Exploring Subnetworks for Efficient Neural Model Training

	こばやし　そうすけ
氏　　名	小　林　颯　介
研究科、専攻	東北大学大学院情報科学研究科（博士課程）システム情報科学専攻
学位論文題目	Exploring Subnetworks for Efficient Neural Model Training （ニューラルモデルの効率的訓練のためのサブネットワーク探索）
論文審査委員	（主査）東北大学教　授　乾 健太郎 東北大学教　授　岡谷 貴之　　東北大学教　授　大町 真一郎 東北大学教　授　鈴木 潤

要約

第１章　序論

機械学習は訓練データから予測システムを作るための情報処理パラダイムとして広く使われるようになった特にニューラルネットワークによる深層学習モデルが、画像、音声、言語を始めとした多様なドメインのタスにおいて高い性能を発揮できることが発見され、大きな発展を遂げた。深層学習モデルは、表現力が非常に高く複雑なタスクや大規模なデータセットから効果的に学習が行えるほか、モデルサイズを大きくした際にも高い化能力を持つことが明らかになってきた。また、データサイズ、モデルサイズ、訓練時間の三要素からなるベ則によって性能向上が支配されている可能性について実験的に示唆されている。このような知見はシステム工的にも性能向上の余地の検討やそのコスト見積もりの観点から非常に有用である一方で、その上限が未だに見ておらず日進月歩でコストと最高性能の増加が続いている。最もボトルネックたりうるデータサイズについても、近年さらなる発展を見せている教師なし学習によってそのべき則が支えられている。例えば、最も成功している事前学習として、自然言語処理における言語モデル的学習が挙げられる。そのような教師なしコーパスの活用手段が見つかった結果、これまで以上に深層学習モデルの訓練にかかるコストが増大し、今後もべき則に従うのであればその増加は免れられないと考えられる。その訓練コストの増大をよそに、機械学習には古くから複数のモデルを構築した上での方法論や応用が存在してきた。例えば、複数モデルのアンサンブルを実施したり、モデルの結果を比較しデータセットや訓練設定の影響を探索したり、あるいは、入力や運用状況に応じて特徴的なモデルを切り替えたりすることも実用的なシステムの構築としては非常に重要であった。しかし、一つのモデルの構築すら困難な状況下においては、このような複数モデルでの方法論は非常に扱いづらくなりつつある。本論文では、大量のモデルを少ないコストで構築する新たな手段として、ニューラルネットワーク内に潜在的に生まれるサブネットワーク群について着目し、その活用と可能性について解明する。

第２章　ニューラルネットワークとサブネットワーク

はじめにニューラルネットワークとその学習の基礎を概観する。その後、その訓練コストがいかにして増大し

てきたか、そして今後も増大していくという見通しを、特に近年盛んな事前学習を中心にして論じる。本論文で注目するサブネットワークについても、枝刈りによるモデル圧縮、宝くじ仮説、dropout などの代表的な先行研究と周辺分野について関連付けてまとめる。特に、本論文とは異なるモチベーションのもとに実施されてきたそれらの先行研究の中から、共通して昇華できるサブネットワークに関する知見と応用例をまとめる。

