Feature Grouping as a Stochastic Regularizer for High Dimensional Structured Data

Sergül Aydöre
(Stevens Institute of Technology, USA)

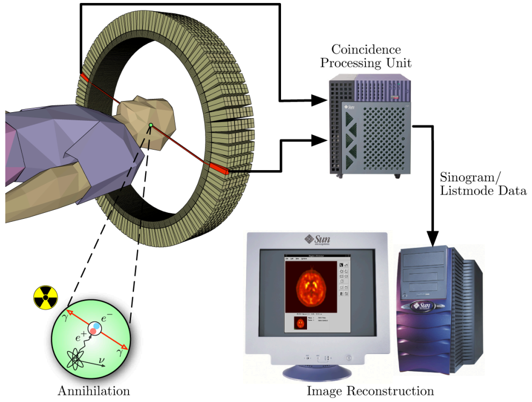
Bertrand Thirion
(INRIA, France)

Gaël Varoquaux
(INRIA, France)

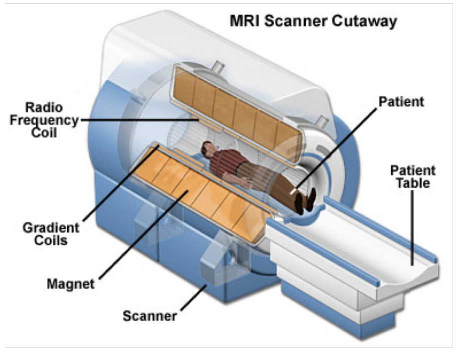
POSTER: Pacific Ballroom #121, 06/11, Tuesday

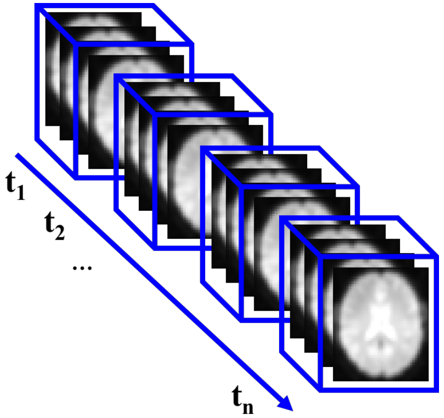
High Dimensional and Small-Sample Data Situations

- Brain imaging, Genomics, Seismology, Astronomy, Chemistry, etc.

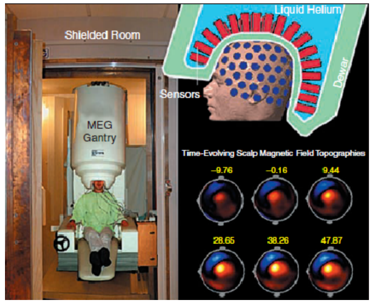


PET acquisition process wikipedia

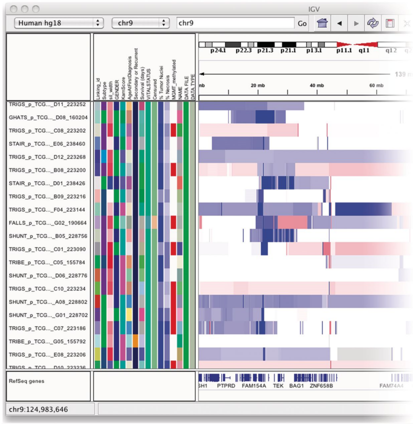




MRI Scanner and rs-fMRI time series acquisition [NVIDIA]

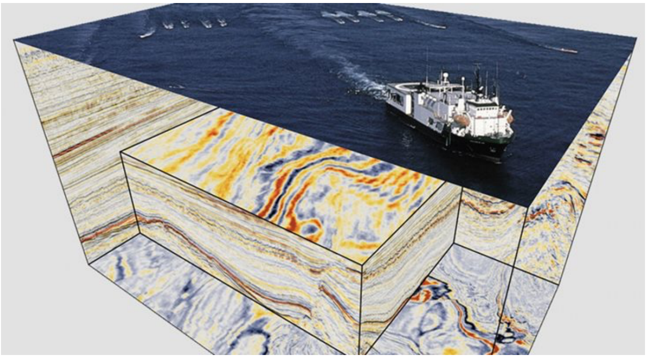


A typical MEG equipment [BML2001]



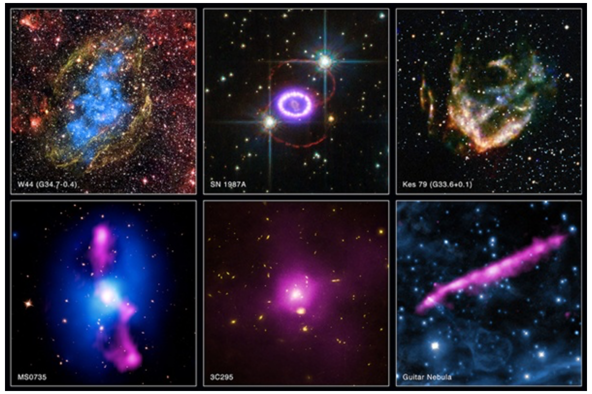
Genomics

Integrative Genomics Viewer, 2012



Seismology

https://www.mapnagroup.com



Astronomy

Astronomy Magazine, 2015

Fitting Complex Models in These Situations

Challenges

1. **Large feature dimension**: due to rich temporal and spatial resolution
2. **Noise in the data**: due to artifacts unrelated to the effect of interest
3. **Small sample size**: due to logistics and cost of data acquisition

Regularization Strategies

- **Early Stopping**: [Yao, 2007]
- ℓ_1 **and** ℓ_2 **penalties:** [Tibshirami 1996]
- **Pooling Layers in CNNs:** [Hinton 2012]
- **Group LASSO:** [Yuan 2006]
- **Dropout:** [Srivastana 2014]

