PhD Prospectus
Efficient Hyper-Parameter Search for Hardware-Aware Optimization of Deep Learning Systems

Akshay Chandrashekaran

Carnegie Mellon University
Department of Electrical and Computer Engineering
Proposal Chair: Associate Professor Ian Lane

May 30, 2017

Electrical & Computer ENGINEERING

Outline

Electrical & Computer ENGINEERING

Outline

Electrical & Computer ENGINEERING

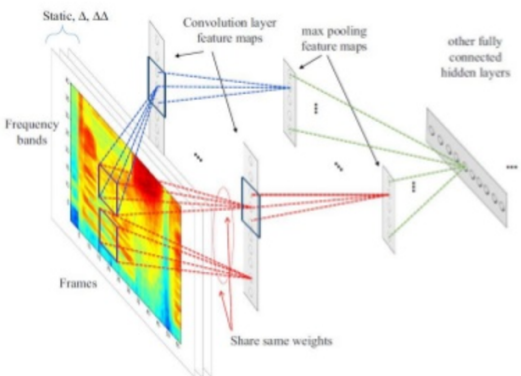
Deep Learning for Perceptual Computing

- Deep Learning is currently state-of-the-art for many machine learning tasks

Electrical & Computer ENGINEERING

Deep Learning for Perceptual Computing

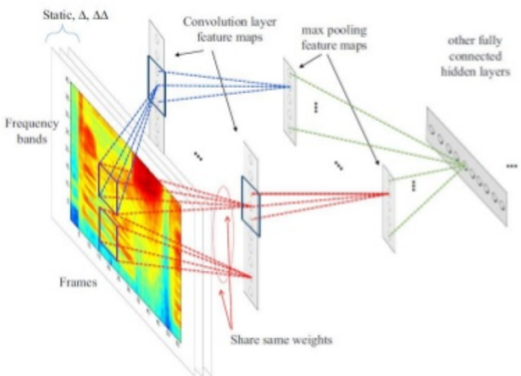
- Deep Learning is currently state-of-the-art for many machine learning tasks