第３章　特定訓練事例の除外サブネットワークによる訓練事例の影響の推定

機械学習モデルの性能および特性を決める最重要要素の一つは訓練データである。ニューラルネットワークにおいては、膨大なデータを用いた複雑な学習過程の末に複雑な関数が推定されるため、ブラックボックス性が問題点として指摘されている。信頼し改善できるようなシステムとしてモデルを構築するためには挙動と性質をよく理解する必要がある。解析のための一方針として、影響関数に代表されるような「ある一つの訓練データを除いたらどうなっていたか？」という反実仮想的なケースを考慮する方法が存在する。一方で、数理的にはニューラルネットワークはその複雑性により直接的な解析は行えず、また、実際に一つの訓練データをそれぞれ除外してモデルを訓練して比較するという総当り的な方法も計算量的に困難である。そのため、何らかの方法で近似的な解や値を求める必要がある。これまでには、解析的な方法の近似を計算する方法が存在したが、主に計算量的な課題が依然として大きく残っていた。本章では、総当り方式の近似という立場で新たな方法を提案する。具体的には、訓練データセット内全てのデータに対して「ある一つの訓練データを除いたら」を実現したようなサブネットワークの集合を、計算量的に現実的な一度のモデル訓練および保存で実現し、推論も軽量に抑える方法を提案する。背景として、これまでに dropout と呼ばれる、訓練時にランダムな重みを枝刈りしながら学習を行う方法があり、これを発展させる。枝刈りを行われた重みはその学習ステップでは一切更新が行われることがない。これを踏まえて、提案手法では dropout の枝刈りマスクを訓練事例ごとに固有に用意し紐付ける方法をとる。これによりある事例で学習する場合においては常に同じ枝刈りが行われ、すなわち、その枝刈られた重みは一度もその事例での学習ステップを経ないことになる。そのため、文字通り裏を返せば、そのマスクを反転したような枝刈りを行えば、そのサブネットワークは「ある一つの訓練データを除いたら」のサブネットワークとして振る舞うモデルになっている。この方法を素朴に検討した場合、２つの計算量的問題点があるが、それらも工夫により解決できる。一つ目は、枝刈りマスクはランダムベクトルかつ固定であるため、実は事例ごとに固定のランダムシードで都度ランダムベクトルを生成すれば、決定的な事例ごとのマスクとして使えるため、保存のコストは一切かからない、というアイデアである。二つ目は、マスクベクトルを少ない有限個の基底となるランダムマスクのベクトルから構成的に構築することで、メモリ消費と速度を抑える方法を提案した。基底となるランダムベクトルの分布を適切に設定することで、構成後には所望の分布のマスクベクトルが得られるほか、構成の組み合わせはハッシュ関数によってランダムに決定することで、異なる事例のマスクが一致する確率も十分に小さくできる。これらの工夫によって軽量なまま実施可能となった提案手法を用いて、ある事例を見ていないサブネット

ワークと見ているサブネットワークの出力の差をスコアとして解析に用いる実験を行った。まずは第一の実験として学習時の損失関数の曲線を観察した。見ているサブネットワークは通常通り過学習の傾向を示した一方で、見ていないサブネットワークについては訓練損失が途中で早めに下げ止まり、訓練に未使用の検証用データにおける損失と似たような曲線を描き、提案手法が理論面だけでなく実験的にも機能していることが確認できた。応用実験として、誤り事例への予測根拠事例の提示とデータフィルタリングを行った。画像物体認識においては予測根拠事例の提示により、形状・シルエット・色合いなどが酷似した悪影響のある訓練事例が特定され、改善のためのモデルやデータに関する知識を得ることができた。文書分類についても、悪影響のある訓練事例とテスト事例の共通点から、特定の単語のような局所的な情報に過学習している可能性を発見できた。データフィルタリングにおいては、訓練データとテストデータの分布が異なるドメイン適応の設定で、少量の検証用データ上での損失への影響を計算し、悪影響を及ぼす訓練データを一定量削除することで、再訓練したモデルの性能が向上することを確認した。さらなる新たな応用として、訓練事例そのものによってその事例自身の予測がどれほど改善するか、という指標を提案し、言語モデルタスクにおける過学習が発生しやすい事例（文脈とトークンと組）を特定して重み付き学習を行うことで汎化性能を向上できる可能性を示唆した。モデルとタスクの組み合わせによってはdropoutで性能が低下することもあるため、重みの固定とブランチネットワークを使うなどして性能を維持する方法についても提案した。

