻な問題をもたらす．

本研究では，このように困難な学習を直接の目的とせず，別のアプローチを検討する．ショートカットにおける分布外データでの問題は，いずれかの特徴量についてラベルとの相関が変化するが，分布外データは通常未知であるので，どの特徴量について変化するかはわからないということである．我々はこれを，どの特徴量に基づいて予測するべきかわからない不確実性のもとでの意思決定の問題として考える．悲観的な想定のもとで最小のリ

スクを達成するようにラベルを選択することで，未知の分布外データでも頑健に予測させる．学習時は mixture-of-experts モデルを，各エキスパートが異なる潜在特徴量に基づいて予測するよう促進して学習し，推論時はどのエキスパートを用いればよいかわからない状況でのリスクを最小化するようにエキスパートの予測を統合する．3 つの自然言語理解タスクで実験を行い，分布内データのみで学習とハイパーパラメータ調整を行う現実的な実験設定において，既存手法を上回る分布外性能を確認した．

2 関連研究と問題点

自然言語理解タスクでは，学習事例の再重み付けがショートカット依存抑制の主要なアプローチとなってきた [9, 10, 11]．再重み付けでは，ショートカットのラベル予測力に応じて，学習時の損失に重みをかける．これにより，ショートカットによって簡単にラベルが予測できる事例での学習を抑え，そうでない事例での学習を促す．主に画像分類タスクでは，データをショートカットに基づいたグループに分割する．各グループはそれぞれ異なるショートカットをもつことになるので，どのグループにおいても同時に最適となるモデルを学習する IRM [12] や，最も損失が高いグループでの損失を最小化するよう学習する GroupDRO [13] は，どのショートカットにも依存しないよう学習することになる．

これらの手法の目標は，ショートカットに依存せず頑健に予測できるモデルの学習である．しかし実際には，学習データとその分布内データでは有効な特徴量をあえて学習から排除することで学習データの分布から逸脱し，分布内外での性能のトレードオフが生じている．分布内データでの性能低下はそれ自体問題であるが，このトレードオフから生じる実用上より深刻な問題として，ハイパーパラメータ調整に分布外データが必要となることが指摘されてきた [9, 11, 14, 15, 16, 17, 18]．これは，トレードオフにより，分布内データでの性能を見るだけでは分布外データでの性能を予測できないことによる．

ショートカットは意図せず生じるものであり，その特定にはデータの詳細な分析が必要となる [1, 2, 3, 4]．したがって，ショートカットの特定自体と，さらにそれに関する分布外データの準備はコストが大きく，これらを常に要求することは現実的ではない．このコストを懸念して学習データに対してはショートカットの特定を必要としない手法も提案されてきたが，依然としてハイパーパラメータ調整のための評価データには事前に特定されたショートカットの情報を必要とする [14, 15, 16, 17, 18]．この情報を用いずにハイパーパラメータ調整を行った場合，分布外データでの性能改善は顕著に小さくなることが報告されている [19]．

3 提案手法

未知のショートカットに関する分布変動では，いずれかの特徴量とラベルとの相関が変化するが，どの特徴量について変化するかはわからない．そこで本研究では，どの特徴量に基づいて予測するべきかわからないという不確実性のもとで最小のリスクを達成するようにラベルを選択することで，未知の分布外データでも頑健に予測する手法を検討する．学習時の目標はそれぞれ異なる特徴量に基づいて予測するモデルの学習となり，既存手法のような直接的なショートカット依存抑制ではなくなるため，上述の実用上の困難が緩和される．提案手法は，学習時と推論時の 2 つのパートから成る．

3.1 学習時：異なる特徴量のモデル化

Mixture-of-Experts モデル　自然言語の構成的な性質やデータ作成プロセスの偏りから，自然言語理解タスクのデータでは，タスクに本質的な特徴量に加えて複数の特徴量がラベルと相関しうる [1, 5, 6]．学習時はこのデータ構造を mixture-of-experts [20] によってモデル化する．推論時の後処理を容易にするため，分類器の最終層でのみエキスパートに分岐する **mixture-of-softmax (MoS)** [21] を特に用いる．

