分布外データの棄却機能を持つ モダンホップフィールドネットワーク

森合遼[1], 井上中順[1], 田中正行[1], 川上玲[1], 池畑諭[1,2], 佐藤育郎[1,3]

[1]東京工業大学, [2]国立情報学研究所, [3]デンソーITラボラトリ

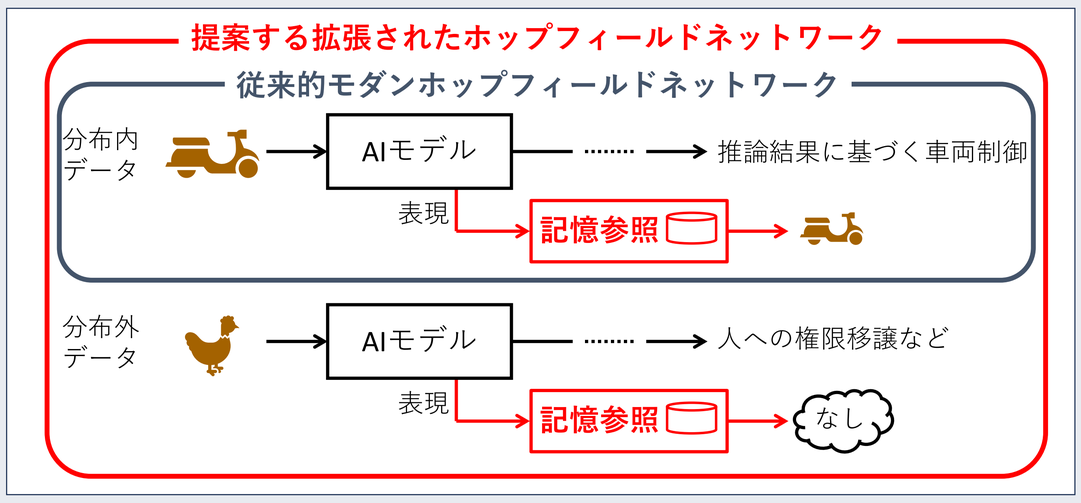
Recognition and Learning Algorithm Laboratory

東京工業大学 Tokyo Institute of Technology

1. 導入

動機 セーフティクリティカルなAIの応用（自動運転など）において、**学習データに帰着された推論根拠**を提示することが重要と考える。

目的 分布内クエリについては従来通り記憶データとの関連付けを行い、**分布外クエリについては記憶データとの関連付けを棄却する**、記憶参照モデルを開発する。



2. モダンホップフィールドネットワーク(MHN)

概要 入力データ（クエリ）に対して、記憶させたデータを関連付ける

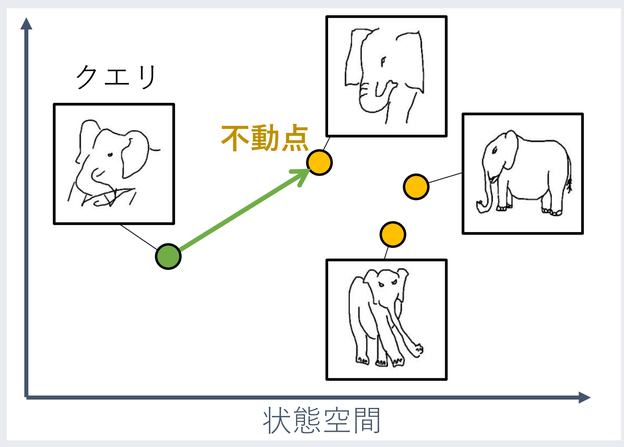
仕組 重みとして記憶データを保持したRNNの状態更新により、不動点（収束した状態）と記憶データを対応付ける

記憶データの重みを持つ２層RNN

$$v \leftarrow W\mathrm{softmax}(\beta h), \quad h \leftarrow W^{\mathsf{T}} v$$

$$W^{\mathsf{T}} = \begin{bmatrix} \text{記憶データ 1} \\ \text{記憶データ 2} \\ \cdots \\ \text{記憶データ N} \end{bmatrix}$$