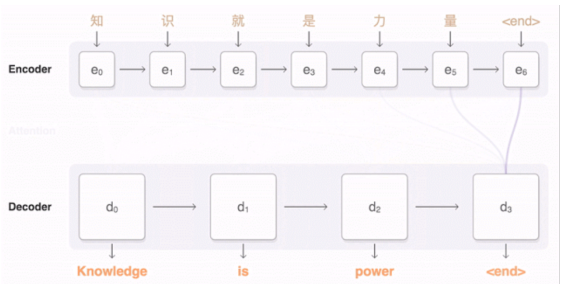


Electrical & Computer ENGINEERING

Deep Learning for Perceptual Computing

- Deep Learning is currently state-of-the-art for many machine learning tasks

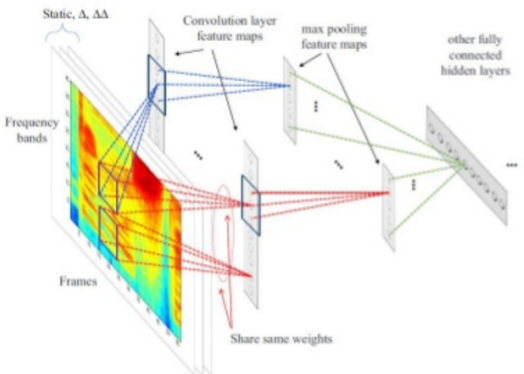


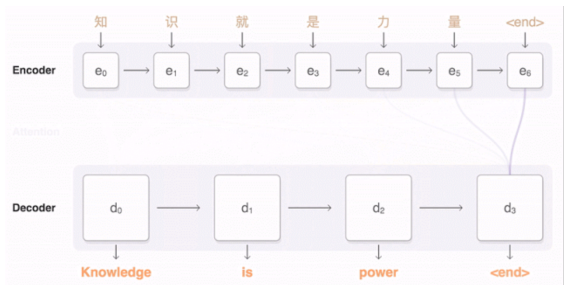


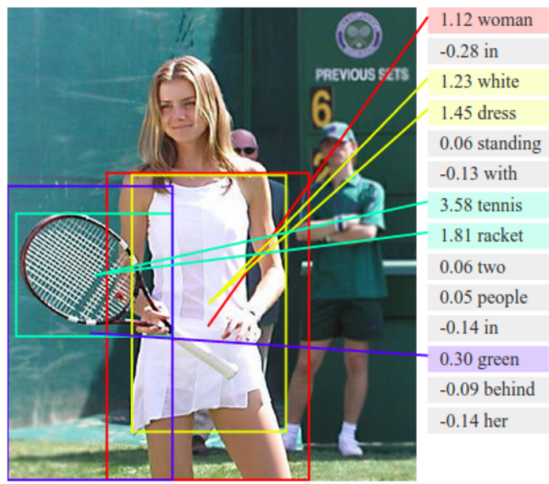
Electrical & Computer ENGINEERING

Deep Learning for Perceptual Computing

- Deep Learning is currently state-of-the-art for many machine learning tasks







Electrical & Computer ENGINEERING

Deep Learning - State of the Art

Task: Speech Recognition

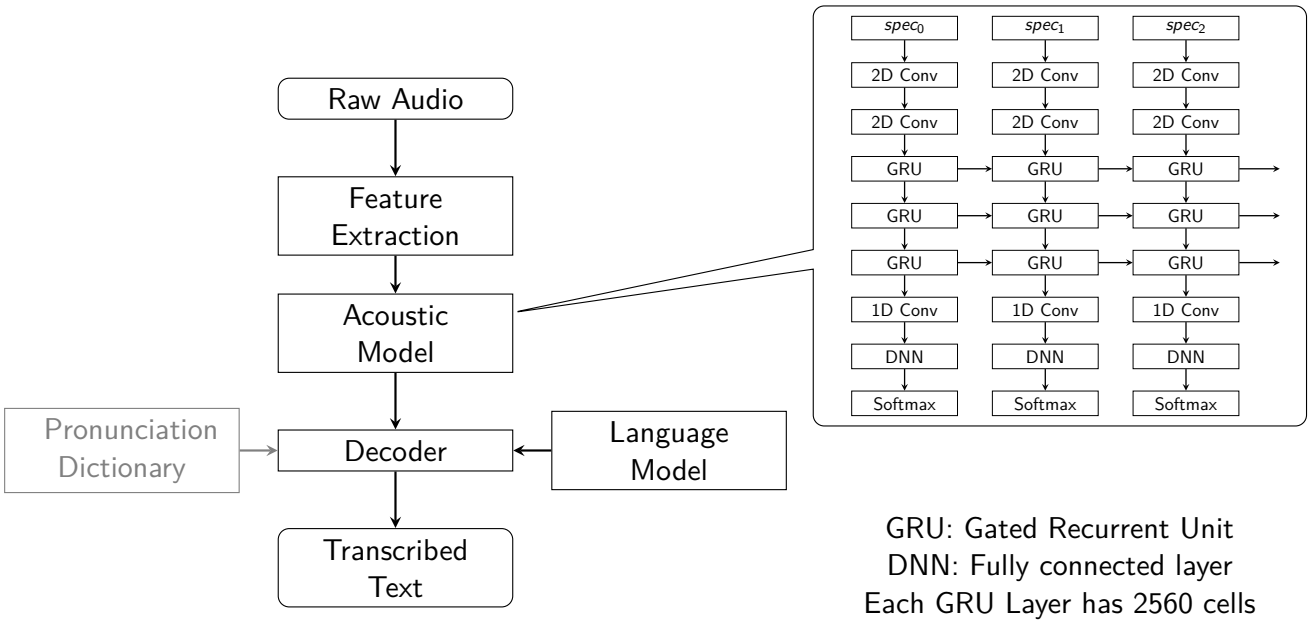


Figure: Deep Speech 2 ASR system architecture

Number of floating point operations per speech frame: 0.15 GFlops

Dario Amodei et al. "Deep speech 2: End-to-end speech recognition in english and mandarin". In:arXiv preprint arXiv:1512.02595(2015)

Electrical & Computer ENGINEERING

Deep Learning - State of the Art

Task: Machine Translation

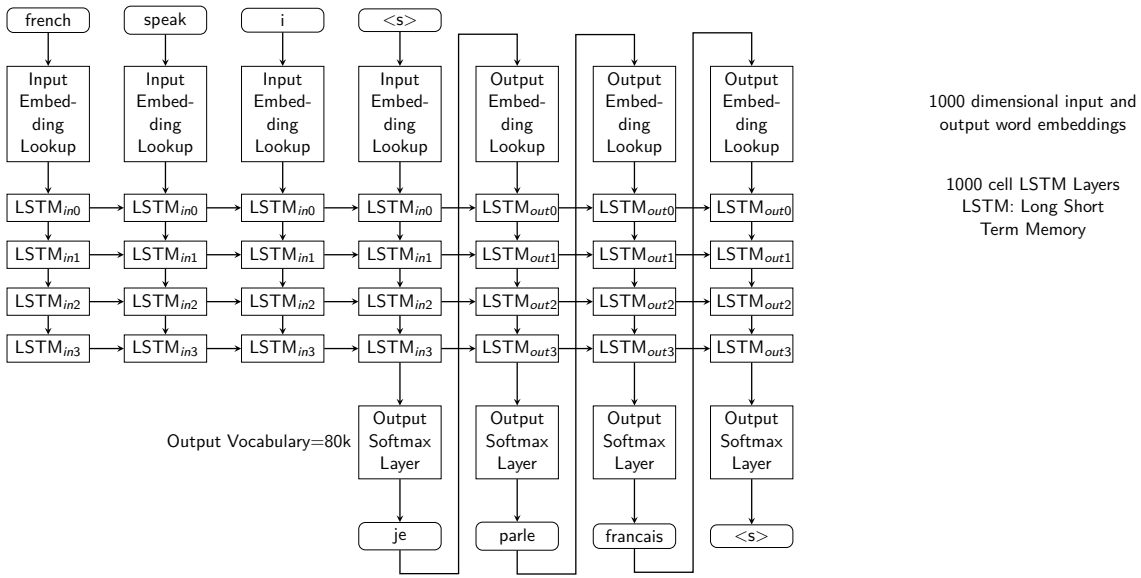


Figure: Machine Translation LSTM model architecture

Number of floating point operations per word: 0.216 GFlops

Sutskever, Ilya, Vinyals, Oriol , and Le, Quoc V. . "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