入力を $x \in \mathcal{X}$，ラベルを $y \in \mathcal{Y}$，潜在特徴量の数を K としたとき，MoS での条件付き確率 $p(y \mid x)$ は，

$$p_{\boldsymbol{\theta}}(y \mid x) = \sum_{k=1}^{K} p^k(y \mid \phi(x))\, \pi_k(\phi(x)). \tag{1}$$

$p^k \in \Delta^{|\mathcal{Y}|-1}$ は k 番目のエキスパートの予測で，$\pi \in \Delta^{K-1}$ はエキスパートの混合率，$\pi_k \in \mathbb{R}$ は π の k 番目の要素である．ただし，Δ^N は N 次元の単体を指す．ϕ はエンコーダを，$\boldsymbol{\theta}$ は p^k，π，ϕ のパラメータ全体を指す．学習は，事例数 M のミニバッチに対して以下の負の対数尤度を最小化するよう行う．

$$L_{\text{main}}(\boldsymbol{\theta}) = -\frac{1}{M} \sum_{i=1}^{M} \log p_{\boldsymbol{\theta}}(y_i \mid x_i). \tag{2}$$

補助損失　この MoS モデルでは，π が各事例で顕著な潜在特徴量を推定し，p^k はその潜在特徴量に

基づいてラベルを予測することが期待される．しかし，学習中に事例をまたいで特定の π_k に確率が集中したり，あるいは π が常に一様分布であったりした場合，実質的に単一のエキスパートしか存在しない状態となる [22]．これを避けるため，学習に次の補助損失を加える．

$$L_{\text{aux}}(\boldsymbol{\theta}) = \frac{\|d_s(\boldsymbol{\Pi}^\top \boldsymbol{\Pi} - \mathbf{I})\|_F}{\|d_s(\mathbf{J} - \mathbf{I})\|_F}. \tag{3}$$

$\boldsymbol{\Pi} \in \mathbb{R}^{K \times M}$ は，事例数 M のミニバッチにおける $\pi(x_1), ..., \pi(x_M)$ の連結とする．簡略化のため，ϕ は省略している．$\boldsymbol{I} \in \mathbb{R}^{M \times M}$ は単位行列，$\boldsymbol{J} \in \mathbb{R}^{M \times M}$ はすべての要素が 1 の行列，$\|\cdot\|_F$ は行列のフロベニウスノルムを指す．$\|\boldsymbol{\Pi}^\top \boldsymbol{\Pi} - \mathbf{I}\|_F$ は Li ら [23] の自己注意機構に対する制約に基づいており，$M \leq K$ であれば，事例ごとに異なる k において π が one-hot である場合に値が 0 になる．これにより，π が事例ごとに異なる潜在特徴量を推定するよう促す．しかし，$M > K$ である場合，すべての π が one-hot であっても同一の π が必ず発生し，値が 0 にならない．そこで，各行でスコア $\pi(x_i)^\top \pi(x_j)$ が上位 s 個に入る事例同士では同じ潜在特徴量が顕著であるとして，その s 個の要素の値を 0 にする関数 d_s を導入する．分母は損失の値を [0, 1] に収めるよう正規化する．

最終的に最小化する損失は，重みのハイパーパラメータ α を用いて以下となる．

$$L(\boldsymbol{\theta}) = L_{\text{main}}(\boldsymbol{\theta}) + \alpha L_{\text{aux}}(\boldsymbol{\theta}). \tag{4}$$

3.2 推論時：不確実性下の意思決定

分布内データであれば π により予測に用いるべき潜在特徴量がわかるが，分布外データではこれがわからない．そこで，決定理論に基づいて，この不確実性のもとでリスクを最小化するよう意思決定を行うことを考える [24, 25]．

