SvMoE: MoE ルータの教師あり学習

村田栄樹　河原大輔
早稲田大学理工学術院
{eiki.1650-2951@toki., dkw@}waseda.jp

概要

Mixture-of-Experts はそのパラメタ数に対して計算コストが小さく, 大規模言語モデルの実用に向けて重要な技術である. しかし, エキスパートを選択するルーティングでは選択が偏り, 効率的なパラメタの利用が難しいという問題がある. それに対して, エキスパート選択を均一にする追加損失が使われるが言語モデル性能に干渉することがわかっている. 本研究では, TF-IDF を教師信号とした教師あり学習で Mixture-of-Experts のルーティングを訓練することを提案する. ケーススタディとして法律へのドメイン特化を扱い, 追加損失なしの提案手法は追加損失を使用するベースラインに比肩する結果を得た.

1　はじめに

Mixture-of-Experts (MoE) は, エキスパートと呼ばれる複数のモジュールの出力をアンサンブルする技術である [1, 2]. アンサンブルには, ルータと呼ばれるモジュールの出力を重みとして使用する. 深層学習 [3] や言語モデリング [4] にも適用され, 特にエキスパートの選択を疎にしたもの [4] はそのパラメタ数に対して計算コストが小さいという利点がある.

しかし, MoE のルータによるエキスパートの選択ではその分布が偏る問題がある. 近年の言語モデルにおける研究ではこれを避けるために, ルーティングを均一に近づけるための Load-Balancing 損失 (LB 損失) を言語モデル損失に加えて訓練される [4, 5, 6]. 一方で, LB 損失は言語モデルの性能に対して干渉するという問題もある [7].

この問題を踏まえて本研究では, MoE 全体の訓練の前にルータを教師あり学習で訓練する, **SvMoE** (**S**upervised **M**ixture-**o**f-**E**xperts) を提案する. 各エキスパートの担当するドキュメントを事前に定義し, 正規化された TF-IDF を教師信号として使用することでエキスパートの学習した知識とルーティングを対応させる. ここでは訓練データの TF-IDF がそのトークンが得意なエキスパートに対応することを仮定している. つまり, (1) 定義されたドキュメントでエキスパートをそれぞれ訓練し (2) それらをマージして MoE モデルとして, (3) TF-IDF でルータを訓練したのちに (4) MoE 全体を訓練することにより LB 損失を使用せずに適切な学習を実現する.

ケーススタディとして日本の法律へのドメイン特化を扱う. 提案手法による日本の法律文を使用した継続事前学習を事前学習済みモデルに対して行い, 司法試験のベンチマーク [8] で評価する. 提案手法により訓練されたルータでは TF-IDF に対応した選択が確認され, LB 損失を用いない提案手法は LB 損失を使用したベースラインとほぼ同等の結果を得た.

2　関連研究

2.1　Mixture-of-Experts

MoE は Transformer ベースの言語モデルにも多く活用されており [5, 6, 9, 10, 11, 12, 13, 14], そのほとんどは Transformer ブロックの MLP 層を並列に並べた形式で言語モデリングを目的として訓練される.

MoE はパラメタ数に対して効率的であるが, 訓練時に特定のエキスパートのみが多く活性化されることが知られている [4]. これはエキスパートの専門性を損ない, パラメタの効率的な利用を妨げるため, 均一なルーティングを促す LB 損失が採用される [4, 5, 6]. LB 損失は訓練時に通常の言語モデル損失に加えられる.

しかし, LB 損失は言語モデルの性能に悪い影響を及ぼすことを報告した研究 [7] もある. この研究では, 言語モデル性能と LB 損失のトレードオフがあるためハイパーパラメタの調整が必要であることを指摘し, LB 損失を使用せずにルーティングを改善している. 具体的には, 動的に学習されるバイアス項をルータの出力に加えることで言語モデル損失の勾配に干渉しない Load-Balancing を提案し, 後続の研究でも利用されている [14]. 本研究もこの立場から,

LB 損失を使用せずに MoE モデルの性能を引き出すことを目指す.

2.2 ドメイン特化モデル

事前学習済み Transformer モデルの登場以降, 医療や金融などのドメインに特化したモデルが開発されている [15, 16, 17, 18]. これらは, ドメイン特化のための事前学習や通常の事前学習済みモデルを継続事前学習することで構築される.