Electrical & Computer ENGINEERING

Deep Learning - State of the Art

Task: Computer Vision

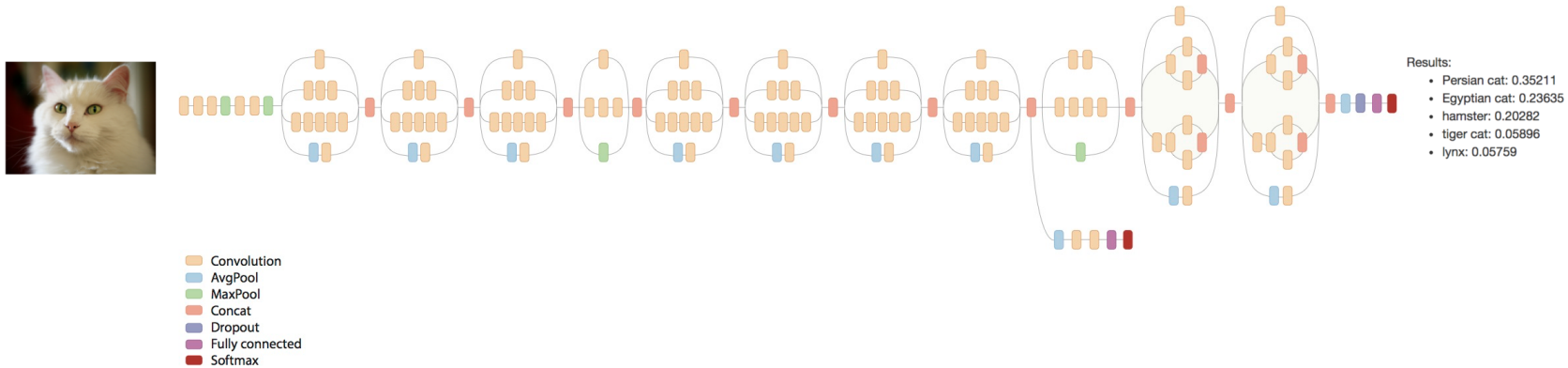


Figure: Inception Neural Network Architecture for the ImageNet classification task

Number of floating point operations per image: 5 GFlops

Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." arXiv preprint arXiv:1512.00567 (2015).

Electrical & Computer ENGINEERING

Current Approach to Designing Deep Learning Systems

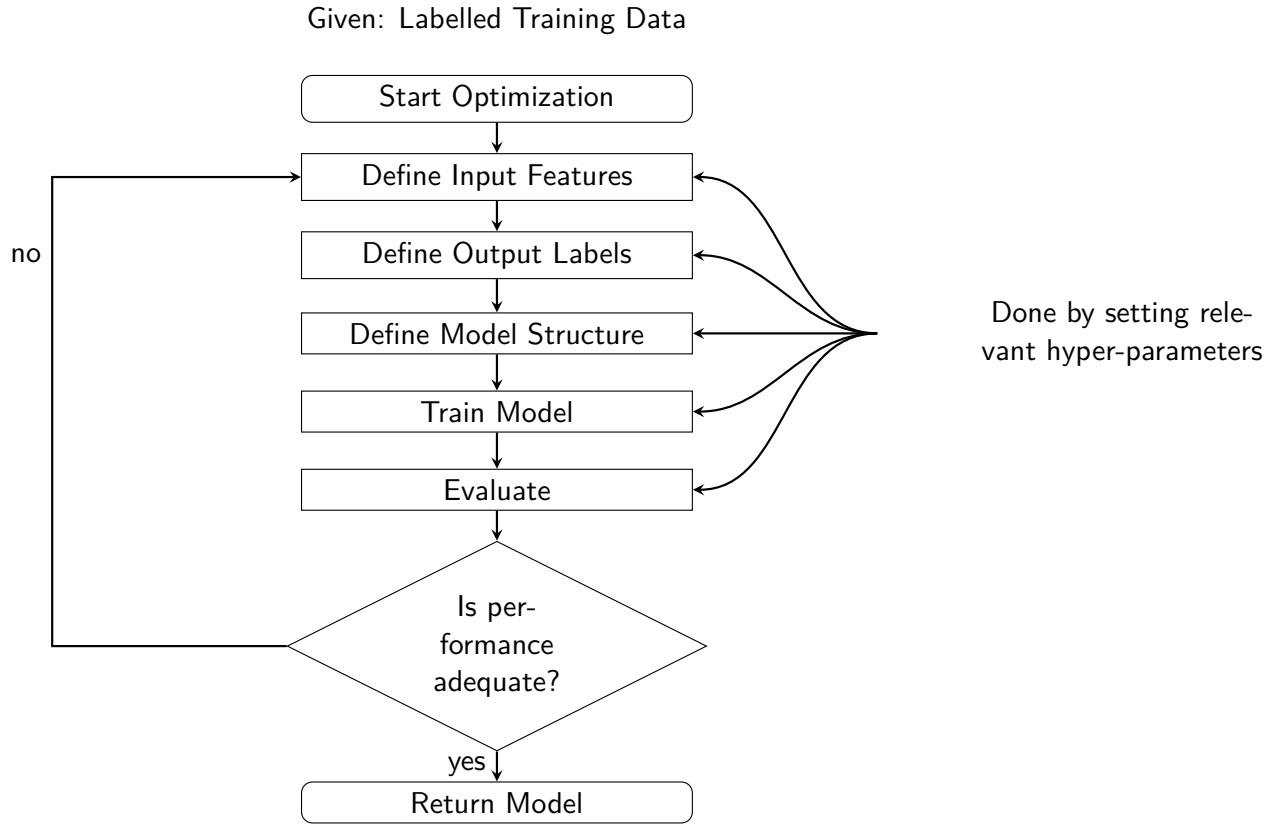


Figure: Current Design Methodology for training deep learning models

Electrical & Computer ENGINEERING

Hyper-parameters

- Set of "bells and whistles" that determine ML model architecture and performance
- Categorical/Integer Valued/Real Valued
- Come into effect at various stages of the training and inference process.

Electrical & Computer ENGINEERING

Hyper-parameters

- Example: K nearest neighbors
- Hyper-parameters
  - Number of neighbors.
  - Number of elements in training data.
  - Distance metric.