Fitting Complex Models in These Situations

Challenges

1. **Large feature dimension**: due to rich temporal and spatial resolution
2. **Noise in the data**: due to artifacts unrelated to the effect of interest
3. **Small sample size**: due to logistics and cost of data acquisition

Regularization Strategies

- **Early Stopping**: [Yao, 2007]
- **ℓ_1 and ℓ_2 penalties:** [Tibshirami 1996]
- **Pooling Layers in CNNs:** [Hinton 2012]…………………… **TRANSLATION INVARIANCE**
- **Group LASSO:** [Yuan 2006]………………………………… **STRUCTURE + SPARSITY**
- **Dropout:** [Srivastana 2014]……………………………………… **STOCHASTICITY**

Fitting Complex Models in These Situations

Challenges

1. **Large feature dimension**: due to rich temporal and spatial resolution
2. **Noise in the data**: due to artifacts unrelated to the effect of interest
3. **Small sample size**: due to logistics and cost of data acquisition

Regularization Strategies

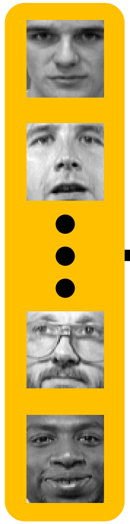
- **Early Stopping**: [Yao, 2007]
- ℓ_1 **and** ℓ_2 **penalties:** [Tibshirami 1996]
- **Pooling Layers in CNNs:** [Hinton 2012]…………………… **TRANSLATION INVARIANCE**
- **Group LASSO:** [Yuan 2006]……………………………… **STRUCTURE + SPARSITY**
- **Dropout:** [Srivastana 2014]……………………………………… **STOCHASTICITY**

- **PROPOSED**: **Use STRUCTURE & STOCHASTICITY**

Feature Grouping to Capture Structure

Algorithm

Training Data



Recursive Nearest Agglomeration (ReNA)

[Hoyos et al 2016]

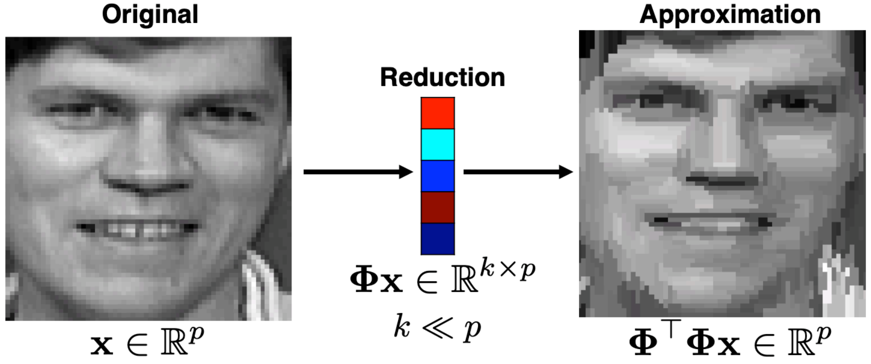
- **ReNA:** a data-driven, graph constrained feature grouping algorithm
- Each feature (pixel) is assigned to a cluster. Clusters are then recursively merged until the desired number of clusters remain.
- Benefits of ReNA: (i) a fast clustering algorithm (ii) leads to good signal approximations.

Feature Grouping Matrix $\mathbf{\Phi} \in \mathbb{R}^{k \times p}$

$$\mathbf{\Phi} = \begin{bmatrix} \alpha_1 \cdots \alpha_1 & 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 \\ 0 \cdots 0 & \alpha_2 \cdots \alpha_2 & 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 \\ 0 \cdots 0 & 0 \cdots 0 & \alpha_3 \cdots \alpha_3 & 0 \cdots 0 & 0 \cdots 0 \\ 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 & \alpha_4 \cdots \alpha_4 & 0 \cdots 0 \\ 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 & \alpha_5 \cdots \alpha_5 \end{bmatrix}$$

Each row captures a different structure

Reduction and Low-rank Approximation



Proposed Approach

Consider fully connected neural network with H layers

Algorithm 1 Training of a Neural Network with Feature Grouping as a Stochastic Regularizer

---

Require: Learning Rate η

Require: Initial Parameters for H layers

$$\boldsymbol{\Theta} \triangleq \{\mathbf{W}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\}$$

Ensure: Generate a bank of feature grouping matrices where each is generated by randomly sampling r samples from the training data set with replacement

$$\mathit{\Phi} = \left\{\boldsymbol{\Phi}^{(1)}, \boldsymbol{\Phi}^{(2)}, \cdots, \boldsymbol{\Phi}^{(b)}\right\}$$

while stopping criteria not met **do**

Sample a minibatch of m samples from the training set $\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}\}$ with corresponding labels $y^{(i)}$

Sample $\boldsymbol{\Phi}$ from the bank $\mathit{\Phi}$.

Define $\boldsymbol{\Xi} \triangleq \left\{\hat{\mathbf{W}}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\right\}$ where $\hat{\mathbf{W}}_0 \triangleq \mathbf{W}_0\boldsymbol{\Phi}^T$.