Branch-**T**rain-**M**erge (BTM) [19] は, 個別に訓練された複数のドメイン特化モデルをマージすることにより, 性能を改善している. ただし, MLP 層を並列化するのではなく Transformer モデル全体を並列化し, トークンロジットの段階でのマージとなっている. また, C-BTM [20] ではクラスタリングされた学習コーパスで BTM の手法を用いることで, 大規模なデータセットでの並列化された訓練を実現した. **B**ranch-**T**rain-Mi**X** (BTX) [21] では, BTM と同様にドメインで分割されたデータを使用して事前学習済みの Transformer をそれぞれ fine-tuning したのちに, それらの MLP 層を並列に統合することで MoE モデルとする. ルーティングのための MoE モデル全体の訓練では, LB 損失が使用される.

本研究では, BTX をベースラインとして LB 損失を使用しないルーティングの学習を試みる.

3 Supervised MoE

本研究では LB 損失の悪影響を考慮し, 事前にルータをドキュメントの TF-IDF によって教師あり学習する **SvMoE** を提案する. 特に, ドメインとしてさらにサブドメインに分割できるようなものを想定し, エキスパートは各サブドメインに特化する設定とする. SvMoE は, エキスパートの訓練, エキスパートのマージ, ルータの教師あり学習, MoE モデル全体の fine-tuning の 4 段階からなる.

3.1 ルータの教師データ作成

SvMoE のエキスパートは分割されたサブドメインに対してそれぞれ特化するように訓練される. ここでは, エキスパートの訓練されたデータに合わせてルーティングするための教師データを作成する. まず, N 個のサブドメインを定義する. あるドメインのデータ $\mathcal{D}_0$ に対して, M 個のカテゴリ ($M > N$) が定義されていることを前提とする. カテゴリをそれぞれドキュメントとして TF-IDF を計算する. この TF-IDF を特徴量として、クラスタリングを実行することで N 個のサブドメインデータ d_n $\left(1 \le n \le N,\ d_i \cap d_j = \emptyset\ (\forall i \neq j)\right)$ を得る. そして, $\mathcal{D} = \bigcup_{n=1}^{N} d_n$ を学習対象のデータセットとする.

次に, ルータ訓練時の教師信号となる TF-IDF を取得する. サブドメインデータをそれぞれドキュメントとして TF-IDF を計算することで, $N \times |V|$ の行列を得る (V は語彙). 以降、各トークン $t \in V$ に対応する正規化された TF-IDF ベクトルを

$$\mathrm{TFIDF}_t = \left(\mathrm{tfidf}_{t,1}, \mathrm{tfidf}_{t,2}, \ldots, \mathrm{tfidf}_{t,N}\right) \in \mathbb{R}^N$$

と表す. ただし, $\sum_{n=1}^{N} \mathrm{tfidf}_{t,n} = 1 \quad (\forall t \in V)$ である.

3.2 モデルの構築

エキスパートの訓練 3.1 節で用意した $\mathcal{D}$ で N 個のエキスパート $\{E_1, \ldots, E_N\}$ を訓練する. 1 つの事前学習済み Transformer モデルを言語モデリングを目的としてそれぞれ訓練することで, 各サブドメインに特化したエキスパート群を得る.

エキスパートのマージ 得られたエキスパート群をマージして MoE モデルとする. BTX に倣い, 各 Transformer ブロックについてエキスパート群の FFN 層を並列に並べることでエキスパート数が N の MoE 層とする. Attention 層などの他の層については平均処理によってマージする. ただし, MoE 層のルータについてはランダムに初期化する.

ルータの教師あり学習 マージされた MoE モデルを 3.1 節で得た TF-IDF 信号で訓練する. モデルのブロック数を L とすると, 各ブロックの MoE 層がルータを持つため, L 個のルータが存在する. 訓練バッチ $b = t_1 t_2 \ldots t_{|b|} (t_i \in V)$ を入力したときの目的損失 $\mathcal{L}_{\mathrm{SvMoE}}$ を以下で定める.

$$\mathcal{L}_{\mathrm{SvMoE}} = \sum_{t \in b} \sum_{l=1}^{L} \ell_{\mathrm{CE}}(\mathrm{softmax}(RL_l), \mathrm{TFIDF}_t)$$

ただし, $RL_l \in \mathbb{R}^N$ は l ブロック目のルータのロジット, $\mathrm{softmax}(\cdot)$ はソフトマックス関数, $\ell_{\mathrm{CE}}(\cdot,\cdot)$ はクロスエントロピー損失である. $\mathcal{L}_{\mathrm{SvMoE}}$ によってルータのパラメタを訓練することにより, 訓練されたデータに多く含まれたトークンをそのエキスパートに割り当てることを促す. ただし, ルータを除くモデルのパラメタはこの段階では更新しない.