Electrical & Computer ENGINEERING

Cloud Computing Architecture for Deep Learning

- Large input features
- Big deep models
- Optimized to run on servers

Current Approach: Cloud

Benefits:

- Large compute capacity
- Large memory capacity

Issues:

- Network Coverage
- Network Latency
- Privacy concerns

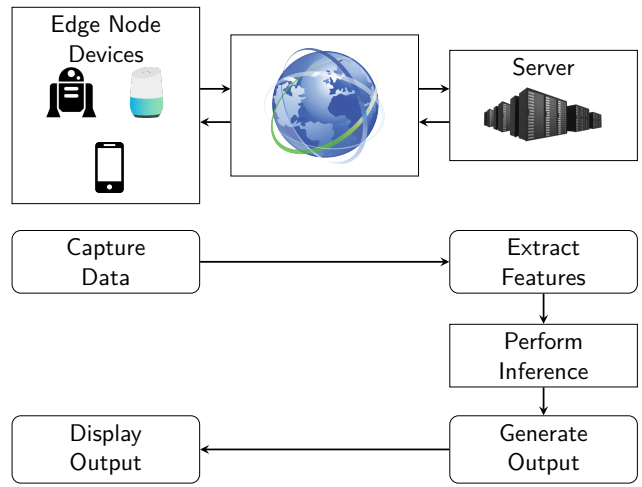


Figure: Deep learning inference cloud computing architecture

Electrical & Computer ENGINEERING

Cloud Computing Architecture for Deep Learning

- Large input features
- Big deep models
- Optimized to run on servers

Current Approach: Cloud

Benefits:

- Large compute capacity
- Large memory capacity

Issues:

- Network Coverage
- Network Latency
- Privacy concerns

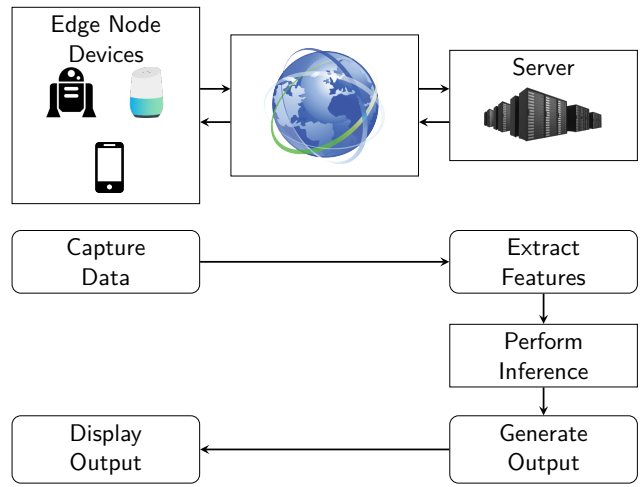


Figure: Deep learning inference cloud computing architecture

Electrical & Computer ENGINEERING

Cloud Computing Architecture for Deep Learning

- Large input features
- Big deep models
- Optimized to run on servers

Current Approach: Cloud

Benefits:

- Large compute capacity
- Large memory capacity

Issues:

- Network Coverage
- Network Latency
- Privacy concerns

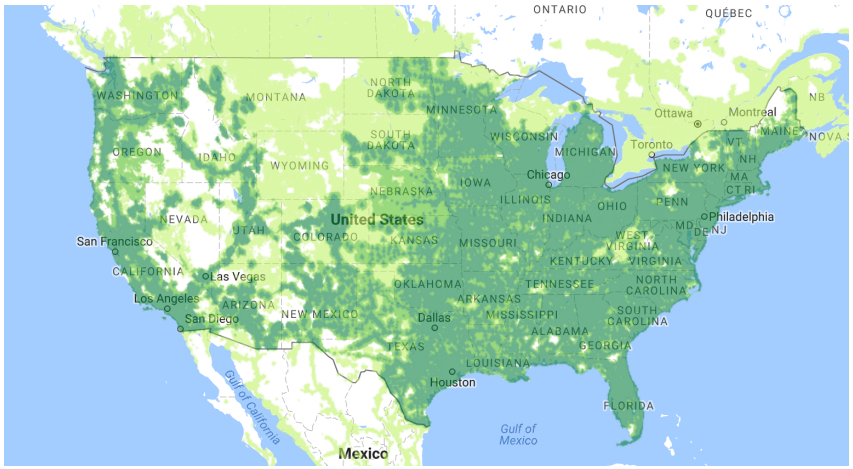


Figure: Project Fi network coverage

https://fi.google.com/coverage?u=0

Electrical & Computer ENGINEERING

Cloud Computing Architecture for Deep Learning

- Large input features
- Big deep models
- Optimized to run on servers

Current Approach: Cloud

Benefits:

- Large compute capacity
- Large memory capacity

Issues:

- Network Coverage
- Network Latency
- Privacy concerns

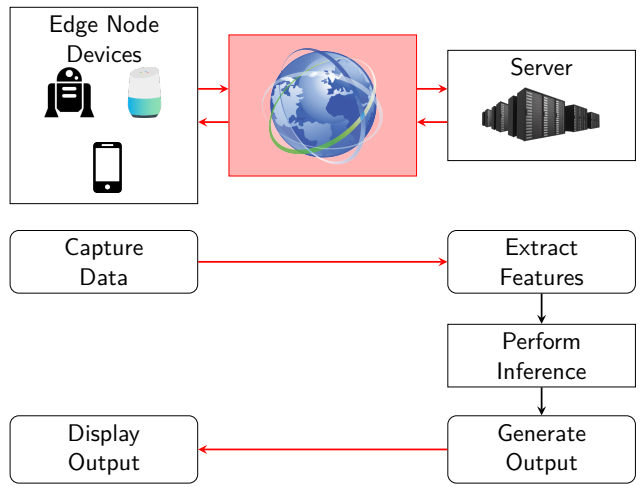


Figure: Latency in cloud computing architecture

Electrical & Computer ENGINEERING

Cloud Computing Architecture for Deep Learning

- Large input features
- Big deep models
- Optimized to run on servers

Current Approach: Cloud

Benefits:

- Large compute capacity
- Large memory capacity

Issues:

- Network Coverage
- Network Latency
- Privacy concerns

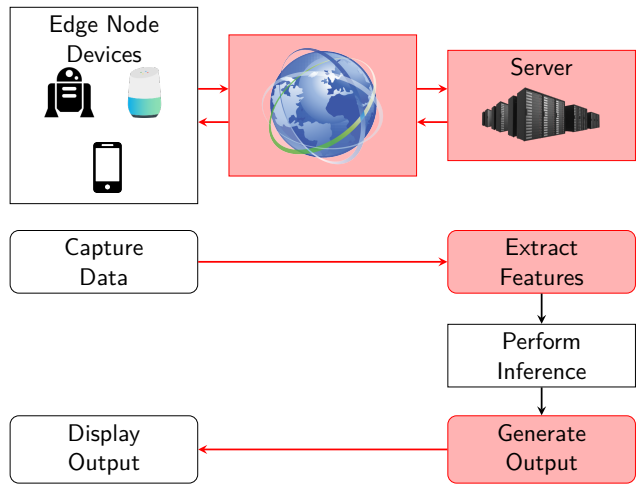


Figure: Privacy in cloud computing architecture

Electrical & Computer ENGINEERING

Edge Computing Architecture for Deep Learning

- Pushes computing to edge of networks.
- Processing and inference happens at data source.
- Advantages:
  - No network latency.
  - Higher degree of privacy.
- Disadvantages:
  - Constrained compute capacity.
  - Constrained memory capacity.

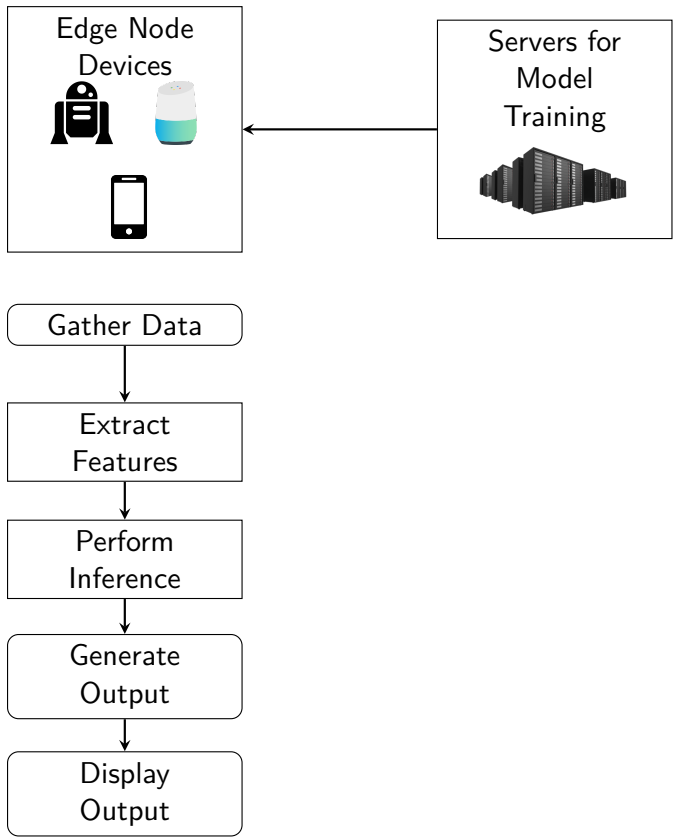


Figure: Deep learning inference in edge computing architecture

Electrical & Computer ENGINEERING

Edge Computing Architecture for Deep Learning

- Pushes computing to edge of networks.
- Processing and inference happens at data source.
- Advantages:
  - No network latency.
  - Higher degree of privacy.
- Disadvantages:
  - Constrained compute capacity.
  - Constrained memory capacity.

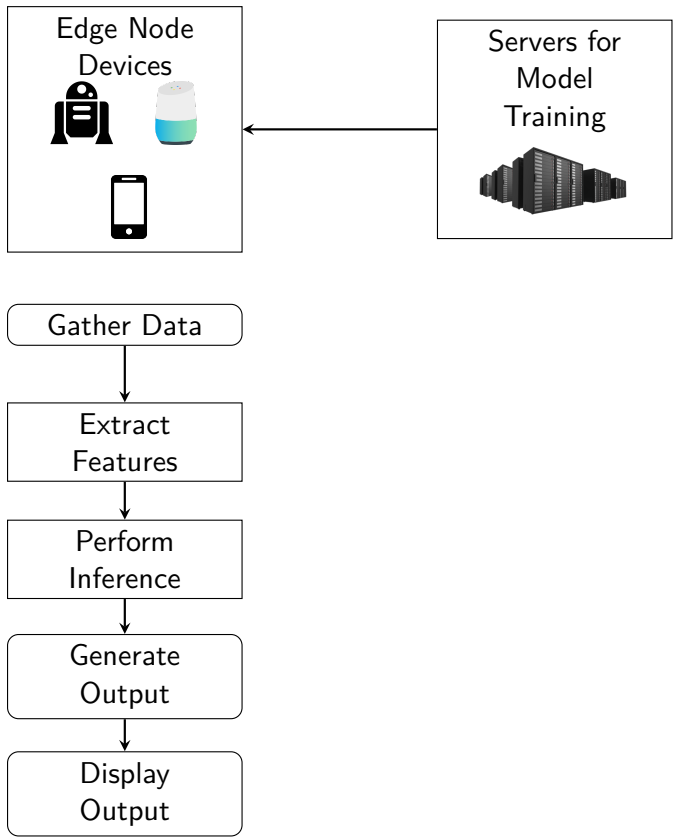


Figure: Deep learning inference in edge computing architecture

Electrical & Computer ENGINEERING

Edge Computing Architecture for Deep Learning

- Pushes computing to edge of networks.
- Processing and inference happens at data source.
- Advantages:
  - No network latency.
  - Higher degree of privacy.
- Disadvantages:
  - Constrained compute capacity.
  - Constrained memory capacity.

