

マルチタスク学習モデルにおける有効な補助タスクの選定

古田 朋也[†] 鈴木 優^{††}

[†] 岐阜大学大学院自然科学技術研究科知能理工学専攻 〒501-1193 岐阜県岐阜市柳戸1番1

^{††} 岐阜大学工学部電気電子・情報工学科 〒501-1193 岐阜県岐阜市柳戸1番1

E-mail: †a4525067@edu.gifu-u.ac.jp, ††ysuzuki@gifu-u.ac.jp

あらまし マルチタスク学習とは、主となるタスク以外のタスクも同時に解くことができるようにモデルの学習を進めることによって、特定のタスクのみに限定されない特徴量を獲得し、汎化性能を向上させる学習手法である。マルチタスク学習に関する研究は、効率良く学習を進めるための手法の提案や最適なタスクの組合せを効率良く探索する手法の提案など様々行われている。しかし、補助タスクの選定は感覚的に行っている研究が多く、採用する補助タスクの最適な数や選定基準については明らかになっていない。そこで本研究では、マルチタスク学習モデルを実装し、どのような補助タスクをどの程度採用すると推定精度向上に効果的なのかを調査するために実験を行った。実験では、商品タイトルの興味を惹く度合い推定を主タスクとしたマルチタスク学習モデルを複数構築して、推定精度を比較・分析することによって補助タスクの採用数による推定精度の変動や推定精度向上に効果的な補助タスクの特徴について調査した。本実験を通して、マルチタスク学習モデル構築時に選定する補助タスクの採用数を増やすことによって推定精度向上が期待できること、シングルタスク学習において十分な推定精度を発揮できるタスクであるという前提の下であれば、主タスクとの相関が強い補助タスクほど推定精度向上に有効であるという知見が得られた。

キーワード マルチタスク学習, 補助タスク選定, ニューラルネットワーク

本稿では、商品タイトルの興味を惹く度合い推定を主タスク

1 はじめに

マルチタスク学習とは、主となるタスク以外のタスクも同時に解くことができるようにモデルの学習を進めることによって、特定のタスクのみに限定されない特徴量を獲得し、汎化性能を向上させる学習手法である。マルチタスク学習に関する研究は、画像処理やテキスト処理に適用した研究以外にも、効率良く学習を進めるための手法の提案をしている研究や最適なタスクの組合せを効率良く探索する手法の提案をしている研究など様々なものが行われている。しかし、採用する補助タスクについては主タスクと関連していると思われるタスクを感覚的に選定している研究が多く、明確な基準に従って選定されているわけではない。また、採用する補助タスクの数について言及している研究や補助タスクの明確な選定基準を示している研究は調べた限りでは行われていない。そのため、マルチタスク学習モデルを構築する際にどのような補助タスクをどの程度採用すべきなのかは明らかになっていない。選定基準が明確ではないため、実際に各タスク用のデータを作成してモデルを構築してみないと補助タスクの良し悪しが判断できない。それでは、補助タスクが効果的ではなかった場合に作成したデータが無駄になってしまう。そこで、本研究ではマルチタスク学習モデルを構築する際にどのような補助タスクをどの程度採用すると推定精度向上に効果的なのかを調査する。どのような補助タスクをどの程度採用すると推定精度向上に効果的なのかを調査して補助タスクの選定基準を定めることによって、構築するモデルの推定精度向上に役立つだけでなく、効果的でないタスクのデータを作成する必要がなくなるためデータ作成の面でも役立つと考える。

としたマルチタスク学習モデルを複数構築して、推定精度を比較・分析することによって補助タスクの採用数による推定精度の変動や推定精度向上に効果的な補助タスクの特徴について調査する実験を行った。実験では、主タスクである興味を惹く度合い推定に加えて5つの補助タスクを用意した。複数人にアンケートを取るによって作成したデータセットを使用して、用意した5つの補助タスク全通りの組合せである32種類のマルチタスク学習モデルを10個ずつ構築した。その際、BERT [1]を用いた興味を惹く度合い推定を行うシングルタスク学習モデルをベースにしてマルチタスク学習モデルを構築した。その後、構築したモデルのテスト用データ推定精度に対して採用した補助タスクの組合せごとの平均値を算出し、それを比較・分析することによって、補助タスクの採用数による推定精度の変動や推定精度向上に効果的な補助タスクの特徴について調査した。

補助タスクの採用数についての分析では、補助タスクの採用数と採用数ごとの推定精度の間に相関が見られるかどうかを調査した。推定精度向上に効果的な補助タスクの特徴についての分析では、補助タスク用データと主タスク用データの相関係数の値と補助タスクを採用した際の効果の大きさに相関が見られるかどうかを調査した。それぞれの分析結果から、マルチタスク学習モデル構築時に選定する補助タスクの採用数を増やすことによって推定精度向上が期待できること、シングルタスク学習において十分な推定精度を発揮できるタスクであるという前提の下であれば、主タスクとの相関が強い補助タスクほど推定精度向上に有効であるという知見が得られた。この知見は、マルチタスク学習モデル構築時における補助タスク選定方法の一基準となり得るものであると考える。

本論文における貢献は以下の通りである。

- マルチタスク学習モデル構築時に選定する補助タスクの採用数の増加に伴う推定精度向上を確認した。
- シングルトask学習において十分な推定精度を発揮できるタスクであるという前提の下であれば、主タスクとの相関が強い補助タスクほど推定精度向上に有効であることを確認した。