Compute gradient estimate:

$$\mathbf{g} \leftarrow \frac{1}{m}\nabla_{\boldsymbol{\Xi}} \sum_i \mathcal{L}\left(f(\boldsymbol{\Phi}\mathbf{x}^{(i)}; \boldsymbol{\Xi}), y^{(i)}\right)$$

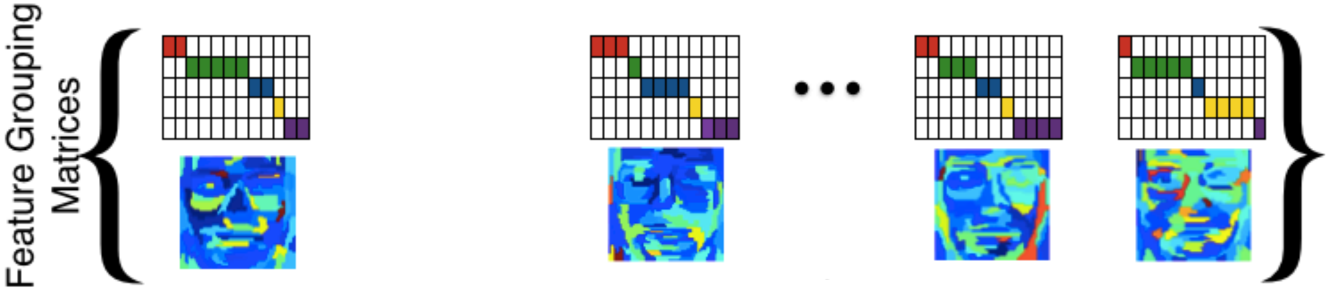
Apply updates:

- $\mathbf{W}_0 \leftarrow \mathbf{W}_0 - \eta \mathbf{g}_{\mathbf{w}_0}\boldsymbol{\Phi}$
  where $\mathbf{g}_{\mathbf{w}_0} \triangleq \frac{1}{m}\nabla_{\hat{\mathbf{W}}_0} \sum_i \mathcal{L}\left(f(\boldsymbol{\Phi}\mathbf{x}^{(i)}; \boldsymbol{\Xi}), y^{(i)}\right)$
- $\mathbf{b}_j \leftarrow \mathbf{b}_j - \eta \mathbf{g}_{b_j}$
  where $\mathbf{g}_{b_j} \triangleq \frac{1}{m}\nabla_{\mathbf{b}_j} \sum_i \mathcal{L}\left(f(\boldsymbol{\Phi}\mathbf{x}^{(i)}; \boldsymbol{\Xi}), y^{(i)}\right)$
  for $j \in \{0, \cdots, H\}$
- $\mathbf{W}_j \leftarrow \mathbf{W}_j - \eta \mathbf{g}_{\mathbf{w}_j}$
  where $\mathbf{g}_{\mathbf{w}_j} \triangleq \frac{1}{m}\nabla_{\mathbf{W}_j} \sum_i \mathcal{L}\left(f(\boldsymbol{\Phi}\mathbf{x}^{(i)}; \boldsymbol{\Xi}), y^{(i)}\right)$
  for $j \in \{1, \cdots, H\}$

end while

Proposed Approach

Pre-compute a bank of feature grouping matrices



Algorithm 1 Training of a Neural Network with Feature Grouping as a Stochastic Regularizer

Require: Learning Rate η

Require: Initial Parameters for H layers

$$\mathbf{\Theta} \triangleq \{\mathbf{W}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\}$$

Ensure: Generate a bank of feature grouping matrices where each is generated by randomly sampling r samples from the training data set with replacement

$$\varPhi = \left\{\mathbf{\Phi}^{(1)}, \mathbf{\Phi}^{(2)}, \cdots, \mathbf{\Phi}^{(b)}\right\}$$

while stopping criteria not met **do**

Sample a minibatch of m samples from the training set $\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}\}$ with corresponding labels $y^{(i)}$

Sample $\mathbf{\Phi}$ from the bank $\varPhi$.

Define $\Xi \triangleq \left\{\hat{\mathbf{W}}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\right\}$ where $\hat{\mathbf{W}}_0 \triangleq \mathbf{W}_0\mathbf{\Phi}^T$.

Compute gradient estimate:

$$\mathbf{g} \leftarrow \frac{1}{m}\nabla_{\Xi} \sum_i \mathcal{L}\left(f(\mathbf{\Phi}\mathbf{x}^{(i)}; \Xi), y^{(i)}\right)$$

Apply updates:

- $\mathbf{W}_0 \leftarrow \mathbf{W}_0 - \eta \mathbf{g}_{\mathbf{w}_0}\mathbf{\Phi}$

  where $\mathbf{g}_{\mathbf{w}_0} \triangleq \frac{1}{m}\nabla_{\hat{\mathbf{W}}_0} \sum_i \mathcal{L}\left(f(\mathbf{\Phi}\mathbf{x}^{(i)}; \Xi), y^{(i)}\right)$

- $\mathbf{b}_j \leftarrow \mathbf{b}_j - \eta \mathbf{g}_{b_j}$

  where $\mathbf{g}_{b_j} \triangleq \frac{1}{m}\nabla_{\mathbf{b}_j} \sum_i \mathcal{L}\left(f(\mathbf{\Phi}\mathbf{x}^{(i)}; \Xi), y^{(i)}\right)$ for $j \in \{0, \cdots, H\}$

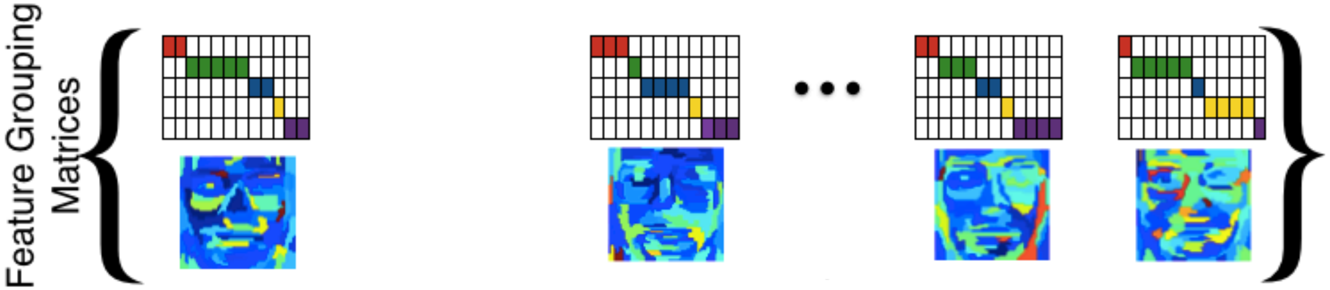
- $\mathbf{W}_j \leftarrow \mathbf{W}_j - \eta \mathbf{g}_{\mathbf{w}_j}$

