Sequential Attention for Feature Selection

Taisuke Yasuda
Carnegie Mellon University
Computer Science Department

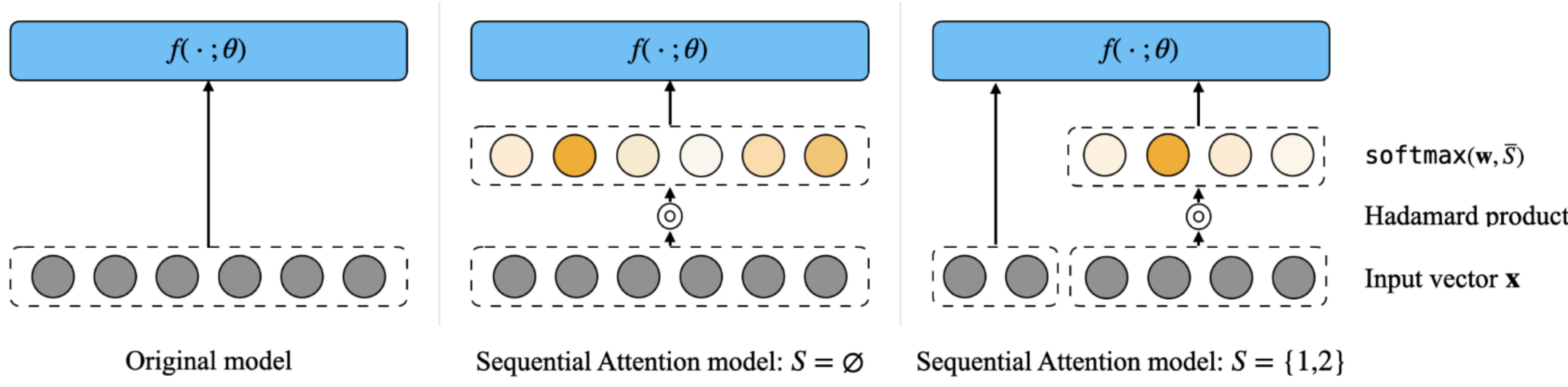
MohammadHossein Bateni, Lin Chen, Matthew Fahrbach, Thomas Fu, Vahab Mirrokni
Google

Feature Selection

- **Feature selection**
  - Given d features, select a subset of k features that maximizes model quality
    - Improves model interpretability
    - Reduces training/inference resources
    - Improve generalization by removing noisy features
  - Prior approaches
    - **Greedy algorithm**: requires training many models
      - Requires k rounds, $O(d)$ models trained per round
    - **L1 regularization**: selection occurs in one round, so it ignores residual/marginal values of features
    - **Attention-based feature selection**: same problem as ^

Main Contribution: Sequential Attention

- **Sequential Attention:** a new **efficient** and **greedy** feature selection algorithm
  - Efficiently simulate the greedy algorithm during training by **evaluating all candidate features at once** using an **attention/softmax mask**
    - Let $S \subseteq [n]$ be the currently selected features
    - Features $i \in S$ are weighted by 1 (unweighted)
    - Features $i \notin S$ are weighted by a softmax mask $\texttt{softmax}(\mathbf{w}, S)_i := \dfrac{\exp(\mathbf{w}_i)}{\sum_{j \in S} \exp(\mathbf{w}_j)}$
    - Train the model and add the feature $i \in S$ with largest attention weight to S



Theoretical Analysis

- We show that a variant of Sequential Attention that we use in practice has **provable guarantees** for the **sparse linear regression problem**
  - **Sparse linear regression**: Given an $n \times d$ design matrix $\mathbf{X}$, target vector $\mathbf{y}$, and a sparsity parameter k, output a k-sparse vector β that minimizes $\|\mathbf{X}\beta - \mathbf{y}\|_2^2 = \sum_{i=1}^{n} (\langle \mathbf{x}_i, \beta \rangle - \mathbf{y}_i)$
  - Our analysis shows the equivalence between three feature selection algorithms:
    - **Sequential Attention**
    - **Sequential LASSO** [Luo-Chen 2014]
      - Very little known guarantees
    - **Orthogonal Matching Pursuit** [Pati-Rezaiifar-Krishnaprasad 1993]
      - Has provable guarantees for sparse linear regression via weak submodularity arguments [Das-Kempe 2011]