2 関連研究

マルチタスク学習 [2] [3] とは、主となるタスク以外のタスクも同時に解くことができるようにモデルの学習を進めることによって、特定のタスクのみに限定されない特徴量を獲得し、汎化性能を向上させる学習手法である。モデルのパラメータ更新時には、各タスクの損失を算出した後、すべてのタスクの損失の重み付き和を用いることによって、主タスク以外のタスクも同時に解くことが可能になるように学習を進めていく。重み付き和を算出する際の重みを大きくすることによって、そのタスクの損失を下げることを重視した学習が可能になる。あるタスクの学習で得た知識を別のタスクに活用するという考え方は転移学習 [4] に類似している。しかし、転移学習ではあるタスクの学習が完了した後に別のタスクの学習を開始するのに対して、マルチタスク学習では一つのモデルの中で同時に複数のタスクの学習を進めていく。そのため、転移学習とマルチタスク学習ではアプローチの方法が異なっている。また、複数のタスクを同時に解くことが可能になるように学習を進めていくため、モデルの出力時には主タスクの結果に加えて補助タスクの結果も同時に出力される点も転移学習と異なっている。マルチタスク学習を適用することによって、主タスクの汎化性能の向上が期待できる、一つのモデルで複数のタスクを解くことが可能となる、学習時間や総パラメータ数の削減ができるなどのメリットが得られる。

マルチタスク学習は画像やテキストなど様々な分野の研究にて適用されている。Lamprinisら [5] は、ニュース記事の見出しからニュース記事の人気予測を行う研究にマルチタスク学習を適用している。この研究ではニュース記事の人気予測という主タスクに加えて、品詞情報のタグ付けタスクとニュース記事のトピック予測という二つの補助タスクも同時に学習することによって予測性能の向上を図っている。Luら [6] は、テキスト要約を行う研究にマルチタスク学習を適用している。この研究ではテキスト要約を行う主タスクに加えて、構文解析タスクとテキストのカテゴリ予測という二つの補助タスクも同時に学習することによって予測性能の向上を図っている。Mulyarら [7] は、自由記述形式の電子カルテから投薬履歴や治療履歴などの様々な情報を抽出する研究にマルチタスク学習を適用している。この研究では、投薬履歴や治療履歴などの抽出タスクをすべて一つのモデルで解くことによって、予測性能の向上と共にモデル構築コストの削減も図っている。この他にも、マルチモーダルモデルによる感情分析にマルチタスク学習を適用している Akhtarら [8] の研究や、係り受け解析を行うモデルにマルチタスク学習を適用している Duongら [9] の研究などが行われている。

る。このようにマルチタスク学習は画像やテキストなど様々な分野の研究にて適用されている。

また、上記のようなマルチタスク学習を適用して推定精度向上を図っている研究以外に、効率良くマルチタスク学習を行うための研究も行われている。Yuら [10] は、効率良くマルチタスク学習を行うためのパラメータ更新時の損失重み付き和の算出方法を提案している。各タスクの勾配が逆方向を向いていた場合に勾配の向きを補正することによって効率良くパラメータを更新していく手法となっている。強化学習モデルにマルチタスク学習を適用した実験において、この手法を用いることによって精度が大幅に改善している。Zhangら [11] は、タスクごとの学習の難易度に着目した学習手法を提案している。マルチタスク学習モデル構築時にタスクごとに Early Stopping を設定して簡単なタスクの学習を途中で打ち切ることによって、簡単なタスクの過学習を防止し、主タスクの収束を促進する手法となっている。顔のランドマーク検出を主タスクとするマルチタスク学習モデルにおいて、この手法を適用することによって検出の精度が向上することが確認されている。Fiftyら [12] は、マルチタスク学習における補助タスクの最適な組合せを効率良く探索する手法を提案している。全タスクを採用したマルチタスク学習モデルの構築時の勾配情報を基に算出したタスク同士の親和性を用いて最適な組合せを決定する手法となっている。この手法は最適な組合せを決定するためにモデルを複数構築する必要がないため、補助タスクの組合せ探索時の大幅なコスト削減になる。Standleyら [13] は、一つの入力に対して解きたいタスクが複数ある場合における最適なタスクの割り当てを行うフレームワークを提案している。この手法では、関連するタスクは同一ネットワークにて、競合するタスクは別のネットワークにて学習が行われる。そのため、構築するモデルの総数は増加する一方で、最適なタスクの組合せによるマルチタスク学習モデルが構築できるため各タスクの推定精度向上が期待できる。

上記のようにマルチタスク学習に関して様々な研究が行われているが、採用する補助タスクについては主タスクと関連していると思われるタスクを感覚的に選定している研究が多く、明確な基準に従った選定が行われているわけではない。また、採用する補助タスクの数について言及している研究や補助タスクの明確な選定基準を示している研究は調べた限りでは行われていない。そのため、マルチタスク学習モデルを構築する際にどのような補助タスクをどの程度採用すべきなのかは明らかになっていない。そこで本稿では、マルチタスク学習モデルの構築時にどのような補助タスクをどの程度採用すると推定精度向上に効果的なのかを調査するために5つの補助タスクを用意して実験を行った。

3 構築モデル

本稿では、商品タイトルの興味を惹く度合い推定を主タスクとしたマルチタスク学習モデルを構築して実験を行う。モデル構築の際には主タスクの学習に加えて補助タスクの学習も同時に行う。モデル構築に使用するデータセットについてを3.1節