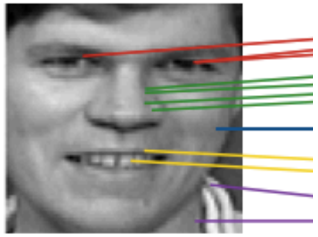
  where $\mathbf{g}_{\mathbf{w}_j} \triangleq \frac{1}{m}\nabla_{\mathbf{W}_j} \sum_i \mathcal{L}\left(f(\mathbf{\Phi}\mathbf{x}^{(i)}; \Xi), y^{(i)}\right)$ for $j \in \{1, \cdots, H\}$

end while

Proposed Approach

Sample from the training set





Algorithm 1 Training of a Neural Network with Feature Grouping as a Stochastic Regularizer

Require: Learning Rate η

Require: Initial Parameters for H layers

$$\boldsymbol{\Theta} \triangleq \{\mathbf{W}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\}$$

Ensure: Generate a bank of feature grouping matrices where each is generated by randomly sampling r samples from the training data set with replacement

$$\mathit{\Phi} = \left\{\boldsymbol{\Phi}^{(1)}, \boldsymbol{\Phi}^{(2)}, \cdots, \boldsymbol{\Phi}^{(b)}\right\}$$

while stopping criteria not met **do**

Sample a minibatch of m samples from the training set $\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}\}$ with corresponding labels $y^{(i)}$

Sample $\boldsymbol{\Phi}$ from the bank $\mathit{\Phi}$.

Define $\boldsymbol{\Xi} \triangleq \left\{\hat{\mathbf{W}}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\right\}$ where $\hat{\mathbf{W}}_0 \triangleq \mathbf{W}_0 \boldsymbol{\Phi}^T$.

Compute gradient estimate:

$$\mathbf{g} \leftarrow \frac{1}{m} \nabla_{\boldsymbol{\Xi}} \sum_i \mathcal{L}\left(f(\boldsymbol{\Phi}\mathbf{x}^{(i)}; \boldsymbol{\Xi}), y^{(i)}\right)$$

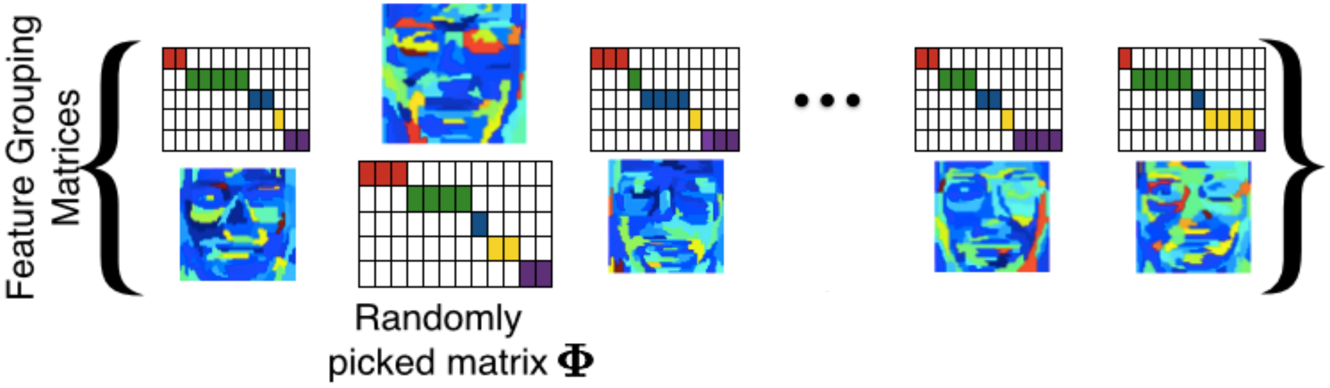
Apply updates:

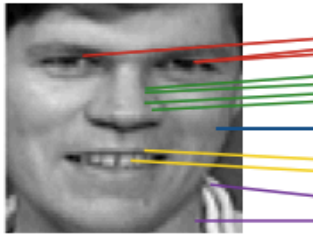
- $\mathbf{W}_0 \leftarrow \mathbf{W}_0 - \eta \mathbf{g}_{\mathbf{w}_0} \boldsymbol{\Phi}$
  where $\mathbf{g}_{\mathbf{w}_0} \triangleq \frac{1}{m} \nabla_{\hat{\mathbf{W}}_0} \sum_i \mathcal{L}\left(f(\boldsymbol{\Phi}\mathbf{x}^{(i)}; \boldsymbol{\Xi}), y^{(i)}\right)$
- $\mathbf{b}_j \leftarrow \mathbf{b}_j - \eta \mathbf{g}_{\mathbf{b}_j}$
  where $\mathbf{g}_{\mathbf{b}_j} \triangleq \frac{1}{m} \nabla_{\mathbf{b}_j} \sum_i \mathcal{L}\left(f(\boldsymbol{\Phi}\mathbf{x}^{(i)}; \boldsymbol{\Xi}), y^{(i)}\right)$
  for $j \in \{0, \cdots, H\}$
- $\mathbf{W}_j \leftarrow \mathbf{W}_j - \eta \mathbf{g}_{\mathbf{w}_j}$
  where $\mathbf{g}_{\mathbf{w}_j} \triangleq \frac{1}{m} \nabla_{\mathbf{W}_j} \sum_i \mathcal{L}\left(f(\boldsymbol{\Phi}\mathbf{x}^{(i)}; \boldsymbol{\Xi}), y^{(i)}\right)$
  for $j \in \{1, \cdots, H\}$

end while

Proposed Approach

Sample $\boldsymbol{\Phi}$ from the bank of feature grouping matrices





Algorithm 1 Training of a Neural Network with Feature Grouping as a Stochastic Regularizer