Sequential Attention

function SEQUENTIALATTENTION(dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$, labels $\mathbf{y} \in \mathbb{R}^n$, model f, loss ℓ, size k)
  Initialize $S \leftarrow \varnothing$
  for $t = 1$ to k **do**
    Let $(\boldsymbol{\theta}^*, \mathbf{w}^*) \leftarrow \arg\min_{\boldsymbol{\theta}, \mathbf{w}} \ell(f(\mathbf{X} \circ \mathbf{W}; \boldsymbol{\theta}), \mathbf{y})$, where $\mathbf{W} = \mathbf{1}_n \mathrm{softmax}(\mathbf{w}, \overline{S})^\top$ for

$$\mathrm{softmax}_i(\mathbf{w}, \overline{S}) := \begin{cases} 1 & \text{if } i \in S \\ \dfrac{\exp(\mathbf{w}_i)}{\sum_{j \in \overline{S}} \exp(\mathbf{w}_j)} & \text{if } i \in \overline{S} := [d] \setminus S \end{cases}$$

    Set $i^* \leftarrow \arg\max_{i \notin S} \mathbf{w}_i^*$
    Update $S \leftarrow S \cup \{i^*\}$
  return S

Select $i^* \in [d]$ with the largest attention weight

Sequential LASSO

function SEQUENTIALLASSO(design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, response $\mathbf{y} \in \mathbb{R}^n$, size constraint k)
  Initialize $S \leftarrow \varnothing$
  for $t = 1$ to k **do**
    Let $\boldsymbol{\beta}^*(\lambda, S)$ denote the optimal solution to

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\arg\min} \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\beta}_{\overline{S}}\|_1$$

    Set $\lambda^*(S) \leftarrow \sup\{\lambda > 0 : \boldsymbol{\beta}^*(\lambda, S)_{\overline{S}} \neq \mathbf{0}\}$
    Let $A(S) = \lim_{\varepsilon \to 0} \{i \in \overline{S} : \boldsymbol{\beta}^*(\lambda^* - \varepsilon, S)_i \neq 0\}$
    Select any $i^* \in A(S)$
    Update $S \leftarrow S \cup \{i^*\}$
  return S

Set $\lambda > 0$ as large as possible without causing all coordinates to be 0

Select $i^* \in [d]$ with nonzero coordinate

Sequential Attention = Sequential LASSO

Lemma [Hoff 2017]. Let $l : \mathbb{R}^d \to \mathbb{R}$ and let $\lambda > 0$. Then,

$$\inf_{\beta, \mathbf{w} \in \mathbb{R}^d} l(\mathbf{w} \odot \beta) + \frac{\lambda}{2}\left(\|\mathbf{w}\|_2^2 + \|\beta\|_2^2\right) = \inf_{\beta \in \mathbb{R}^d} l(\beta) + \lambda \|\beta\|_1$$

Linear attention weights

LASSO

Orthogonal Matching Pursuit

function OMP(design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, response $\mathbf{y} \in \mathbb{R}^n$, size constraint k)
  Initialize $S \leftarrow \varnothing$
  for $t = 1$ to k **do**
    Set $\boldsymbol{\beta}_S^* \leftarrow \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^S} \|\mathbf{X}_S \boldsymbol{\beta} - \mathbf{y}\|_2^2$
    Let $i^* \notin S$ maximize

$$\langle \mathbf{X}_i, \mathbf{y} - \mathbf{X}_S \boldsymbol{\beta}_S^* \rangle^2 = \langle \mathbf{X}_i, \mathbf{y} - \mathbf{P}_S \mathbf{y} \rangle^2 = \langle \mathbf{X}_i, \mathbf{P}_S^\perp \mathbf{y} \rangle^2$$

    Update $S \leftarrow S \cup \{i^*\}$
  return S

Select $i^* \in [d]$ with maximum correlation with residual

Sequential LASSO = OMP

- This is our **main technical contribution**

Theorem [Yasuda-Bateni-Chen-Fahrbach-Fu-Mirrokni 2023] (informal). For sparse linear regression, Sequential LASSO selects some feature $i \in S$ maximizing the correlation with the residual at each step.