表 1 データセット内の各補助タスクのラベルの内訳

タスク	票数										総数
	0	1	2	3	4	5	6	7	8	9	
Sub1	3	5	10	10	35	64	144	525	2,175	7,616	10,587
Sub2	8,070	1,923	422	120	37	11	4	0	0	0	10,587
Sub3	8,227	1,692	329	121	105	57	37	9	9	1	10,587
Sub4	7,539	1,664	481	315	224	153	113	72	22	4	10,587
Sub5	62	3,249	2,089	1,431	1,237	1,027	821	500	147	24	10,587

にて、用意した補助タスクについてを 3.2 節にて、構築するモデルの概要についてを 3.3 節にて述べる。

3.1 使用データセット

本稿では、国立情報学研究所から提供されている楽天データセット¹を使用した。このデータセットは楽天市場に出品されている商品データ約 2 億 8,300 万件によって構成されている。各商品データには、商品タイトル、商品画像、商品説明文、商品カテゴリなどが含まれている。

上記データセットのうち、食品カテゴリに属する商品のタイトルを用いたデータセットを作成して実験を行った。本稿で構築するのは、商品タイトルの興味を惹く度合い推定を行うモデルである。そのため、商品画像などの見た目に関わる要素が興味を惹くかどうか大きく影響する商品データを使用するのは不適切であると考えた。そこで、見た目の影響が小さい食品カテゴリに属する商品データを使用することにした。

3.2 補助タスク

本稿では、商品タイトルの興味を惹く度合い推定を行う主タスクに加えて、5つの補助タスクを用意して実験を行う。先行研究では、主タスクと関連していると思われるタスクを感覚的に選定して補助タスクとする研究が多い。そのため、本稿においても同様の選定方法にて補助タスクを用意した。本稿にて構築するマルチタスク学習モデルの主タスクが商品タイトルの興味を惹く度合い推定であるため、興味を惹く要因となり得る要素についてを推定するタスクである下記 5つを補助タスクとして用意することにした。

- Sub1 販売商品把握の可否の推定
- Sub2 呼びかけ表現の有無の推定
- Sub3 希少性/限定性を示す表現の有無の推定
- Sub4 評価/実績の記載の有無の推定
- Sub5 商品の特徴/アピールポイント記載の有無の推定

本稿にて用意した 5つの補助タスクは、後述のアンケートにおける投票人数を推定する回帰タスクとしている。

用意した商品タイトルに対してアンケートを取り、主タスクのラベルに加えて上記の各補助タスクに対応するラベルも付与する。商品タイトルを読んだ際に興味を惹かれるかどうかの基準は人によって様々であるため、一人の意見を用いて作成したデータセットでは汎化性能が低くなってしまう。そのため、商

品タイトル一つに対して複数人にアンケートを取りデータセットを作成していく。アンケートでは、作業者に商品タイトルを一つずつ提示して、以下に示す主タスク用と補助タスク用の合計 6つの質問をする。

- i. 主タスク用の質問
 - ・ Main その商品について「詳しく知りたい」または「買ってみようかな」と感じますか?
- ii. 補助タスク用の質問
 - ・ Sub1 何を販売しているか理解/把握できますか?
 - ・ Sub2 閲覧者へ呼びかける表現が含まれていますか?
 - ・ Sub3 商品の希少性/限定性(数量限定, 希少部位など)について示す表現が含まれていますか?
 - ・ Sub4 商品の評価/実績(〇〇賞獲得, ランキング 1 位など)について示す表現が含まれていますか?
 - ・ Sub5 商品の特徴やアピールポイントについて示す表現が含まれていますか?

上記の質問はすべて'Yes'か'No'の二択で回答する形式となっており、各質問に対して'Yes'と回答した人数を集計して、その人数を各タスクのラベルとして付与する。そのため、本稿で作成したデータセットには、6種類のタスクそれぞれに対応したラベルが商品タイトル一つに対して付与されている。なお、主タスクの学習に用いる質問の集計に関しては 3 票以上のラベルを同列に扱う。商品タイトルを読んで興味を惹かれるかどうかの基準は人によって様々であり、読んだ人全員の興味を惹く商品タイトルを作成することは困難であると思われる。そのため、読んだ人のおよそ 30%の興味を惹くことができる商品タイトルであれば十分であると考えられる。そして、およそ 30%の人の興味を惹くことができるかどうかモデルの推定によって分かるだけでも有用性はあるのではないかと考えたため、3 票以上のラベルを同列に扱うことにした。

本稿では、上記手順にて商品タイトル 10,587 件に対してラベルを付与してデータセットを作成した。その際に、商品タイトル一つにつき 9 人に対してアンケートを実施した。その結果、主タスクである興味を惹く度合い推定に用いるラベルの内訳は、0 票が 6,952 件、1 票が 2,616 件、2 票が 746 件、3 票以上が 272 件となった。また、各補助タスクに対応するラベルの内訳は表 1 のようになった。表 1 の補助タスクの番号は 3.2 節にて述べた補助タスクの番号と対応している。作成したデータセットを使用して、興味を惹く度合い推定を主タスクとするマルチタスク学習モデルを構築する。

1: 楽天グループ株式会社 (2020): 楽天市場データ. 国立情報学研究所情報学研究データリポジトリ. (データセット). <https://doi.org/10.32130/idr.2.1>