Require: Learning Rate η

Require: Initial Parameters for H layers

$$\boldsymbol{\Theta} \triangleq \{\mathbf{W}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\}$$

Ensure: Generate a bank of feature grouping matrices where each is generated by randomly sampling r samples from the training data set with replacement

$$\mathit{\Phi} = \left\{\boldsymbol{\Phi}^{(1)}, \boldsymbol{\Phi}^{(2)}, \cdots, \boldsymbol{\Phi}^{(b)}\right\}$$

1: **while** stopping criteria not met **do**

2: Sample a minibatch of m samples from the training set $\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}\}$ with corresponding labels $y^{(i)}$

3: Sample $\boldsymbol{\Phi}$ from the bank $\mathit{\Phi}$.

4: Define $\Xi \triangleq \left\{\hat{\mathbf{W}}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\right\}$ where $\hat{\mathbf{W}}_0 \triangleq \mathbf{W}_0\boldsymbol{\Phi}^T$.

5: Compute gradient estimate:

$$\mathbf{g} \leftarrow \frac{1}{m}\nabla_{\Xi}\sum_i \mathcal{L}\left(f(\boldsymbol{\Phi}\mathbf{x}^{(i)}; \Xi), y^{(i)}\right)$$

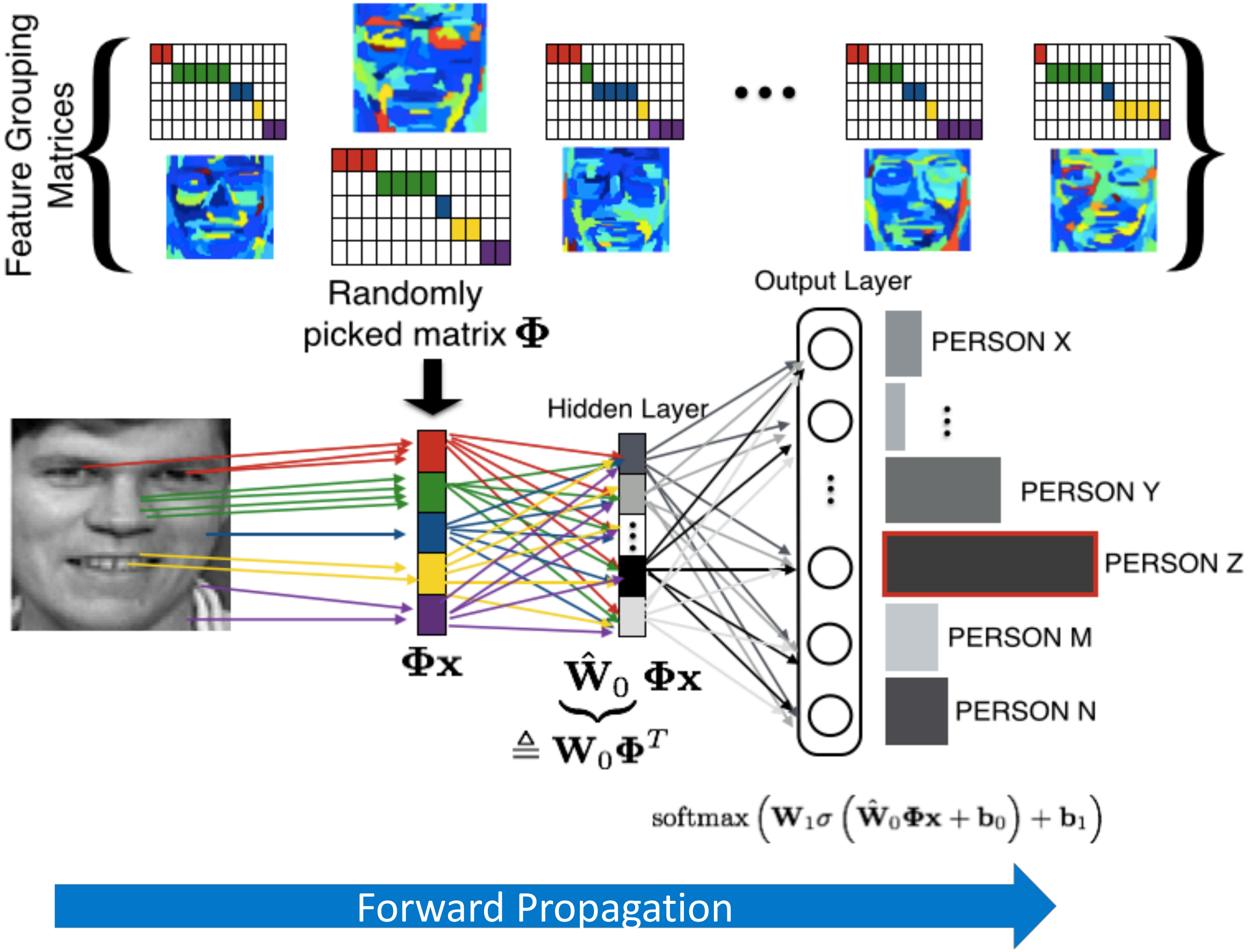
6: Apply updates:

- $\mathbf{W}_0 \leftarrow \mathbf{W}_0 - \eta\mathbf{g}_{\mathbf{w}_0}\boldsymbol{\Phi}$ where $\mathbf{g}_{\mathbf{w}_0} \triangleq \frac{1}{m}\nabla_{\hat{\mathbf{W}}_0}\sum_i \mathcal{L}\left(f(\boldsymbol{\Phi}\mathbf{x}^{(i)}; \Xi), y^{(i)}\right)$
- $\mathbf{b}_j \leftarrow \mathbf{b}_j - \eta\mathbf{g}_{\mathbf{b}_j}$ where $\mathbf{g}_{\mathbf{b}_j} \triangleq \frac{1}{m}\nabla_{\mathbf{b}_j}\sum_i \mathcal{L}\left(f(\boldsymbol{\Phi}\mathbf{x}^{(i)}; \Xi), y^{(i)}\right)$ for $j \in \{0, \cdots, H\}$
- $\mathbf{W}_j \leftarrow \mathbf{W}_j - \eta\mathbf{g}_{\mathbf{w}_j}$ where $\mathbf{g}_{\mathbf{w}_j} \triangleq \frac{1}{m}\nabla_{\mathbf{W}_j}\sum_i \mathcal{L}\left(f(\boldsymbol{\Phi}\mathbf{x}^{(i)}; \Xi), y^{(i)}\right)$ for $j \in \{1, \cdots, H\}$

7: **end while**

Proposed Approach

Re-define parameter space and project input onto lower dimensional space



Algorithm 1 Training of a Neural Network with Feature Grouping as a Stochastic Regularizer

Require: Learning Rate η

Require: Initial Parameters for H layers

$\mathbf{\Theta} \triangleq \{\mathbf{W}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\}$

Ensure: Generate a bank of feature grouping matrices where each is generated by randomly sampling r samples from the training data set with replacement

$\mathit{\Phi} = \left\{\mathbf{\Phi}^{(1)}, \mathbf{\Phi}^{(2)}, \cdots, \mathbf{\Phi}^{(b)}\right\}$

while stopping criteria not met **do**

Sample a minibatch of m samples from the training set $\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}\}$ with corresponding labels $y^{(i)}$

Sample $\mathbf{\Phi}$ from the bank $\mathit{\Phi}$.

Define $\mathbf{\Xi} \triangleq \left\{\hat{\mathbf{W}}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\right\}$ where $\hat{\mathbf{W}}_0 \triangleq \mathbf{W}_0 \mathbf{\Phi}^T$.

Compute gradient estimate:

$$\mathbf{g} \leftarrow \frac{1}{m} \nabla_{\Xi} \sum_i \mathcal{L}\left(f(\mathbf{\Phi x}^{(i)}; \Xi), y^{(i)}\right)$$

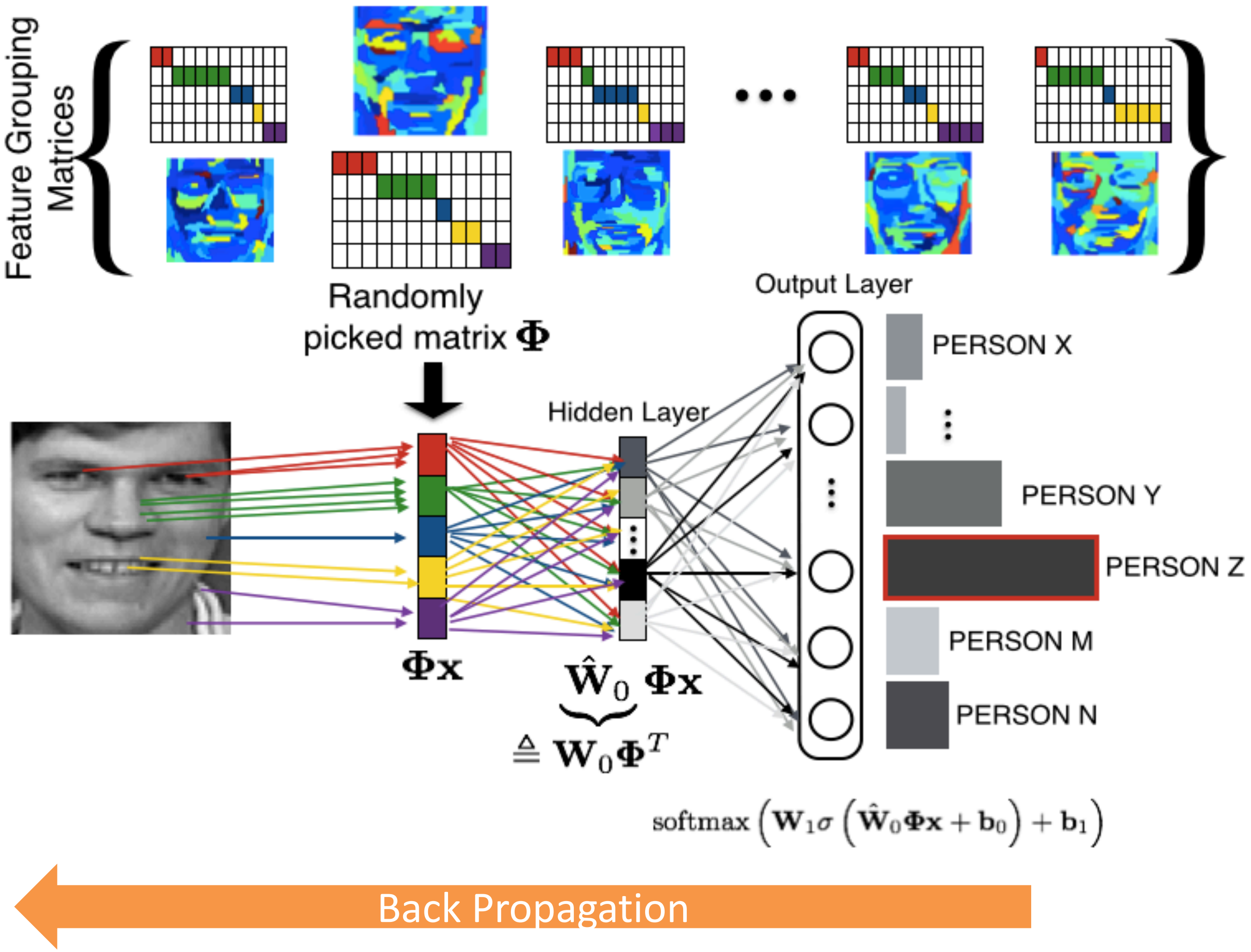
Apply updates:

- $\mathbf{W}_0 \leftarrow \mathbf{W}_0 - \eta \mathbf{g}_{\mathbf{W}_0} \mathbf{\Phi}$
  where $\mathbf{g}_{\mathbf{W}_0} \triangleq \frac{1}{m} \nabla_{\hat{\mathbf{W}}_0} \sum_i \mathcal{L}\left(f(\mathbf{\Phi x}^{(i)}; \Xi), y^{(i)}\right)$
- $\mathbf{b}_j \leftarrow \mathbf{b}_j - \eta \mathbf{g}_{\mathbf{b}_j}$
  where $\mathbf{g}_{\mathbf{b}_j} \triangleq \frac{1}{m} \nabla_{\mathbf{b}_j} \sum_i \mathcal{L}\left(f(\mathbf{\Phi x}^{(i)}; \Xi), y^{(i)}\right)$
  for $j \in \{0, \cdots, H\}$
- $\mathbf{W}_j \leftarrow \mathbf{W}_j - \eta \mathbf{g}_{\mathbf{W}_j}$
  where $\mathbf{g}_{\mathbf{W}_j} \triangleq \frac{1}{m} \nabla_{\mathbf{W}_j} \sum_i \mathcal{L}\left(f(\mathbf{\Phi x}^{(i)}; \Xi), y^{(i)}\right)$
  for $j \in \{1, \cdots, H\}$

end while

Proposed Approach

Apply back propagation



Algorithm 1 Training of a Neural Network with Feature Grouping as a Stochastic Regularizer

Require: Learning Rate η

Require: Initial Parameters for H layers

$$\mathbf{\Theta} \triangleq \{\mathbf{W}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\}$$

Ensure: Generate a bank of feature grouping matrices where each is generated by randomly sampling r samples from the training data set with replacement

$$\mathit{\Phi} = \left\{\mathbf{\Phi}^{(1)}, \mathbf{\Phi}^{(2)}, \cdots, \mathbf{\Phi}^{(b)}\right\}$$

1: **while** stopping criteria not met **do**

2: Sample a minibatch of m samples from the training set $\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}\}$ with corresponding labels $y^{(i)}$

3: Sample $\mathbf{\Phi}$ from the bank $\mathit{\Phi}$.

4: Define $\mathbf{\Xi} \triangleq \left\{\hat{\mathbf{W}}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\right\}$ where $\hat{\mathbf{W}}_0 \triangleq \mathbf{W}_0 \mathbf{\Phi}^T$.

5: Compute gradient estimate:

$$\mathbf{g} \leftarrow \frac{1}{m} \nabla_{\mathbf{\Xi}} \sum_i \mathcal{L}\left(f(\mathbf{\Phi x}^{(i)}; \mathbf{\Xi}), y^{(i)}\right)$$

6: Apply updates:

- $\mathbf{W}_0 \leftarrow \mathbf{W}_0 - \eta \mathbf{g}_{\mathbf{W}_0} \mathbf{\Phi}$
  where $\mathbf{g}_{\mathbf{W}_0} \triangleq \frac{1}{m} \nabla_{\hat{\mathbf{W}}_0} \sum_i \mathcal{L}\left(f(\mathbf{\Phi x}^{(i)}; \mathbf{\Xi}), y^{(i)}\right)$
- $\mathbf{b}_j \leftarrow \mathbf{b}_j - \eta \mathbf{g}_{\mathbf{b}_j}$
  where $\mathbf{g}_{\mathbf{b}_j} \triangleq \frac{1}{m} \nabla_{\mathbf{b}_j} \sum_i \mathcal{L}\left(f(\mathbf{\Phi x}^{(i)}; \mathbf{\Xi}), y^{(i)}\right)$
  for $j \in \{0, \cdots, H\}$
- $\mathbf{W}_j \leftarrow \mathbf{W}_j - \eta \mathbf{g}_{\mathbf{W}_j}$
  where $\mathbf{g}_{\mathbf{W}_j} \triangleq \frac{1}{m} \nabla_{\mathbf{W}_j} \sum_i \mathcal{L}\left(f(\mathbf{\Phi x}^{(i)}; \mathbf{\Xi}), y^{(i)}\right)$
  for $j \in \{1, \cdots, H\}$

7: **end while**

Proposed Approach

Update parameters

To update $\mathbf{W}_0$, project gradients back to the original space.

Other terms are updated in a standard way.

Algorithm 1 Training of a Neural Network with Feature Grouping as a Stochastic Regularizer

Require: Learning Rate η

Require: Initial Parameters for H layers

$$\mathbf{\Theta} \triangleq \{\mathbf{W}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\}$$

Ensure: Generate a bank of feature grouping matrices where each is generated by randomly sampling r samples from the training data set with replacement

$$\mathit{\Phi} = \left\{\mathbf{\Phi}^{(1)}, \mathbf{\Phi}^{(2)}, \cdots, \mathbf{\Phi}^{(b)}\right\}$$

1: **while** stopping criteria not met **do**

2: Sample a minibatch of m samples from the training set $\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}\}$ with corresponding labels $y^{(i)}$

3: Sample $\mathbf{\Phi}$ from the bank $\mathit{\Phi}$.

4: Define $\mathbf{\Xi} \triangleq \left\{\hat{\mathbf{W}}_0, \mathbf{b}_0, \mathbf{W}_1, \mathbf{b}_1, \cdots, \mathbf{W}_H, \mathbf{b}_H\right\}$ where $\hat{\mathbf{W}}_0 \triangleq \mathbf{W}_0\mathbf{\Phi}^T$.