第４章　多様かつ高性能なサブネットワークによるアンサンブル

機械学習モデルの性能向上手段は様々であるが、中でも多くのケースで特に大きな工夫や苦労もなく汎用的に使える方法論も存在する。教師データの追加やハイパーパラメータの調整のほか、アンサンブルや事前学習も非常に重要である。アンサンブルでは、同等程度の性能を持ったモデルを複数用意し、それらの出力を平均などで統合したものを最終的な出力とする。訓練と推論がモデルの数の分だけかかってしまうものの、これをするだけで一般的に性能が上がることが多い。異なる判断基準を持つモデルが独立した予測を行うことによって同時に誤る確率が下がると言われている。事前学習は近年特に重要になった処理で、主に教師ラベルのない大量のデータセットにおいて自動化可能な人為的なタスクを設計し、それを大規模に学習することで良いモデル初期値を得ることができる。その初期から教師あり学習であるファインチューニングを始めることで大幅な性能向上が見込める。これらアンサンブルと事前学習は本来併用可能であるが、しかし事前学習のコストを考えると複数のモデルをアンサンブル用に構築するのは困難である。そのため、現実的には事前学習のモデルを一つだけ用意して、そこからファインチューニングを数回異なるランダムシードで行うことでアンサンブルを構築せざるをえないケースが多い。しかし、典型的な言語モデルからのファインチューニングではモデルの多様性が不足し、アンサンブルの性能が伸び悩むことが報告されている。本章では、そのようなケースについて新たな多様性を導入し、アンサンブルの性能向上を助ける方法を提案する。事前学習済みの言語モデル内には多様な知識が存在し、様々なタ

スクに転移ができ、場合によっては枝刈りによってのみでも性能の良いモデルに変換できるケースもある。そのため、本章では複数のサブネットワークを事前学習済みモデルから抽出し、それぞれをファインチューニングすることで多様なモデル群を生成し、アンサンブルの性能を向上させることを目指す。まず、通常のファインチューニングと同等性能に達することができる単一のサブネットワークの探索を、宝くじ仮説の立証に使われてきた枝刈り法 iterative magnitude pruning によって実現する。しかし、実はそのままでは別のランダムシードを用いてもほとんど同じ枝刈りに到達してしまうことを示した。そこで、次にアンサンブルに重要とされる多様性をさらに高めるために、できるだけ異なる重みが枝刈りされるような工夫を提案する。Iterative magnitude pruning では、学習後に最も絶対値が小さい方から一定割合の重みを枝刈り対象に選ぶ。そのため、多様な枝刈りを求めるため、本章では絶対値を減少させやすくする L1 正則化を、ランダムシードごとに異なるマスクおよび強さで重みに適用する。その適用方法についても、積極的に他のモデルと排他的になるようにマスクを設定する手法と、完全に独立にランダムなマスクを生成する手法の双方を検証した。実験では、文書分類あるいは文書ペアの類似度回帰問題において、提案したサブネットワークによるアンサンブルがベースラインに対して一定の性能向上を示した。特に正則化によって多様性を増したアンサンブルが最も高い性能を達成した。また、正則化によって枝刈りの多様性が実際に増していることも示した。一方で、実験するタスクによってはそもそも元のネットワークと同等性能を示すサブネットワークが見つかりづらく、事前学習済み言語モデルでは、厳しい性能水準については宝くじ仮説が成り立たないケースがあることを先行研究の問題点の指摘とともに明らかにした。このようなタスクでもアンサンブル時の向上幅自体は改善したものの、アンサンブルのトータルの性能がベースラインを有意に上回る設定は見つけられなかった。それらの結果を通して、宝くじ仮説の関連研究へ実用上の新たな視点や研究目標も提示した。

第５章　結論

本論文では、ニューラルネットワークのサブネットワークの特性、特に一つのモデルを訓練した後に潜在的に構築される大量のサブネットワーク群の特性について着目し、その新たな応用方法を提案した。これまではあくまでネットワークを軽量化したものとしてや、訓練時に行われるノイズ的な処理の結果として扱われることが主だったサブネットワークについて新たな応用性を実証した。サブネットワーク群の中に潜む同質性と異質性について着目し、また、それらを訓練時から積極的にコントロールするような訓練方法や、訓練後にさえも所望の特性を持つサブネットワークを得るための探索方法を提案した。それぞれの手法は、本来複数のモデルを要するような様々な応用について、モデル群の代替としてのサブネットワーク群という観点を提示し、実験的にもその有用性を多角的に示した。