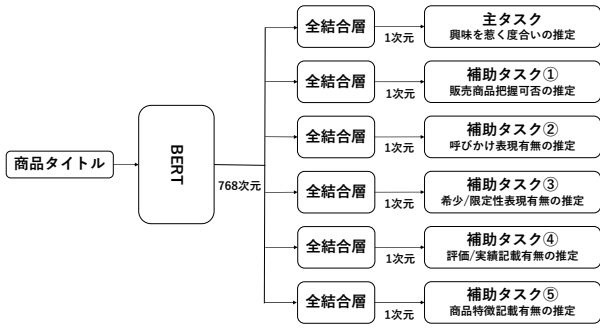


図 1 構築するモデルの概要

3.3 モデル概要

本稿では、BERT を用いた興味を惹く度合い推定を行うシングルタスク学習モデルをベースにしてマルチタスク学習モデルを構築する。BERT は、Transformer [14] の Encoder を使用したニューラルネットワークモデルであり、モデルの構造を修正せずとも転移学習することによって、様々な自然言語処理のタスクに応用できるモデルとなっている。転移学習前の事前学習として MLM (Masked Language Modeling) と NSP (Next Sentence Prediction) が長い文章を含むデータセットを用いて行われている。これら二種類の事前学習にて獲得したネットワークのパラメータを別のタスク用にファインチューニングすることによって、高い推定精度を発揮することが期待できるモデルである。

構築するモデルの概要を図 1 に示す。まず、入力として与えられた商品タイトルを BERT モデルへ入力する。その出力となる 768 次元のベクトルを各タスクへ分岐させる。分岐する数は補助タスクの採用数によって異なり、図 1 は用意した補助タスク 5 つをすべて採用した場合のモデル構造となっている。分岐した後、768 次元のベクトルは各タスクに応じた全結合層に入力され、それぞれ 1 次元の数値が出力される。そのため、図 1 のモデルの場合は商品タイトルを一つ入力すると 6 つのタスクの推定結果が同時に出力されることになる。なお、本稿にて用意したタスクは主タスクも含めてすべて回帰タスクとなっている。上記のようなマルチタスク学習モデルを 3.2 節に従って作成したデータセットを用いて構築する。

本稿では、訓練済み BERT モデルとして、東北大学の乾・鈴木研究室の訓練済み日本語 BERT モデル²を使用した。訓練済みモデルは日本語版 Wikipedia にて事前学習が行われており、語彙数は 32,000 となっている。訓練時には BERT モデルの最終層以外のパラメータは更新しないように設定してファインチューニングを行った。そのため、訓練時に更新されるのは BERT モデルの最終層と各タスクに応じた全結合層のパラメータのみとなっている。

また、マルチタスク学習時のパラメータ更新には採用したタ

スクの損失すべての重み付き和を用いる。重み付き和を算出する際の重みを大きくすることによって、そのタスクの損失を下げることを重視した学習が可能になる。マルチタスク学習を適用する目的は主タスクの損失を下げることであるため、更新時には主タスクの損失を下げることを重視したい。そこで、本稿ではマルチタスク学習時のパラメータ更新に用いる主タスクと補助タスクの損失の比率が 6 : 4 になるように設定をした。そのため、パラメータ更新に用いる損失すべての重み付き和の算出式を式 1, 2 の通りに設定をしている。

$$L_{Total} = L_{Main} + \frac{2}{3N} L_{ST} \quad (1)$$

$$L_{ST} = L_{Sub1} + L_{Sub2} + \dots + L_{SubN} \quad (2)$$

上式において、パラメータ更新に用いる損失すべての重み付き和を L_{Total} 、主タスクの損失を L_{Main} 、補助タスクの損失の合計値を L_{ST} としている。また、 $L_{Sub1}, L_{Sub2}, \dots, L_{SubN}$ はそれぞれの補助タスクにおける損失となっている。補助タスクの損失の合計値である L_{ST} を $\frac{2}{3N}$ 倍することによって、主タスクと補助タスクの損失の比率が 6 : 4 になるように調整している。ここで、 N はモデル構築時に採用した補助タスクの総数であり、本稿では $1 \leq N \leq 5$ の範囲の整数値となる。上式のように計算した L_{Total} を用いて学習時のパラメータ更新を行い、モデルを構築する。

4 実 験

本稿にて実験を行う目的は二つある。一つ目は、マルチタスク学習モデル構築時に採用する補助タスクの数によってモデルの推定精度が変動するかどうかを調査することである。二つ目は、マルチタスク学習モデル構築時にどのような補助タスクを採用すると推定精度向上に効果的なのかを調査することである。

4.1 実験 1: 有効な補助タスクの特徴と採用数についての分析

本実験では、用意した補助タスク全通りの組合せによるマルチタスク学習モデルを構築してテスト用データの推定精度を比較・分析することによって、補助タスクの採用数による推定精度の変動や推定精度向上に効果的な補助タスクの特徴についてを調査する。

4.1.1 実験手順

まず、3.2 節にて作成したデータセットを用いて、3.3 節にて述べたマルチタスク学習モデルを構築する。本稿では、3.2 節にて述べた 5 つの補助タスク全通りの組合せである 32 種類のマルチタスク学習モデルを構築する。モデル構築に用いるデータセットのうち、8 割を訓練用データ、1 割を評価用データ、1 割をテスト用データとして扱い、訓練用データに対しては同じデータを複製することによって主タスクのラベル間の偏りを解消した。ここで分割した訓練用データ、評価用データ、テスト用データを用いて、すべてのモデルを構築する。そのため、どのような補助タスクの組合せのモデルを構築する場合であっても訓練用データ、評価用データ、テスト用データの中身は同一と