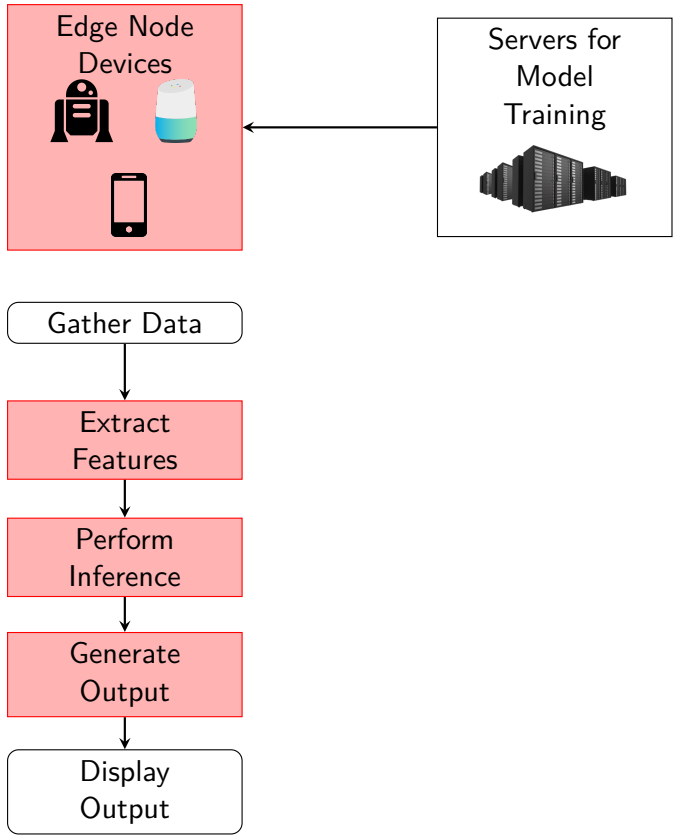


Figure: Disadvantages of edge computing

Electrical & Computer ENGINEERING

Problem Statement

- Given: Training data, validation data, evaluation data, Edge Device.
- Problem:
  - Design a neural network capable of running on the edge device, but with similar accuracy compared to cloud system.
- Approach:
  - Treat the hyper-parameter selection as a search process.
  - Design an efficient algorithm to optimize hyper-parameters for a given hardware.

Electrical & Computer ENGINEERING

Problem Statement

- Given: Training data, validation data, evaluation data, Edge Device.
- Problem:
  - Design a neural network capable of running on the edge device, but with similar accuracy compared to cloud system.
- Approach:
  - Treat the hyper-parameter selection as a search process.
  - Design an efficient algorithm to optimize hyper-parameters for a given hardware.

Electrical & Computer ENGINEERING

Problem Statement

- Given: Training data, validation data, evaluation data, Edge Device.
- Problem:
  - Design a neural network capable of running on the edge device, but with similar accuracy compared to cloud system.
- Approach:
  - Treat the hyper-parameter selection as a search process.
  - Design an efficient algorithm to optimize hyper-parameters for a given hardware.

Electrical & Computer ENGINEERING

Problem Statement

- Given: Training data, validation data, evaluation data, Edge Device.
- Problem:
  - Design a neural network capable of running on the edge device, but with similar accuracy compared to cloud system.
- Approach:
  - Treat the hyper-parameter selection as a search process.
  - Design an efficient algorithm to optimize hyper-parameters for a given hardware.

Electrical & Computer ENGINEERING

Hardware Awareness

- DL Model training happens on servers
- Performance metric can be computed on server
- Computational metrics obtained by performing inference on edge device.

Electrical & Computer ENGINEERING

Challenges

- Need to optimize multiple objectives
- Need to perform efficient search
  - Over the hyper-parameter search space
  - Reduce the wall-time for optimization

Electrical & Computer ENGINEERING

Challenges

- Need to optimize multiple objectives
- Need to perform efficient search
  - Over the hyper-parameter search space
  - Reduce the wall-time for optimization

Electrical & Computer ENGINEERING

Challenges

- Need to optimize multiple objectives
- Need to perform efficient search
  - Over the hyper-parameter search space
  - Reduce the wall-time for optimization

Electrical & Computer ENGINEERING

Challenges

- Need to optimize multiple objectives
- Need to perform efficient search
  - Over the hyper-parameter search space
  - Reduce the wall-time for optimization

Electrical & Computer ENGINEERING

Proposed Solutions

- Usage of Multi-Objective Hyper-parameter Optimization
- Hierarchical Hyper-parameter Optimization
- Early Stopping Techniques

Electrical & Computer ENGINEERING

Proposed Solutions

- Usage of Multi-Objective Hyper-parameter Optimization
- Hierarchical Hyper-parameter Optimization
- Early Stopping Techniques

Electrical & Computer ENGINEERING

Proposed Solutions

- Usage of Multi-Objective Hyper-parameter Optimization
- Hierarchical Hyper-parameter Optimization
- Early Stopping Techniques

Electrical & Computer ENGINEERING

Outline

Electrical & Computer ENGINEERING

Multi-Objective Optimization

Given multiple objectives $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ...f_M(\boldsymbol{x})]$

Find $\boldsymbol{x}^* = \text{argmin}_{\mathbf{x}}[f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ...f_M(\boldsymbol{x})]$

Electrical & Computer ENGINEERING

Pareto Front

- Non-Dominated Solution:$\boldsymbol{x}_1 \preceq \boldsymbol{x}_2$ if
  - $f_i(\boldsymbol{x}_1) \leq f_i(\boldsymbol{x}_2)\ \forall i = 1 : M$
  - $f_i(\boldsymbol{x}_1) < f_i(\boldsymbol{x}_2)$ for at least one $i = 1 : M$
- Pareto Front: Set of all points in the hyper-parameter search space that are not dominated by the other points in the search space.