MoE モデルの fine-tuning ルータを学習した MoE モデル全体を, 言語モデリングを目的として再び学習する. このとき, LB 損失は用いない. この段階を持って, 提案手法の最終的なモデルとする.

表 1: サブドメインごとのトークン数. 単位はすべて百万トークン.

ID	訓練	検証	テスト	合計
1	23.0	3.2	1.9	28.1
2	18.0	2.0	2.6	22.6
3	24.0	3.5	2.4	29.9
4	32.1	3.8	4.7	40.5
5	31.3	4.3	4.1	39.7
6	73.4	5.9	8.1	87.4
7	21.0	2.9	2.5	26.4
8	39.5	2.0	5.3	46.9

4 モデル構築実験

4.1 設定

データ処理と訓練の設定, および評価方法について述べる. $N = 8$ として, モデルとトークナイザはすべて llm-jp/llm-jp-3-1.8b を用いる. このトークナイザは TF-IDF の計算にも使用する.

データセット 日本の法律条文データを対象に実験する. e-Gov から取得した条文に対して法令分類をカテゴリ ($M = 50$) としたものを $\mathcal{D}_0$ とする. トークナイザを使用して TF-IDF を計算し, 要素数が等しくなるようなスペクトラルクラスタリング[1)]によって 8 つのサブドメインに分割されたものを $\mathcal{D}$ とする. テスト分割を 8:1:1 で行う. 表 1 にサブドメインごとのトークン数を示す.

モデル まず, 得られた 8 つの法律サブドメインそれぞれを使用して, 言語モデリングで llm-jp/llm-jp-3-1.8b を訓練し 8 つのエキスパートモデルを得る. これら 8 つのエキスパートを 3.2 節の手続きに従って, MoE モデルへのマージ, ルータの訓練, 全体を fine-tuning したものが提案手法のモデル (**SvMoE**) である. 比較として, 全体の fine-tuning 時に LB 損失を加えたものも評価する.

また, 8 つのエキスパートをマージし, 全体を fine-tuning したモデル (**BTX**) をベースラインとして比較する. BTX ベースラインについても, 全体の fine-tuning 時に LB 損失を加えたものと加えなかったものをそれぞれ評価する.

MoE アーキテクチャを持つ全てのモデルにおいて, LB 損失の係数 α は BTX に則り, $\alpha = 0.01$ とする. また, 選択されるエキスパート数は 2 で固定する.

さらに, MoE ではない通常の Transformer モデルのベースライン (**Dense**) も用意する. これは同じ llm-jp/llm-jp-3-1.8b を, サブドメインごとではなくデータ全量を使用して訓練したものである.

評価 $\mathcal{D}$ に対して評価し, ベースラインと提案手法を比較する. ただし, 継続事前学習による知識の定着度合いを測ることが目的であるため, 訓練・検証・テストセットのそれぞれで評価する. 評価指標として, 以下の 3 つを用いる.

- PPL: 与えられたテキストに対する, モデルのパープレキシティの平均.
- CMR_k (条件付き平均順位): ルータロジットの値による n 番目のエキスパートの順位 r_n^{RL} が k であったときに, TF-IDF による n 番目のエキスパートの順位 r_n^{TFIDF} の平均. つまり, $\mathrm{CMR}_k = \mathbb{E}[r_n^{\mathrm{TFIDF}} \mid r_n^{RL} = k]$ である. この値が k に近いほど TF-IDF に従ってルーティングされているといえる. $k = 1, 2$ について報告する.
- S_{RL}: ルータロジットのエントロピーの平均. この値が小さいほどルータの確信度が大きいと解釈する.

4.2 定量評価

表 2 に $\mathcal{D}$ のテストセットに対しての評価結果を示す. 残りのセットについては付録 A に結果を示す.

まず, パープレキシティについては Dense が最も良く, BTX と SvMoE がそれに続く結果となった. BTX と SvMoE の両方で, LB 損失を使わない設定でより良い結果を得た. これは, 先行研究 [7] で指摘されたように LB 損失が言語モデル性能に影響を与えた結果である.