2 : <https://github.com/cl-tohoku/bert-japanese>

なっている。学習時の最大エポック数は 10,000 として、主タスクの Validation Loss が 100 エポック改善しない場合は Early Stopping を適用して学習を停止するように設定をした。そのため、最終的なモデルのパラメータは主タスクの Validation Loss が最も低かった場合のものとなる。また、BERT モデルのパラメータ更新設定やパラメータ更新に用いる損失すべての重み付き和の設定は 3.3 節にて述べた通りである。

補助タスクの組合せ一つに対して 10 個ずつモデルを構築して、テスト用データ推定時の主タスク RMSE の平均値を算出する。また、補助タスクを全く採用しないシングルタスク学習モデルの RMSE 平均値との間に差があるか否かの検定を行い、 p 値を算出する。その際には、帰無仮説を「構築したマルチタスク学習モデルの主タスク RMSE の平均値とシングルタスク学習モデルの主タスク RMSE の平均値との間に差はない」として対応のない 2 標本 t 検定を行った。そして、得られた各組合せによる結果を比較・分析することによって、補助タスクの採用数による推定精度の変動や推定精度向上に効果的な補助タスクの特徴について調査し、マルチタスク学習モデルにおける補助タスクの選定方法について考察する。

4.1.2 結果・考察

補助タスク各組合せによる構築モデルの主タスク RMSE の平均値、決定係数平均値、検定時の p 値を表 2 に示す。表 2 に記載されている補助タスクの番号は 3.2 節にて述べた補助タスクと対応している。表 2 に示した補助タスクの組合せ全 32 通りの結果を基に、補助タスクの採用数による推定精度の変動や推定精度向上に効果的な補助タスクの特徴についてを分析する。

a) 補助タスク採用数についての分析

表 2 にて補助タスクの採用数に注目すると、タスク同士の組合せによって多少の差はあるが採用数が多いモデルであるほど主タスク RMSE の平均値や p 値が小さくなる傾向がありそうなが読み取れる。補助タスクの採用数と RMSE 平均値の関係について分析した結果を表 3 に示す。ここでは、補助タスクの採用数ごとに表 2 中の主タスク RMSE の平均値を算出している。そして、補助タスクの採用数と採用数ごとの RMSE 平均値の相関係数を算出している。表 3 を見ると、補助タスクの採用数と採用数ごとの RMSE 平均値の相関係数は -0.7847 となっており、この二つの間には強い相関が見られることが分かった。この結果より、補助タスクの採用数を増やすことによって推定精度向上が期待できると言えるのではないかと考える。

本実験にて、マルチタスク学習モデル構築時に選定する補助タスクの採用数を増やすことによって推定精度向上が期待できるという知見が得られた。補助タスクの採用数を増やすことによって推定精度向上が期待できる一方で、データ作成時にかかるコストも増加してしまう。そのため、補助タスクの採用数を増やす際にはデータ作成時にかかるコストも見極めることが重要になると考える。

b) 精度向上に効果的な補助タスクの特徴についての分析

表 2 にて、採用しているタスクに注目すると Sub5 を補助タスクとして採用しているモデルの主タスク RMSE の平均値や p 値が他のモデルよりも小さくなっていることが読み取れる。

各補助タスクの効果の大きさと主タスクとの関係について分析した結果を表 4 に示す。ここでは、まず表 2 を基に各補助タスクを採用した場合の主タスク RMSE の平均値と採用しなかった場合の主タスク RMSE の平均値を算出し、その差を取った。算出した RMSE 平均値の差が大きいタスクほど、採用した場合の効果の大きいタスクであると言える。その結果より、Sub5 のタスクが最も効果が大きいこと、Sub3 のタスクが最も効果が小さいこと、Sub1,2,4 のタスクが同程度の効果を発揮していることが分かる。

効果が大きいタスクと小さいタスクの違いを調査するために、使用したデータセット上での主タスクのラベルと各補助タスクのラベルの相関係数の算出を行った。また、採用した補助タスクがシングルタスク学習モデルにてどの程度学習可能なかを測定するため、補助タスクのみを解くシングルタスク学習モデルを構築してテスト用データ推定時の決定係数を算出した。このシングルタスク学習モデルはマルチタスク学習モデル構築時と同じ訓練用データ、評価用データ、テスト用データを使用しており、学習時の最大エポック数は 10,000、Validation Loss が 100 エポック改善しない場合は Early Stopping を適用して学習を停止するように設定をして構築した。表 4 の決定係数の値はモデル 10 個分の平均値となっている。そして、RMSE 平均値の差 (表 4 の①) とデータセット上での主タスクとの相関係数の値 (表 4 の②) の相関係数を算出した。算出した相関係数は -0.6822 となり、強い相関があるとは言いがたい結果となった。しかし、補助タスクのみを解くシングルタスク学習モデルの推定精度が低い Sub1 と Sub2 を除外して同様の相関係数を算出したところ -0.9311 となり、強い相関が見られた。この結果から、シングルタスク学習において十分な推定精度を発揮できるタスクであるという前提の下であれば、主タスクとの相関が強いタスクほどマルチタスク学習時の補助タスクとして推定精度向上に有効と言えるのではないかと考える。