提案モデルのもとで，x から y を決定する関数 $\delta : \mathcal{X} \to \mathcal{Y}$ の集合 $\mathcal{D}$ から，リスクを最小化する δ を選びたい．このリスクを次のように定義する．

$$R(\pi, \delta) = 1 - \mathbb{E}_x\Big[\sum_{k=1}^{K} p^k\big(\delta(x)|x\big)\, \pi_k(x)\Big]. \tag{5}$$

分布内データであれば，π を用いて以下のようにラベルを決定することでリスクを最小化できる．

$$\delta^*_\pi(x) = \arg\max_{y \in \mathcal{Y}} \sum_{k=1}^{K} p^k\big(y|x\big)\, \pi_k(x). \tag{6}$$

しかし，分布外データではこの限りでない．不確実性のもとでリスクを最小化する方法として，まず **uniform weighting** を導入する．これは，どの潜在特徴量でも同程度に正しく予測できると仮定して，エキスパートの予測の平均をとる手法である．

$$\delta^*_u(x) = \arg\max_{y \in \mathcal{Y}} \frac{1}{K} \sum_{k=1}^{K} p^k\big(y|x\big). \tag{7}$$

この仮定は必ずしも成り立たないため，より悲観的な状況を考える手法として，**argmin weighting** を導入する．ラベルごとに最大のリスクを考えて，その最悪の場合で最小のリスクを達成する minimax 問題 $\min_{\delta \in \mathcal{D}} \max_{\pi \in \Delta^{K-1}} R(\pi, \delta)$ を考え，以下のようにラベルを決定する（図 3 参照）．ただし，$\mathcal{H} \in \{1, ..., K\}$．

$$\delta^*_w(x) = \arg\max_{y \in \mathcal{Y}} \min_{k \in \mathcal{H}} p^k\big(y|x\big). \tag{8}$$

これらはいずれも軽量なルールベースの後処理であるため，推論時に容易に適用可能である．

4 実験

4.1 実験設定

データセット　比較対象とする先行研究にしたがい，含意関係認識 **MNLI** [29]，言い換え同定 **QQP**，事実検証 **FEVER** [30] の 3 つのデータセットで実験を行った．いずれのデータセットも，分布内開発データ (**Dev**) と，ショートカットとラベルの相関が大きく変動した分布外テストデータから成る．分布外テストデータはそれぞれ，**HANS** [2]，**PAWS** [3]，**FEVER Symmetric v1 and v2** [30] を用いた．

比較対象　ベースラインは $\text{BERT}_{\text{base-uncased}}$ [31] とし，提案モデルのエンコーダ ϕ にもこれを用いた．また，主要な既存手法を選び比較を行った．詳細は付録 A を参照．いずれも学習時にショートカットが未知である設定で提案された手法だが，評価データにおいてショートカットの情報が利用可能であった．そこで，本実験ではすべての手法について，分布内開発データでハイパーパラメータ調整を行う設定 [19] で再評価を行った．

ハイパーパラメータ調整　提案手法のハイパーパラメータは，エキスパート数 K，損失の重み α，補助損失の s の 3 つである．データセットごとに以下の探索を行った．まず，最も自然にデータ構造を表現できる K を特定するため，$\alpha = 0$ とした上で，$K \in \{5, 10, 15\}$ から，Dev スコアが最大となるエポッ

表 1　ベースライン・既存手法との比較結果．スコアは accuracy で，5 回の試行の平均と標準偏差である．背景色がついているのが提案手法で，破線から下は既存手法．太字は最も高い平均スコアを示す．