- **Proof sketch** (first step of selection only)
  - Primal problem: minimize $\|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda\|\beta\|_1$ over $\beta \in \mathbb{R}^d$
  - Dual problem: minimize $\|\mathbf{y} - \mathbf{u}\|_2^2$ over $\mathbf{u} \in \mathbb{R}^n$ s.t. $\|\mathbf{X}^\top \mathbf{u}\|_\infty \leq \lambda$
  - If $\lambda \geq \|\mathbf{X}^\top \mathbf{y}\|_\infty \to$ projection residual is 0 $\to \beta = 0$
  - If $\lambda < \|\mathbf{X}^\top \mathbf{y}\|_\infty \to$ projection residual is orthogonal to $\mathbf{X}_{i^*} \to \beta_{i^*} \neq 0$
    - Here, i^* witnesses the max of $\|\mathbf{X}^\top \mathbf{y}\|_\infty$

- **Remarks**
  - Prior known guarantees for Sequential LASSO only apply to statistical settings
  - Our result gives the first connection between **LASSO** and **submodularity**

Experiments

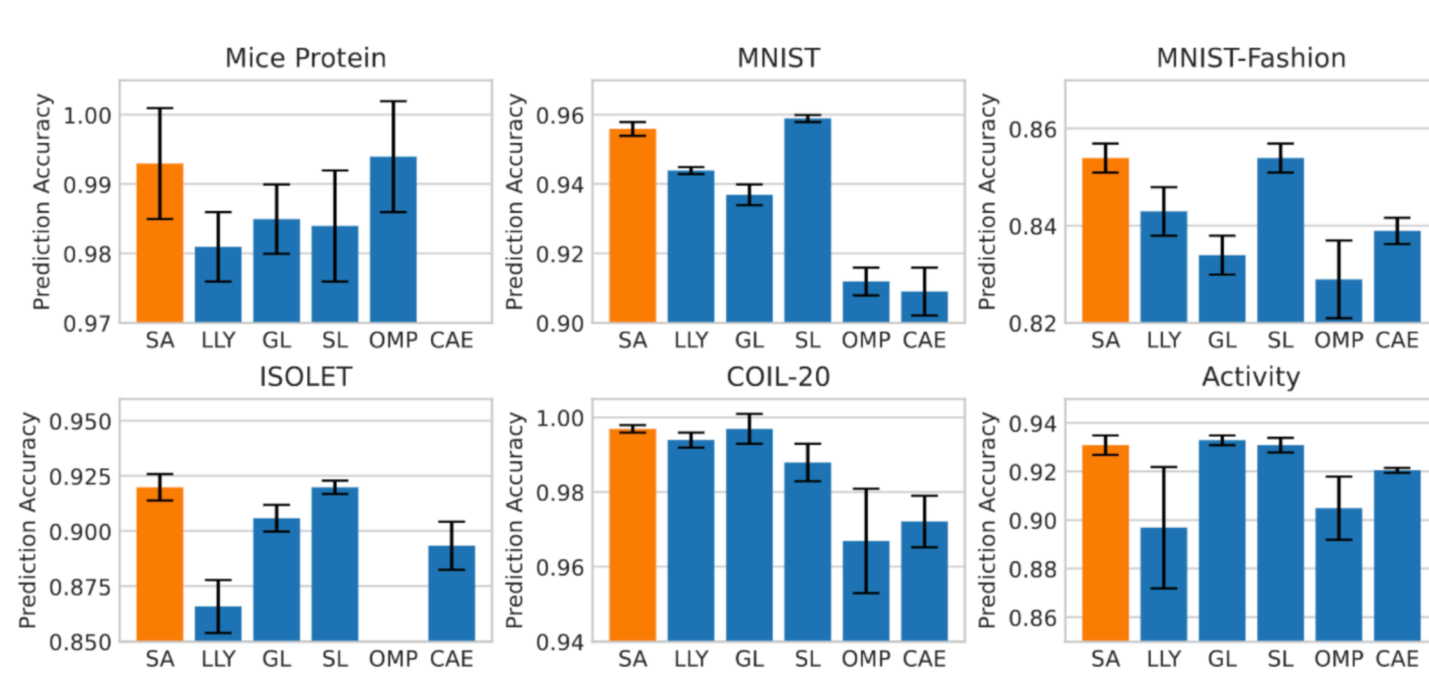


Figure 3: Feature selection results for small-scale neural network experiments. Here, SA = Sequential Attention, LLY = (Liao et al., 2021), GL = Group LASSO, SL = Sequential LASSO, OMP = OMP, and CAE = Concrete Autoencoder (Balın et al., 2019).