本実験にて、シングルタスク学習において十分な推定精度を発揮できるタスクであるという前提の下であれば主タスクとの相関が強い補助タスクほど推定精度向上に有効であるという知見が得られた。しかし、本稿では一つのデータセットを使用した実験しか行っていない。そのため、別のデータセットにおいても同様の結果が得られるかどうかを確かめるために、別のデータセットを使用した同様の実験を行う必要がある。また、本実験にて構築したマルチタスク学習モデルに採用しているタスクはすべて回帰タスクとなっている。本実験にて得られた知見である主タスクとの相関の強さを基にした補助タスク選定方法が、分類タスクを解くマルチタスク学習モデルにおいても推定精度向上に有効であるかどうかは不明である。そのため、分類タスクを解くモデルの場合においても同様の実験を行い、上記について確かめる必要がある。

4.2 実験 2: 補助タスクがノイズとして扱われているかの検証

表 4 において、シングルタスク学習モデルの推定精度が共に低い Sub1 と Sub2 は主タスクとの相関に差があるにもかかわらず同程度の推定精度向上効果を発揮している。この結果から、

表 2 構築したモデルの主タスク RMSE の平均値, 決定係数平均値, p 値

構築したモデル	RMSE(Main)	決定係数 (R^2)						p -value
		Main	Sub1	Sub2	Sub3	Sub4	Sub5	
Single-Task	0.8684	-0.4328	-	-	-	-	-	-
Multi-Task(Sub1)	0.7262	0.0790	0.1659	-	-	-	-	0.1185
Multi-Task(Sub2)	0.7278	0.0755	-	0.3242	-	-	-	0.1217
Multi-Task(Sub3)	0.7330	0.0622	-	-	0.5843	-	-	0.1353
Multi-Task(Sub4)	0.7228	0.0878	-	-	-	0.5856	-	0.1102
Multi-Task(Sub5)	0.6815	0.1890	-	-	-	-	0.5688	0.0448
Multi-Task(Sub1,2)	0.7094	0.1214	0.0395	0.2667	-	-	-	0.0830
Multi-Task(Sub1,3)	0.7293	0.0714	0.0239	-	0.5754	-	-	0.1259
Multi-Task(Sub1,4)	0.7058	0.1302	0.0574	-	-	0.5485	-	0.0768
Multi-Task(Sub1,5)	0.6912	0.1654	0.0451	-	-	-	0.5343	0.0562
Multi-Task(Sub2,3)	0.7166	0.1034	-	0.2842	0.5261	-	-	0.0969
Multi-Task(Sub2,4)	0.7089	0.1225	-	0.2620	-	0.5392	-	0.0823
Multi-Task(Sub2,5)	0.7022	0.1397	-	0.2464	-	-	0.4970	0.0705
Multi-Task(Sub3,4)	0.7196	0.0960	-	-	0.5464	0.5586	-	0.1029
Multi-Task(Sub3,5)	0.6897	0.1698	-	-	0.5210	-	0.5100	0.0535
Multi-Task(Sub4,5)	0.7088	0.1221	-	-	-	0.5433	0.4765	0.0829
Multi-Task(Sub1,2,3)	0.6969	0.1520	0.0355	0.1987	0.5161	-	-	0.0633
Multi-Task(Sub1,2,4)	0.7036	0.1362	0.0238	0.2223	-	0.5347	-	0.0728
Multi-Task(Sub1,2,5)	0.6869	0.1766	0.0527	0.2320	-	-	0.4939	0.0502
Multi-Task(Sub1,3,4)	0.7045	0.1334	-0.0075	-	0.5422	0.5376	-	0.0748
Multi-Task(Sub1,3,5)	0.7127	0.1086	0.0298	-	0.5350	-	0.4906	0.0949
Multi-Task(Sub1,4,5)	0.7095	0.1641	0.0332	-	-	0.5208	0.4826	0.0662
Multi-Task(Sub2,3,4)	0.7095	0.1210	-	0.2132	0.4922	0.5062	-	0.0834
Multi-Task(Sub2,3,5)	0.6965	0.1533	-	0.2290	0.4955	-	0.4607	0.0625
Multi-Task(Sub2,4,5)	0.6915	0.1653	-	0.2009	-	0.5026	0.4512	0.0559
Multi-Task(Sub3,4,5)	0.6891	0.1710	-	-	0.5053	0.4946	0.4363	0.0531
Multi-Task(Sub1,2,3,4)	0.6998	0.1443	0.0010	0.2054	0.5114	0.5221	-	0.0681
Multi-Task(Sub1,2,3,5)	0.6950	0.1560	0.0163	0.1659	0.5016	-	0.4705	0.0613
Multi-Task(Sub1,2,4,5)	0.6996	0.1454	0.0205	0.1767	-	0.5183	0.4544	0.0671
Multi-Task(Sub1,3,4,5)	0.6849	0.1810	0.0322	-	0.4996	0.4978	0.4612	0.0483
Multi-Task(Sub2,3,4,5)	0.6874	0.1751	-	0.1791	0.5065	0.4974	0.4275	0.0511
Multi-Task(Sub1,2,3,4,5)	0.6790	0.1948	0.0072	0.1031	0.5007	0.4873	0.4288	0.0424