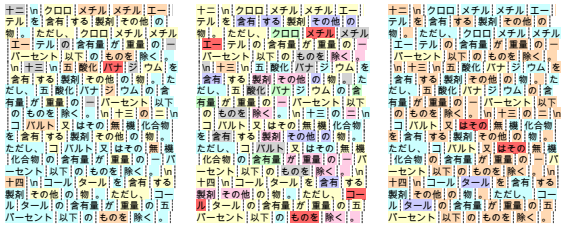
条件付き平均順位 CMR_k については, SvMoE が BTX より良い結果を示した. これは, TF-IDF を教師信号とした訓練により, それに従ったルーティングが言語モデルの推論時にも可能だということを示している. ルータロジットのエントロピー S_{RL} も同様に, SvMoE がより良い結果を示した. ルータの訓練により, 対応するエキスパートを確実にアクティブにすることができていると考えられる.

まとめると, 提案手法は想定通り TF-IDF に従ったルーティングが可能だが, それはパープレキシティの改善には繋がらなかったといえる. MoE はしばしば統語情報によってエキスパートを選択する [13, 22]. 本研究ではより頻度情報を重視したことが, パープレキシティの改善に繋がらなった要因として考えられる.

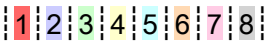
1) https://github.com/anamabo/Equal-Size-Spectral-Clustering

表 2: 訓練に使用したデータのテストセットに対する評価結果. 最も良い値と次に良い値をそれぞれ**太字**と下線で示す. 提案手法は最下列である.

モデル	LB 損失	PPL	CMR_1	CMR_2	S_{RL}
Dense	-	**1.156±0.462**	-	-	-
BTX	✓	1.187±0.407	4.063±2.237	4.742±2.206	1.805±0.198
	×	1.169±0.401	2.847±2.025	4.058±2.152	1.705±0.261
SvMoE	✓	1.245±0.443	2.538±2.097	**3.503±2.374**	**0.465±0.316**
	×	1.199±0.434	**2.533±2.092**	3.518±2.365	0.466±0.319



(a) SvMoE (b) BTX (c) TF-IDF



(d) 配色の凡例

図 1: エキスパート選択の比較. 図 1a,1b は $l = 15$ を報告. 特定化学物質障害予防規則より抜粋.

4.3 定性評価

図 1 にエキスパート選択の例を示す. 入力した文章は d_4 のテストセットのものである.

SvMoE では入力トークンの TF-IDF が高くなっている E_5, E_6 が多く選択されている. 一方で BTX では, やや E_4 が多いもののどのエキスパートも比較的均等に選択された. 例えば “クロロ” は d_4, d_5 の TF-IDF が高く d_1 ではほとんど現れない. SvMoE では “クロロ” とその次の単語を推論する際にそれぞれ E_4, E_6 が使用されるが, BTX では E_3, E_1 が使用されることが観察された. 定性的にも, SvMoE は BTX と比較して入力トークンの頻度情報を用いた選択がされることが確認された.

5 ダウンストリームタスクでの評価

4.2 節では SvMoE による改善が見られなかった. そこで, ダウンストリームタスクにおいて提案手法とベースラインを比較する. 先行研究 [8] で構築された司法試験ベンチマークを対象とする.

5.1 設定

司法試験は正しいものの組み合わせを選ぶなど複雑な形式であるため, 緩和タスクとして設定された各文の正誤を問う正誤判定タスクを扱う. 正誤判定タスクは, ある文に対してその内容があっていれば 1 をそうでなければ 0 を出力するものである. Few-shot の設定で, 4 節で構築した 5 つのモデルを評価する. ショットとして 2019 年度の問題をランダムに 5 つ与え, 2023 年度の 180 問を評価対象とする.

表 3: 司法試験正誤判定タスクの結果.

モデル	LB 損失	正解率 (%)
Dense	-	48.33
BTX	✓	**51.67**
	×	47.22
SvMoE	✓	48.33
	×	51.11

5.2 結果

表 3 に正誤判定タスクの結果を示す. BTX ベースラインは, LB 損失を使用した場合に最も高いスコアとなり, 使用しない設定では 4 ポイント以上スコアが下落した. 一方で SvMoE では, LB 損失を使用しな場合でも高いスコアを示し, LB 損失を使用した BTX ベースラインとほぼ同等の結果となった. LB 損失を使用する SvMoE では, 使用しないものと比較して 3 ポイント程度のスコア下落となった.

4.2 節のパープレキシティでの評価で SvMoE はベースラインと比較して良い結果ではなかったが, ダウンストリームタスクではベースラインに比肩する性能を示した. 表 3 からも LB 損失がモデルの最終的な性能に小さくない影響を与えることを考慮すると, 提案手法はハイパーパラメタの設定も必要ないことなどから, 一定の有用性があると言える.