	MNLI		QQP		FEVER		
	Dev	HANS	Dev	PAWS	Dev	Symm. v1	Symm. v2
$\text{BERT}_{\text{base-uncased}}$	$84.4_{\pm 0.2}$	$55.2_{\pm 4.2}$	$91.5_{\pm 0.1}$	$36.7_{\pm 3.1}$	$86.7_{\pm 0.2}$	$58.5_{\pm 1.4}$	$65.1_{\pm 1.5}$
MoS	$84.4_{\pm 0.1}$	$59.4_{\pm 5.5}$	$91.4_{\pm 0.1}$	$34.9_{\pm 1.6}$	$87.0_{\pm 0.5}$	$58.9_{\pm 1.2}$	$65.5_{\pm 1.0}$
→ uniform	$83.0_{\pm 1.0}$	$63.6_{\pm 5.7}$	$89.1_{\pm 2.4}$	$47.0_{\pm 8.6}$	$87.6_{\pm 1.2}$	$\mathbf{62.2}_{\pm 0.7}$	$\mathbf{68.2}_{\pm 1.0}$
→ argmin	$81.0_{\pm 3.1}$	$\mathbf{67.2}_{\pm 4.6}$	$83.8_{\pm 7.3}$	$\mathbf{55.7}_{\pm 8.5}$	$85.3_{\pm 6.8}$	$61.8_{\pm 1.2}$	$67.4_{\pm 2.2}$
Conf-reg ♠ self-debias [26]	$84.5_{\pm 0.2}$	$63.7_{\pm 2.4}$	$90.5_{\pm 0.2}$	$31.0_{\pm 1.7}$	$87.1_{\pm 0.7}$	$59.7_{\pm 1.3}$	$66.5_{\pm 1.1}$
JTT [16]	$80.7_{\pm 0.3}$	$57.3_{\pm 2.2}$	$89.4_{\pm 0.2}$	$36.0_{\pm 0.6}$	$82.7_{\pm 1.1}$	$53.0_{\pm 2.6}$	$60.3_{\pm 2.6}$
RISK [27]	$83.9_{\pm 0.3}$	$56.3_{\pm 4.2}$	$90.5_{\pm 0.1}$	$34.8_{\pm 3.2}$	$87.6_{\pm 0.8}$	$58.9_{\pm 2.6}$	$65.9_{\pm 1.6}$
EIIL [17]	$83.9_{\pm 0.2}$	$61.5_{\pm 2.4}$	$91.1_{\pm 0.2}$	$31.0_{\pm 0.6}$	$86.8_{\pm 1.1}$	$56.2_{\pm 1.9}$	$63.8_{\pm 1.7}$
BAI [18]	$83.7_{\pm 0.2}$	$62.0_{\pm 2.1}$	$91.2_{\pm 0.2}$	$31.2_{\pm 0.3}$	$86.3_{\pm 1.2}$	$56.0_{\pm 2.1}$	$63.6_{\pm 1.9}$
$\text{GroupDRO}_{\text{label-group}}$ [13, 19]	$84.3_{\pm 0.3}$	$57.7_{\pm 2.9}$	$\mathbf{91.6}_{\pm 0.1}$	$34.6_{\pm 3.7}$	$\mathbf{89.3}_{\pm 0.2}$	$62.1_{\pm 1.1}$	$67.9_{\pm 1.3}$
ReWeightCRT [28]	$\mathbf{84.6}_{\pm 0.1}$	$55.8_{\pm 0.3}$	$91.5_{\pm 0.0}$	$32.0_{\pm 0.3}$	$88.5_{\pm 0.0}$	$61.3_{\pm 0.4}$	$66.9_{\pm 0.2}$

表 2　Ablation study. 背景色部分は表 1 と同じ結果．

		Dev	HANS
$\alpha = 0.5$	MoS	$84.4_{\pm 0.1}$	$59.4_{\pm 5.5}$
	→ argmin	$81.0_{\pm 3.1}$	$\mathbf{67.2}_{\pm 4.6}$
$\alpha = 0.0$	MoS	$\mathbf{84.5}_{\pm 0.0}$	$57.6_{\pm 4.8}$
	→ argmin	$76.0_{\pm 9.5}$	$60.7_{\pm 4.4}$
$\text{DeBERTa}_{\text{v3-large}}$	MoS	$\mathbf{91.8}_{\pm 0.0}$	$66.3_{\pm 1.8}$
	→ argmin	$86.5_{\pm 4.9}$	$\mathbf{74.4}_{\pm 8.4}$