Electrical & Computer ENGINEERING

Multi-Objective Optimization

- Type I: Formulation all but one of the objectives as constraints
- Type II: Simultaneous optimization of all objectives

Electrical & Computer ENGINEERING

Multi-Objective Optimization

- Type I: Formulation all but one of the objectives as constraints
- Type II: Simultaneous optimization of all objectives

Electrical & Computer ENGINEERING

Constrained Multi-Objective Optimization

- Constrained Random Sampling
- Constrained Evolutionary Algorithms
  - Constrained Genetic Algorithms
- Constrained Bayesian Optimization

Electrical & Computer ENGINEERING

Constrained Multi-Objective Optimization

- Constrained Random Sampling
- Constrained Evolutionary Algorithms
  - Constrained Genetic Algorithms
- **Constrained Bayesian Optimization**

Electrical & Computer ENGINEERING

Pareto Front based Multi-Objective Optimization

- Random Sampling
- Evolutionary Algorithms
  - Non Dominated Sort Genetic Algorithms-II
  - Covariance Mean Adaptation - Evolutionary Strategy
- ParEGO

Electrical & Computer ENGINEERING

Pareto Front based Multi-Objective Optimization

- Random Sampling
- Evolutionary Algorithms
  - Non Dominated Sort Genetic Algorithms-II
  - Covariance Mean Adaptation - Evolutionary Strategy
- **ParEGO**

Electrical & Computer ENGINEERING

Multi-Objective Optimization

- For Type I Problems: Constrained Bayesian Optimization
- For Type II Problems: ParEGO

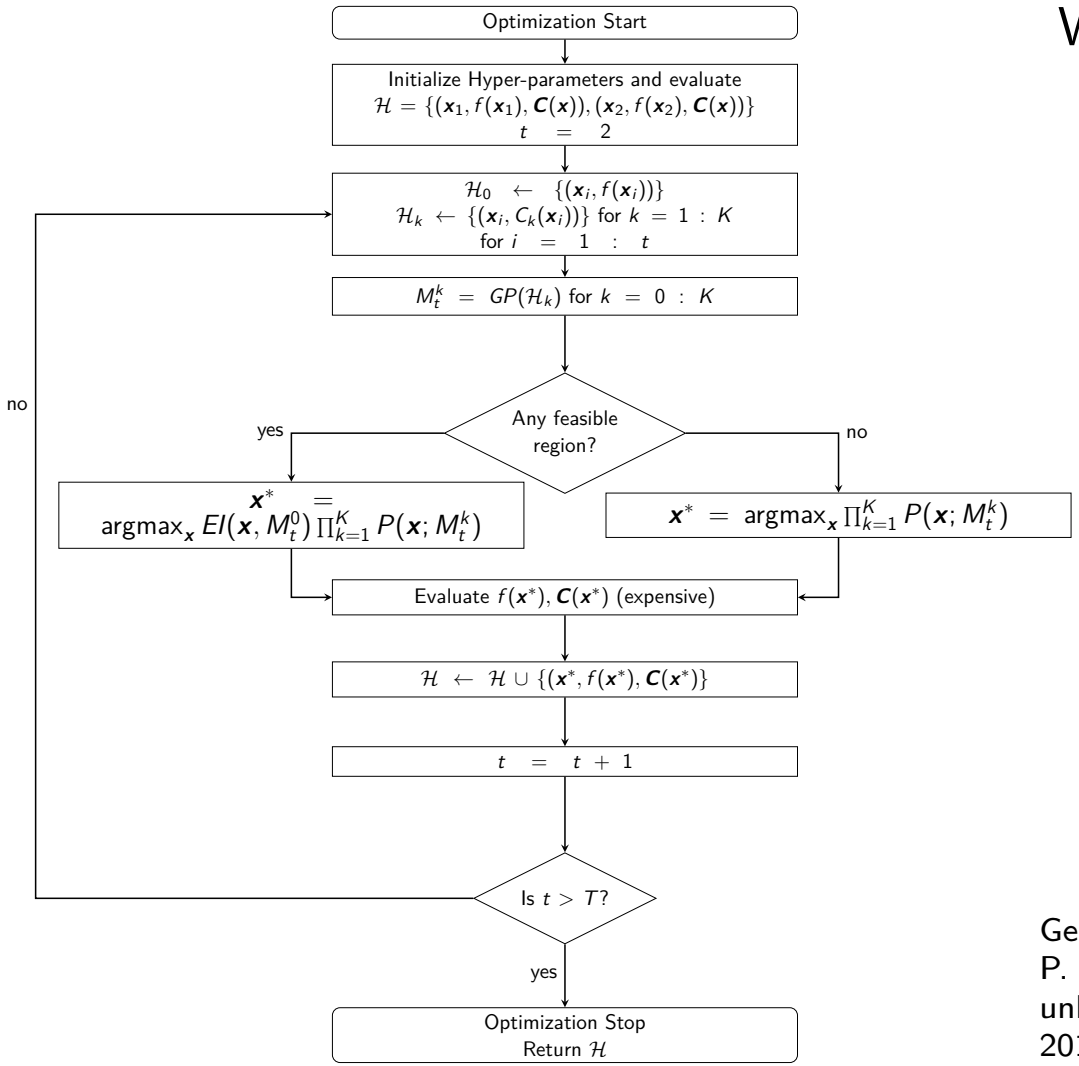
Electrical & Computer ENGINEERING

Multi-Objective Optimization

- For Type I Problems: Constrained Bayesian Optimization
- For Type II Problems: ParEGO

Electrical & Computer ENGINEERING

Constrained Bayesian Optimization



Where:

- $\mathcal{H}$: set of all evaluated points
- $GP(.)$: Gaussian Process
- EI: Expected improvement
- $C_k(\boldsymbol{x})$: k^{th} constraint
- $P(\boldsymbol{x}; M_t^k)$: Probability of improvement of k^{th} constraint

Gelbart, Michael A., Jasper Snoek, and Ryan P. Adams. "Bayesian optimization with unknown constraints." Proceedings of UAI 2014.