6 おわりに

本研究では LB 損失がモデルの性能に与える影響を考慮して, LB 損失を使用せずに適切にルータを学習するフレームワーク SvMoE を提案した. 提案手法に基づいた MoE モデルを構築し, TF-IDF に従ったルーティングが可能であることを確認した. また, ダウンストリームタスクでは LB 損失を使用するベースラインに比肩する結果を得た.

謝辞

本研究は「戦略的イノベーション創造プログラム (SIP)」「統合型ヘルスケアシステムの構築」JPJ012425 の補助を受けて実施した.

参考文献

[1] 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.

[2] M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the em algorithm. In **Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)**, Vol. 2, pp. 1339–1344 vol.2, 1993.

[3] David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. arXiv, 2013. abs/1312.4314.

[4] 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 **International Conference on Learning Representations**, 2017.

[5] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. {GS}hard: Scaling giant models with conditional computation and automatic sharding. In **International Conference on Learning Representations**, 2021.

[6] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. **Journal of Machine Learning Research**, Vol. 23, No. 120, pp. 1–39, 2022.

[7] Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts, 2024. abs/2408.15664.

[8] チェジョンミン, 笠井淳吾, 坂口慶祐. 日本の司法試験を題材とした gpt モデルの評価. 言語処理学会第 30 回年次大会発表論文集, 2024.

[9] Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. In **The Eleventh International Conference on Learning Representations**, 2023.

[10] Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. Glam: Efficient scaling of language models with mixture-of-experts. arXiv, 2021. abs/2112.06905.

[11] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. arXiv, 2022. abs/2207.04672.

[12] Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Y Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. Mixture-of-experts meets instruction tuning: A winning combination for large language models. In **The Twelfth International Conference on Learning Representations**, 2024.

[13] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. arXiv, 2024. abs/2401.04088.

[14] DeepSeek-AI. Deepseek-v3 technical report. arXiv, 2024. abs/2412.19437.

[15] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models. arXiv, 2023. abs/2305.09617.

[16] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. arXiv, 2023. abs/2303.17564.

[17] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Climatebert: A pretrained language model for climate-related text. arXiv, 2022. abs/2110.12010.

[18] Masanori Hirano and Kentaro Imajo. Construction of domain-specified japanese large language model for finance through continual pre-training. arXiv, 2024. abs/2404.10555.

[19] Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. Branch-train-merge: Embarrassingly parallel training of expert language models. In **First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022**, 2022.

[20] Suchin Gururangan, Margaret Li, Mike Lewis, Weijia Shi, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. Scaling expert language models with unsupervised domain discovery, 2023. abs/2303.14177.

[21] Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Roziere, Jacob Kahn, Shang-Wen Li, Wen tau Yih, Jason E Weston, and Xian Li. Branch-train-mix: Mixing expert LLMs into a mixture-of-experts LLM. In **First Conference on Language Modeling**, 2024.

[22] Dongyang Fan, Bettina Messmer, and Martin Jaggi. TOWARDS AN EMPIRICAL UNDERSTANDING OF MOE DESIGN CHOICES. In **ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models**, 2024.

A　モデル構築の詳細

クラスタリングの詳細　本文でも触れたように, e-Gov[2)]から取得したカテゴリを TF-IDF を特徴としてクラスタリングする. ただしデータ量に偏りがあるため, ‘国税’, ‘金融・保険’ および ‘地方財政’ については 1 カテゴリを 1 サブドメインとして固定する. つまり, 47 カテゴリについて 5 クラスタを作るようにクラスリングする. クラスタリングの結果を表 4 に示す.

モデル訓練と評価の詳細　モデル訓練時の損失曲線は図 2 のようになった. 訓練の最初期には, ルータを訓練済みの SvMoE がランダムに初期化する BTX より低い損失を示した. しかし, 最終的な損失は LB 損失を使用しない BTX が最も良い結果を示した. ただし 5 節で示したように損失はモデルの性能に必ずしも直結しない.

次に表 5 に訓練および検証セットでのモデルの評価を示す. テストセットについて表 2 で示した結果と同様の評価である. 全体としてテストセットでの結果と同様の傾向となったが, 訓練セットで他のセットと比較して良い結果を示した.