状態更新と記憶データの特定



状態更新が最小化するエネルギー関数 [D. Krotov+ ICLR2020, M. Widrich + NeurIPS2020]

$$E(v,h) = v^{\mathsf{T}} \frac{\partial L_v}{\partial v} - L_v + h^{\mathsf{T}} \frac{\partial L_h}{\partial h} - L_h + v^{\mathsf{T}} W h$$

ラグランジアン

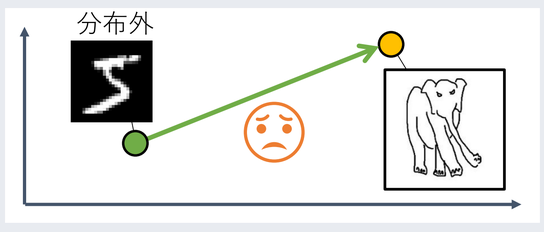
$$L_v = \frac{1}{2}\sum_{i=1}^{M} v_i^2$$

$$L_h = \frac{1}{\beta}\log \sum_{j=1}^{N} e^{\beta h_j}, \ \beta > 0$$

課題

分布外のクエリを無理やり記憶データに関連付けてしまう。

理由）エネルギー関数が記憶データに対応する局所最小**しか**持たない



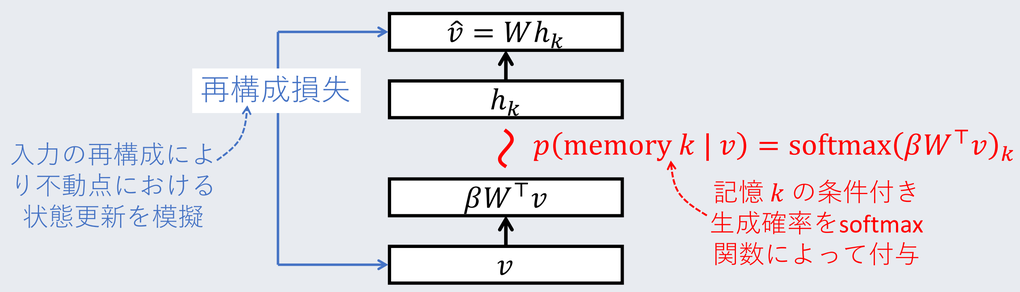
3. 提案手法

従来手法との差分

1. 重み行列 W の最適化により、MHNの**状態空間上の確率密度**を計算可能に
2. ラグランジアンの再定式化により、**分布外クエリを原点へ遷移**させる新たな更新則を導入

1. 状態空間上の確率密度関数

ステップ１ RNNをアンロール展開し、隠れ層を記憶ベクトルの確率的サンプリングによって置き換えたオートエンコーダを形成



ステップ２ カテゴリカル分布に従う確率変数を含む上記ネットワークの重みWの更新により、対数尤度の下限を近似的に最大化

対数確率密度を隠れ層ラグランジアンに帰着可能 $\log p(v) \propto \log \sum_{j=1}^{N} \exp(\beta W^{\mathsf{T}} v)_j = L_h(\beta W^{\mathsf{T}} v)$

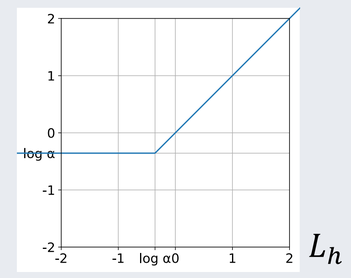
補足：不動点近傍において、入力と番目の記憶データとの同時確率が次式によって与えられる

$$p(k, v) \propto \exp(\beta W^{\mathsf{T}} v)_k$$

2. 分布外クエリを原点に遷移させる更新則

新たな隠れ層ラグランジアン

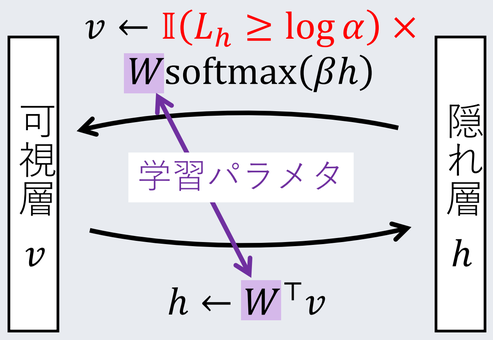
$$L_h^{\alpha} = \mathrm{ReLU}(L_h - \log \alpha) + \log(\alpha)$$