クで $L_{\text{main}} + L_{\text{aux}}$ が最も小さくなる K を選択した．K をこれに固定したうえで，次に $\alpha \in \{0.0, 0.5, 1.0\}$ から同じく $L_{\text{main}} + L_{\text{aux}}$ が最も小さくなる α を選択した．s は 2^n のうち，π の出力が事例ごとに異なる one-hot であった場合に，$K = 5$ でも L_{aux} が 0 になる最小の数に設定した．実験では $M = 32$ のミニバッチを 2 並列処理したので，$s = 8$ とした．選択された値やその他の詳細は付録 B を参照．

4.2　実験結果

分布外性能比較　表 1 に結果を示す．提案モデル（MoS）はいずれのタスクにおいても分布内開発データ（Dev）でベースラインと同等の高い性能を示しているが，提案した後処理 (→ **uniform**/**argmin**) を適用するだけで，分布外テストデータでの性能が顕著に向上している．既存手法は，Dev でのハイパーパラメータ調整では分布外性能の改善が小さく，先行研究 [19] と整合的な結果となっている．

Ablation Study　ハイパーパラメータ調整の結果，MNLI における α の最適な値は 0.5 であったが，これを 0 にした場合との比較を行い，補助損失 L_{aux} の効果を検証した．結果を表 2 に示す．$\alpha = 0.5$ では MoS の性能は $\alpha = 0.0$ のときとほぼ変わらないのに対して，後処理を適用した場合の HANS での性能は大きく向上しており，補助損失が分布外性能の改善に大きく寄与していることがわかる．また，エンコーダ ϕ にサイズのより大きい $\text{DeBERTa}_{\text{v3-large}}$ [32] を用いた場合の結果も示した．大規模なモデルでは分布外性能が高くなるが，後処理によってさらに改善しており，提案手法の有効性を示している．

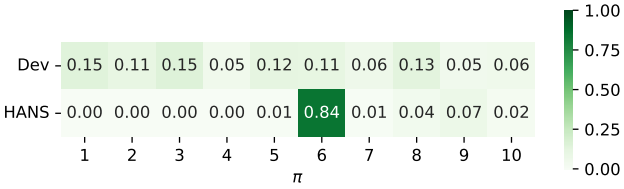

図 2　MNLI データセットにおける π の平均値．$K = 10$.

解釈性　提案モデルの解釈性を検証するため，MNLI データそれぞれでの π の平均値を図 2 に示す．HANS の仮説文は前提文の一部を変更することで作られるため，全事例にわたって両文の単語一致率が高い．図 2 では HANS において特定のエキスパートに割当てが集中しており，単語一致率の高さという特徴量をこのエキスパートが捉えていることを示唆している．付録 C に示すとおり，その他のデータセットについても同様の傾向であった．

5　おわりに

本研究では，mixture-of-experts モデルを悲観的に統合することで，ショートカットに関する分布変動に対して頑健に予測する手法を提案した．実験により，分布内データでのみ学習・調整を行う実用的な設定での高い分布外性能を確認した．今後の課題としては，分布内外での後処理の適応的な使い分けや，解釈性の改善，データ中の異なる潜在特徴量を捉えられることの理論的保証の検討などがある．

参考文献

[1] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In **NAACL-HLT**, 2018.

[2] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In **ACL**, 2019.

[3] Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In **NAACL-HLT**, 2019.

[4] Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. In **EMNLP-IJCNLP**, 2019.

[5] Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In **EMNLP-IJCNLP**, 2019.