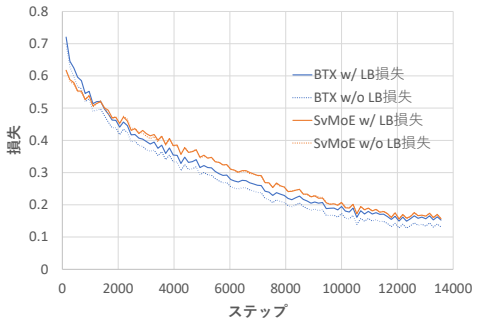


図 2: 提案手法とベースラインの損失曲線.

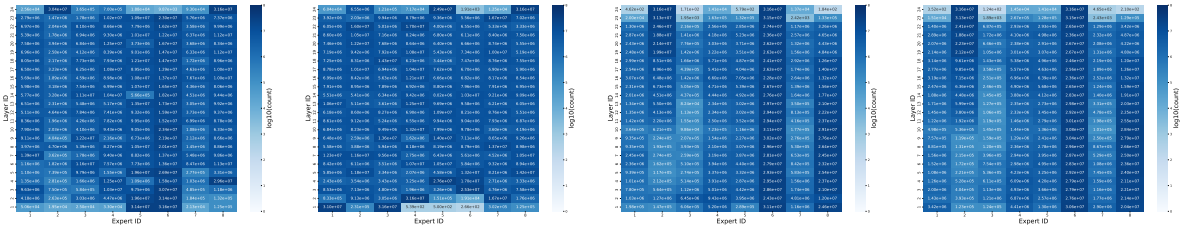
また, テストセットでのエキスパート選択の頻度をカウントした結果を図 3 に示す. 表 1 に示すように入力データ数に偏りがあるため, SvMoE はそれに対応したエキスパートが多く選択されている. 一方で BTX では, 特に LB 損失を使用した場合に入力データの分布によらず均等に選択されていることがわかる. また, 入力層や出力層の近辺ではどの場合においても偏りが出ることがわかる.

表 4: 各サブドメインに割り当てられた法令分類のカテゴリ.

ID	カテゴリ名
1	‘刑事’, ‘国会’, ‘鉱業’, ‘事業’, ‘商業’, ‘土地’, ‘文化’, ‘司法’, ‘民事’, ‘外事’
2	‘水産業’, ‘国債’, ‘地方自治’, ‘産業通則’, ‘災害対策’
3	‘憲法’, ‘財務通則’, ‘郵務’, ‘行政手続’, ‘都市計画’, ‘道路’, ‘社会福祉’, ‘社会保険’, ‘林業’, ‘貨物運送’
4	‘国有財産’, ‘行政組織’, ‘国家公務員’, ‘国土開発’, ‘労働’, ‘統計’, ‘教育’, ‘海運’, ‘農業’, ‘防衛’
5	‘観光’, ‘警察’, ‘消防’, ‘工業’, ‘電気通信’, ‘環境保全’, ‘外国為替・貿易’, ‘厚生’, ‘陸運’, ‘河川’, ‘航空’, ‘建築・住宅’
6	‘国税’
7	‘金融・保険’
8	‘地方財政’

表 5: 訓練に使用したデータに対する評価結果. 各セット内で最も良い値と次に良い値をそれぞれ**太字**と下線で示す.

データ	モデル	LB 損失	PPL	CMR_1	CMR_2	S_{RL}
訓練	Dense	-	**1.042±0.087**	-	-	-
	BTX	✓	1.089±0.159	4.043±2.254	4.755±2.205	1.802±0.201
		×	1.069±0.145	2.802±2.035	4.087±2.150	1.691±0.269
	SvMoE	✓	1.141±0.196	**2.473±2.078**	**3.543±2.369**	**0.449±0.316**
		×	1.099±0.202	2.477±2.079	3.563±2.360	0.450±0.318
検証	Dense	-	**1.154±0.331**	-	-	-
	BTX	✓	1.188±0.320	4.145±2.214	4.749±2.201	1.803±0.199
		×	1.169±0.316	3.057±2.086	4.086±2.152	1.716±0.256
	SvMoE	✓	1.243±0.348	2.687±2.164	**3.558±2.397**	0.477±0.327
		×	1.201±0.340	**2.681±2.161**	3.569±2.385	**0.476±0.329**



(a) BTX (LB 損失なし)　(b) BTX (LB 損失あり)　(c) SvMoE (LB 損失なし)　(d) SvMoE (LB 損失あり)

図 3: 各モデルのレイヤごとのエキスパート選択の頻度.

2)　https://laws.e-gov.go.jp/bulkdownload