導かれる更新則

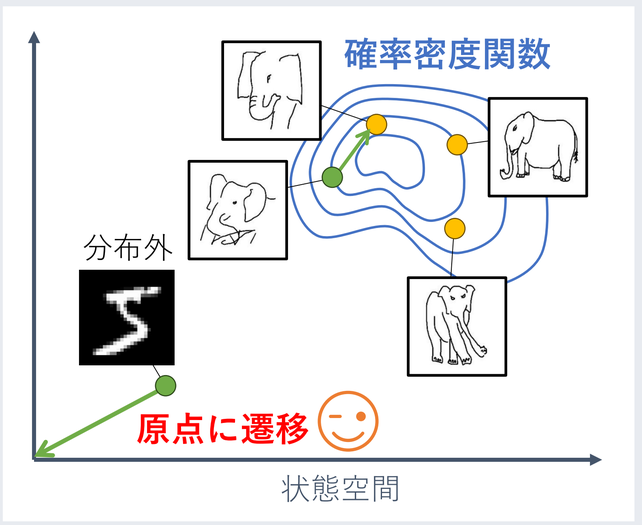
$$v_{t+1} = \mathbb{I}(L_h \geq \log \alpha)\, W\mathrm{softmax}(\beta W^{T} v_t)$$

補足：新たなラグランジアンにより定義されるエネルギー関数に勾配降下を適用することで状態 v の更新則を導くことが可能



$\log p(v) \propto L_h < \log \alpha$
→ 分布内として従来同様の更新

$\log p(v) \propto L_h \geq \log \alpha$
→ 分布外として原点に遷移
（分布外であることを特定可能）



4. 評価

タスク 画像データセットを用いた分布外検出

評価指標 FPR95(TPRが95%となる時のFPRの値)↓

ベースライン MSP[1], Energy[2], React[3], SHE[4], HE[4]
※いずれもエネルギーベースドモデルを利用した分布外検出手法。確率的なモデリングは不使用。

設定 分布内データセット： CIFAR-10
分布外データセット： SVHN, LSUN-C, LSUN-R, iSUN, Places, DTD, Tiny ImageNet, SUN, iNaturalist
※ CIFAR-10で学習されたモデル(ResNet-18)の特徴空間において、分布の内外を判定

結果 8データセットで全てのベースラインを改善

表. 分布外検出精度比較(%)↓

	MSP [1]	Energy [2]	ReAct [3]	SHE [4]	HE [4]	Ours
SVHN	74.99	51.81	54.35	5.66	**5.44**	5.68
LSUN-C	44.59	14.91	14.73	8.32	8.42	**4.93**
LSUN-R	38.93	14.98	14.51	4.75	4.69	**2.43**
iSUN	35.82	11.99	11.78	3.38	3.43	**1.68**
Places	39.16	21.06	17.36	0.44	0.44	**0.40**
DTD	54.93	54.58	48.99	9.05	9.05	**7.38**
Tiny ImageNet	44.49	27.76	28.06	10.90	10.17	**7.69**
SUN	38.34	20.80	15.08	0.00	0.00	**0.00**
iNaturalist	68.40	65.16	52.10	2.81	2.80	**2.43**

[1] D. Hendrycs+, A baseline for detecting misclassified and out-of-distribution examples in neural networks, ICLR2017.
[2] Z. Liu+, Biologically Plausible Sequence Learning with Spiking Neural Networks, NeurIPS2020.
[3] Y. Sun+, React: Out-of-distribution detection with rectified activations, NeurIPS2021.
[4] Z. Jinsong+, Out-of-Distribution Detection based on In-Distribution Data Patterns Memorization with Modern Hopfield Energy, ICLR2023.

本研究は東京工業大学「DENSO IT LAB 認識・学習アルゴリズム共同研究講座」における研究プロジェクト「機械学習の実応用に向けた品質基盤開発」の成果である.