5: Compute gradient estimate:

$$\mathbf{g} \leftarrow \frac{1}{m}\nabla_{\mathbf{\Xi}} \sum_i \mathcal{L}\left(f(\mathbf{\Phi}\mathbf{x}^{(i)}; \mathbf{\Xi}), y^{(i)}\right)$$

6: Apply updates:

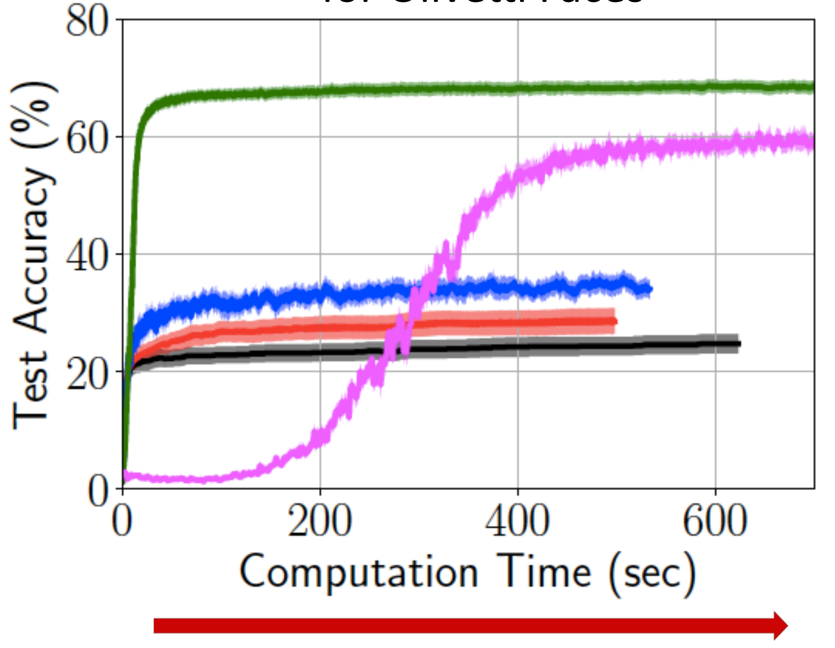
- $\mathbf{W}_0 \leftarrow \mathbf{W}_0 - \eta \mathbf{g}_{\mathbf{w}_0}\mathbf{\Phi}$ where $\mathbf{g}_{\mathbf{w}_0} \triangleq \frac{1}{m}\nabla_{\hat{\mathbf{W}}_0} \sum_i \mathcal{L}\left(f(\mathbf{\Phi}\mathbf{x}^{(i)}; \mathbf{\Xi}), y^{(i)}\right)$
- $\mathbf{b}_j \leftarrow \mathbf{b}_j - \eta \mathbf{g}_{\mathbf{b}_j}$ where $\mathbf{g}_{\mathbf{b}_j} \triangleq \frac{1}{m}\nabla_{\mathbf{b}_j} \sum_i \mathcal{L}\left(f(\mathbf{\Phi}\mathbf{x}^{(i)}; \mathbf{\Xi}), y^{(i)}\right)$ for $j \in \{0, \cdots, H\}$
- $\mathbf{W}_j \leftarrow \mathbf{W}_j - \eta \mathbf{g}_{\mathbf{w}_j}$ where $\mathbf{g}_{\mathbf{w}_j} \triangleq \frac{1}{m}\nabla_{\mathbf{W}_j} \sum_i \mathcal{L}\left(f(\mathbf{\Phi}\mathbf{x}^{(i)}; \mathbf{\Xi}), y^{(i)}\right)$ for $j \in \{1, \cdots, H\}$

7: **end while**

Experimental Results

Noisy Settings

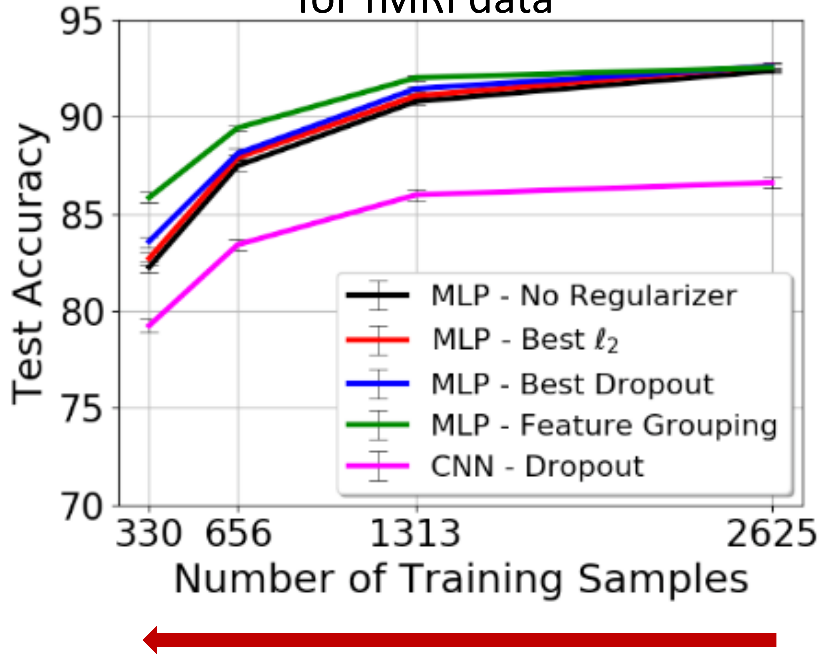
Performance in terms of computation time
for Olivetti Faces



Feature Grouping
is computationally efficient and robust to noise

Small-sample Settings

Performance in terms of sample size
for fMRI data



Feature Grouping
performs best as the sample size decreases

Thank You!

Visit our POSTER TODAY at Pacific Ballroom #121!