シングルタスク学習モデルの推定精度が低い補助タスクは学習時にノイズのようには扱われていないのではないかと考えられる。そこで、ノイズとして扱うタスクを補助タスクとして採用したマルチタスク学習モデルを構築して推定精度を比較する追加実験を行うことによって上記について検証した。

4.2.1 実験手順

ノイズとして扱うタスクとして、0 から 3 の範囲のランダムな数値を予測する回帰タスクを設定した。このタスクを補助タスクとして採用したマルチタスク学習モデルを構築する。モデル構築時には、4.1 節のモデル構築時と同じ訓練用データ、評価用データ、テスト用データを使用しており、学習時の最大エポック数は 10,000、Validation Loss が 100 エポック改善しない場合は Early Stopping を適用して学習を停止するように設

定した。また、BERT モデルのパラメータ更新設定やパラメータ更新に用いる損失すべての重み付き和の設定は 3.3 節にて述べた通りである。上記の設定でモデルを 10 個構築してテスト用データ推定時の主タスク RMSE の平均値を算出する。

モデル構築後、Sub1 を採用した場合の RMSE 平均値、Sub2 を採用した場合の RMSE 平均値との間に差があるか否かの検定を行い、 p 値を算出する。その際には、帰無仮説を「Sub1 を採用した場合の主タスク RMSE の平均値とノイズとして扱うタスクを補助タスクとして採用した場合の主タスク RMSE の平均値との間に差はない」、「Sub2 を採用した場合の主タスク RMSE の平均値とノイズとして扱うタスクを補助タスクとして採用した場合の主タスク RMSE の平均値との間に差はない」として、それぞれ対応のない 2 標本 t 検定を行った。この検定を行うことによって、シングルタスク学習モデルの推定精度が低い補助タスクは学習時にノイズのようには扱われているのか否かについて検証する。

表 3 補助タスク採用数と主タスク RMSE の平均値の関係

採用数	0	1	2	3	4	5
RMSE	0.8684	0.7183	0.7081	0.6983	0.6934	0.6790
相関係数	-0.7847					

表 4 各補助タスクの効果の大きさと主タスクとの関係

		Sub1	Sub2	Sub3	Sub4	Sub5
タスク採用/不採用時の主タスク RMSE の平均値	<i>In</i>	0.7010	0.7007	0.7027	0.7004	0.6930
	<i>Out</i>	0.7162	0.7160	0.7140	0.7164	0.7239
① RMSE 平均値の差 (<i>In</i> - <i>Out</i>)		-0.0153	-0.0153	-0.0112	-0.0160	-0.0388
② データセット上での主タスクとの相関係数		0.1205	0.2944	0.2810	0.3951	0.4824
③ シングルタスク学習時 決定係数 (R^2)		0.2380	0.4734	0.6560	0.7076	0.6512
① と ② の相関係数		-0.6822				
① と ② の相関係数 (Sub3,4,5 のみ)		-0.9311				

表 5 ノイズとして扱う補助タスクを採用したモデルの主タスク RMSE の平均値, 決定係数平均値, Sub1 採用時と Sub2 採用時それぞれとの検定による p 値

採用した補助タスク	RMSE(Main)	決定係数 (R^2)		p -value	
		Main	Noise	Sub1 <i>In</i>	Sub2 <i>In</i>
Multi-Task(Noise)	0.7073	0.1262	-0.4201	0.4793	0.3565

4.2.2 結果・考察

ノイズとして扱う補助タスクを採用したマルチタスク学習モデルにおけるテストデータ推定時の主タスク RMSE の平均値, 決定係数平均値, 検定時の p 値を表 5 に示す. 表 5 の結果を見ると, 主タスク RMSE の平均値は 0.7073 となっている. 表 4 上段のタスク採用時の主タスク RMSE の平均値を見ると, Sub1 は 0.7010, Sub2 は 0.7007 となっており, ノイズとして扱うタスクを採用した場合とほぼ同等の推定精度となっている. そして, p 値においては, Sub1 採用時との検定では 0.4793, Sub2 採用時との検定では 0.3565 というれも有意な差があるとは言えない結果となっている. 以上から, Sub1 と Sub2 の補助タスクが推定精度向上に効果的に作用しているとは言い難く, ノイズのように扱われてパラメータが適度に均されたことによって, 結果的に主タスクの過学習を抑制しているだけである可能性が否めない.

上記より, シングルタスク学習モデルの推定精度が低いタスクであると, どのようなタスクを補助タスクとして採用したとしてもノイズとして扱われてしまうため, 推定精度向上には同程度の効果しか発揮できないのではないかと考える. また, シングルタスク学習において十分な推定精度を発揮できるタスクである Sub4 であっても補助タスクとしては Sub1,2 と同程度の効果しか発揮できておらず, Sub3 に関しては Sub1,2 よりも小さな効果しか発揮できていない. この結果から, シングルタスク学習において十分な推定精度を発揮できるタスクであっても, 補助タスクとして採用した際にノイズのように扱われるタスクと同程度の効果を発揮するには Sub4 程度の主タスクとの相関の強さが必要になるのではないかと考える. そのため, シングルタスク学習において十分な推定精度を発揮できるが主タスクとの相関が低いタスクを採用するのであれば, ノイズとして扱われてしまうタスクを採用してパラメータを適度に均すことを選択した方が高い推定精度を期待できるのではないかと考える.