[6] Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. Competency problems: On finding and removing artifacts in language data. In **EMNLP**, 2021.

[7] Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In **AISTATS**, 2022.

[8] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. **TACL**, Vol. 10, pp. 1138–1158, 2022.

[9] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In **EMNLP-IJCNLP**, 2019.

[10] He He, Sheng Zha, and Haohan Wang. Unlearn dataset bias in natural language inference by fitting the residual. In **Workshop on Deep Learning Approaches for Low-Resource NLP**, 2019.

[11] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. End-to-end bias mitigation by modelling biases in corpora. In **ACL**, 2020.

[12] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. **arXiv:1907.02893v3**, 2019.

[13] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In **ICLR**, 2020.

[14] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Learning to model and ignore dataset bias with mixed capacity ensembles. In **Findings of ACL: EMNLP**, 2020.

[15] Abbas Ghaddar, Phillippe Langlais, Mehdi Rezagholizadeh, and Ahmad Rashid. End-to-end self-debiasing framework for robust NLU training. In **Findings of ACL: ACL-IJCNLP**, 2021.

[16] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In **ICML**, 2021.

[17] Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In **ICML**, 2021.

[18] Sicheng Yu, Jing Jiang, Hao Zhang, Yulei Niu, Qianru Sun, and Lidong Bing. Interventional training for out-of-distribution natural language understanding. In **EMNLP**, 2022.

[19] Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. In **ICML**, 2023.

[20] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. **Neural computation**, Vol. 3, No. 1, pp. 79–87, 1991.

[21] Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In **ICLR**, 2018.

[22] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In **ICLR**, 2017.

[23] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In **ICLR**, 2017.

[24] Abraham Wald. **Statistical Decision Functions**. Wiley: New York, 1950.

[25] James O. Berger. **Statistical Decision Theory and Bayesian Analysis**. Springer-Verlag, New York, 1985.

[26] Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. Towards debiasing NLU models from unknown biases. In **EMNLP**, 2020.

[27] Ting Wu and Tao Gui. Less is better: Recovering intended-feature subspace to robustify NLU models. In **COLING**, 2022.

[28] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In **ICLR**, 2020.

[29] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In **NAACL-HLT**, 2018.

[30] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In **NAACL-HLT**, 2018.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **NAACL-HLT**, 2019.

[32] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In **ICLR**, 2023.

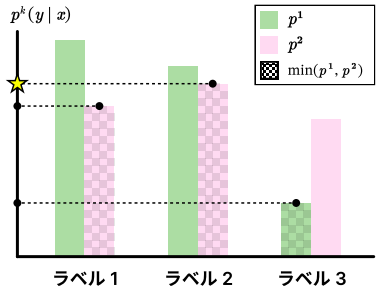

図 3　Argmin weighting を適用した際のラベル選択の例．エキスパート数 $K = 2$，ラベル数 $|\mathcal{Y}| = 3$ としている．Argmin weighting 適用後（網掛け部分）で最もリスクが小さいラベル 2 が選択される．

A　比較対象手法の詳細

比較対象とした手法のうち，**Conf-reg** ♠ $_{\text{self-debias}}$ [26] と **JTT** [16] は再重み付けを行う手法で，弱い分類器がショートカットに依存しやすいという経験則を活用して，弱い分類器で分類が難しい事例に重みをかける．**RISK** [27] は特徴量削減によって，ショートカットとなる冗長な特徴量への依存を防ぐ．**EIIL** [17] はショートカットに基づくグループを推定した後に，それに基づいて IRM [13] で学習する手法である．**BAI** [18] はその拡張で，グループ推定と IRM での学習を複数回行う．**GroupDRO**$_{\text{label-group}}$ [13, 19] は，データをショートカットではなくラベルごとのグループに分割したうえで，GroupDRO で学習する手法である．GroupDRO では事前にショートカットが特定されている必要があるため，この要求を緩和したものとして提案された [19]．**ReWeightCRT** [28] はラベル不均衡データでの学習手法で，学習データでのラベル頻度によって損失を重み付けする．テスト時にラベル頻度が均等である場合に，分布外データでも性能改善があることが報告されている [19]．IRM と GroupDRO の詳細は 2 節を参照．

いずれの手法も，公開されているコードに基づいて，論文記載のハイパーパラメータを用いて再現実験を行った．ショートカットが事前にわからない現実的な設定での比較のため，モデルは分布内開発データでの性能が最も高くなるエポックのものを選択した．Conf-reg ♠ $_{\text{self-debias}}$ では，分布外テストデータで調整していた annealing のためのハイパーパラメータについても分布内データで調整した．

B　ハイパーパラメータ調整の詳細

4.1 節で説明したハイパーパラメータ調整の結果，MNLI ではエキスパート数 $K = 10$，損失への重み $\alpha = 0.5$ が最適となった．QQP では $K = 15$，$\alpha = 1.0$，FEVER では $K = 10$，$\alpha = 1.0$ であった．学習時は，学習率 2e-5 で 10 エポック学習を行った．

DeBERTa$_{\text{v3-large}}$ でも同様にハイパーパラメータ調整を行った結果，MNLI データにおいて，BERT$_{\text{base-uncased}}$ と同じ値が選択された．学習のエポック数は同じく 10 としたが，先行研究にしたがって，学習率は 5e-6 とし，最大勾配ノルム 1.0 として勾配クリッピングを用いた [32]．

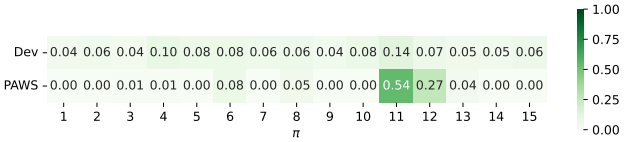

図 4　QQP データセットにおける π の平均値．$K = 15$.

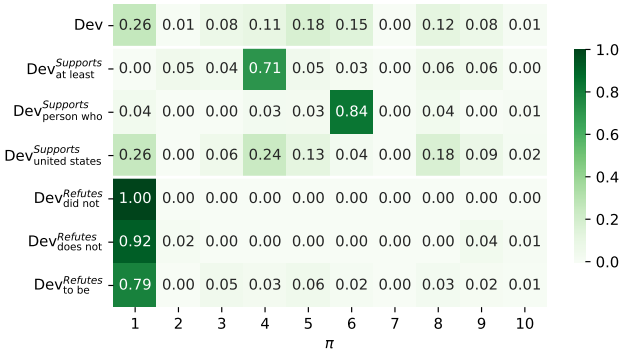

図 5　FEVER データセットにおける π の平均値．$K = 10$. Dev$^{\text{label}}_{\text{bigram}}$ は，そのラベルと相関する bigram を含む事例だけで構成されたデータを示す．

C　その他データでの解釈性

4.2 節での分析に続き，特定の特徴量が顕著に存在するデータにおいて，特定のエキスパートが集中して用いられる傾向があるか検証する．PAWS では，元となる文に対して単語の入れ替えや逆翻訳を行い，これを言い換え候補として用いる．このため，HANS と同様に全事例にわたって文同士の単語一致率が高い．FEVER Symmetric ではこのような顕著な特徴量が存在しないが，以下のようにして特定の特徴量が顕著に存在するデータ分割を作成した．FEVER では特定の bigram がラベルと強く相関することが報告されているため，FEVER の Dev のうち，特定の bigram を含む事例だけを集めて一つの分割とした．図 4, 5 から，Dev 以外の，特定の特徴量が顕著に存在するデータでは，特定のエキスパートに割当てが集中していることがわかる．これは，提案モデルにおいてエキスパートがそれぞれ異なる潜在特徴量を捉えていることを示唆する結果である．