5 おわりに

本稿では, マルチタスク学習モデル構築時にどのような補助タスクをどの程度採用すると推定精度向上に効果的なのかについて調査を行った. マルチタスク学習に関する研究は, 効率良く学習を進めるための手法の提案や最適なタスクの組合せを効率良く探索する手法の提案など様々行われている. しかし, 補助タスクの選定に関しては感覚的に選定している研究が多く, 採用する補助タスクの最適な数や選定基準については明らかになっていない. そこで, 採用する補助タスクの最適な数や選定基準について調査するために, 商品タイトルの興味を惹く度合い推定を主タスクとしたマルチタスク学習モデルを構築して実験を行った. 実験では, 主タスクである興味を惹く度合い推定に加えて 5 つの補助タスクを用意した. 複数人にアンケートを取ることによって作成したデータセットを使用して, 5 つの補助タスク全通りの組合せである 32 種類のマルチタスク学習モデルを構築した. その後, 構築したモデルごとのテスト用データ推定精度を比較・分析することによって, 補助タスクの採用数による推定精度の変動や推定精度向上に効果的な補助タスクの特徴について調査した.

補助タスクの採用数についての分析においては, 補助タスクの採用数ごとに主タスク RMSE の平均値を算出して, 補助タスクの採用数を増やすことによる推定精度の変動を調査した. 補助タスクの採用数と採用数ごとの RMSE 平均値の相関係数を算出したところ -0.7847 となり, これら二つの間に強い相関が見られた. 推定精度向上に効果的な補助タスクの特徴についての分析においては, 各補助タスクを採用した場合の主タスク RMSE の平均値と採用しなかった場合の主タスク RMSE の平均値の差を算出して, 各補助タスクの推定精度向上効果と効果の大きい補助タスクの特徴を調査した. 補助タスク用データと主タスク用データの相関係数の値と補助タスクを採用した際の推定精度向上効果の相関係数を算出したところ -0.6822 となり, 強い相関があるとは言い難い結果であった. しかし, シングルタスク学習において十分な推定精度を発揮できないタスクを除

外して同様の相関係数を算出したところ-0.9311 と強い相関が見られた。それぞれの分析結果から、補助タスクの採用数を増やしたモデルであるほど推定精度が高くなる傾向があること、シングルタスク学習において十分な推定精度を発揮できて主タスクとの相関が強いタスクを採用することによって推定精度が高くなる傾向があることを確認した。また、シングルタスク学習において十分な推定精度を発揮できないタスクであるとノイズとして扱われてしまうため、どのようなタスクを補助タスクとして採用した場合でも同程度の効果しか発揮できないという傾向も見られた。

本実験を通して、マルチタスク学習モデル構築時に選定する補助タスクの採用数を増やすことによって推定精度向上が期待できること、シングルタスク学習において十分な推定精度を発揮できるタスクであるという前提の下であれば、主タスクとの相関が強い補助タスクほど推定精度向上に有効であるという知見が得られた。この知見は、マルチタスク学習モデル構築時における補助タスク選定方法の一基準となり得るものであると考える。

しかし、補助タスクの採用数を増加させることで推定精度向上が見込める一方でデータ作成時にかかるコストも増加してしまう。そのため、補助タスクの採用数を増やす際にはデータ作成時にかかるコストも見極めることが重要になると考える。また、本稿では一つのデータセットを使用した実験しか行っておらず、すべてのタスクが回帰タスクとなっている。本実験にて得られた知見である主タスクとの相関の強さを基にした補助タスク選定方法が、分類タスクを解くマルチタスク学習モデルにおいても推定精度向上に有効であるかどうかは不明である。そのため、別のデータセットを使用した同様の実験、分類タスクを解くマルチタスク学習モデルを使用した実験を行うことによって上記についてを確かめる必要がある。

謝辞 本研究の一部は JSPS 科研費 19H04218 および越山科学技術振興財団の助成を受けたものです。本研究では、国立情報学研究所の IDR データセット提供サービスにより楽天グループ株式会社から提供を受けた「楽天データセット」(https://rit.rakuten.com/data_release/) を利用した。

文 献

- [1] 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, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [2] Rich Caruana. Multitask learning. *Machine learning*, Vol. 28, No. 1, pp. 41–75, 1997.
- [3] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [4] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242–264. IGI global, 2010.
- [5] Sotiris Lamprinidis, Daniel Hardt, and Dirk Hovy. Predicting news headline popularity with syntactic and semantic knowledge using multi-task learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 659–664, 2018.
- [6] Yao Lu, Linqing Liu, Zhile Jiang, Min Yang, and Randy Goebel. A multi-task learning framework for abstractive text summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 9987–9988, 2019.
- [7] Andriy Mulyar, Ozlem Uzuner, and Bridget McInnes. Mtclicbert: scaling clinical information extraction with multitask learning. *Journal of the American Medical Informatics Association*, Vol. 28, No. 10, pp. 2108–2115, 2021.
- [8] Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbali, and Pushpak Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 370–379, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [9] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)*, pp. 845–850, 2015.
- [10] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, Vol. 33, pp. 5824–5836, 2020.
- [11] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pp. 94–108. Springer, 2014.
- [12] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, Vol. 34, pp. 27503–27516, 2021.
- [13] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pp. 9120–9132. PMLR, 2020.
- [14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

- [1] 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, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [2] Rich Caruana. Multitask learning. *Machine learning*, Vol. 28, No. 1, pp. 41–75, 1997.
- [3] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [4] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242–264.