Electrical & Computer ENGINEERING

ParEGO

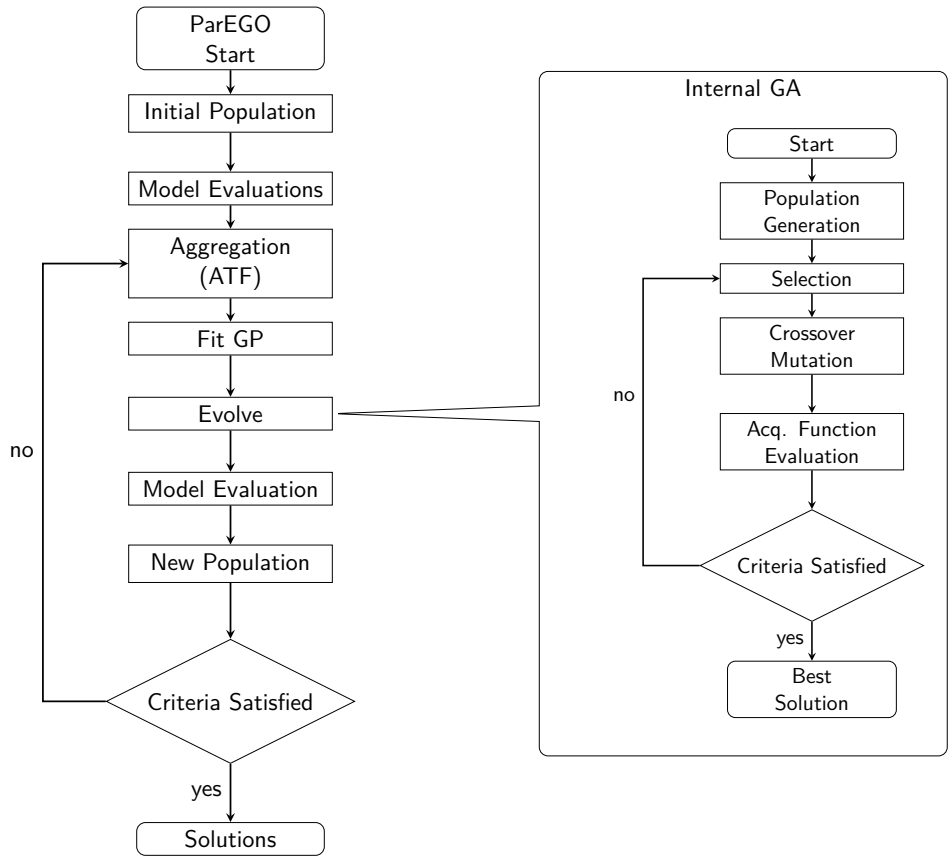


Figure: ParEGO Pipeline

Cristescu, Cristina, and Joshua Knowles. "Surrogate-Based Multiobjective Optimization: ParEGO Update and Test."

Electrical & Computer ENGINEERING

Comparison of Constrained Optimization Techniques

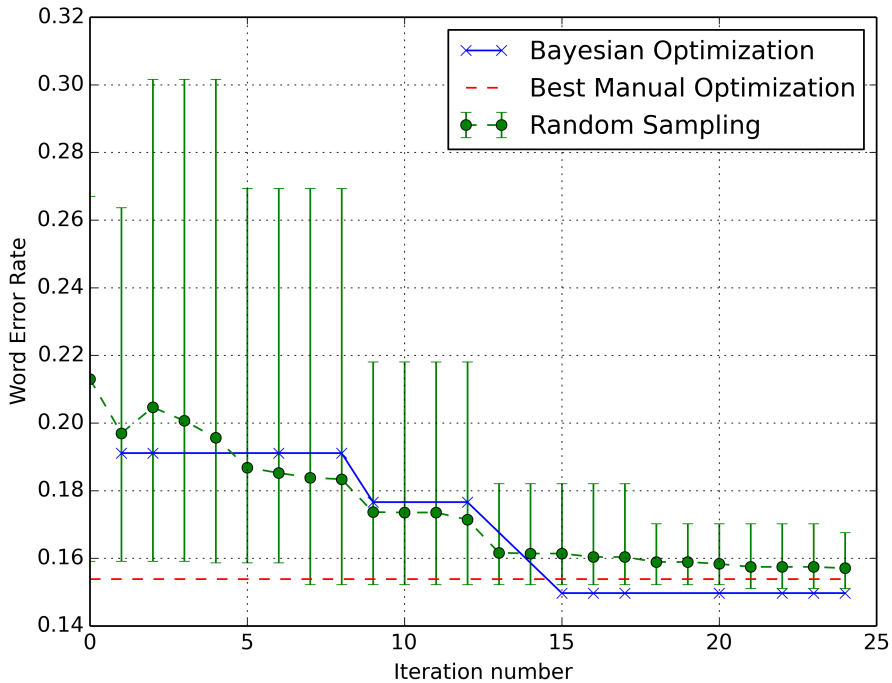


Figure: Comparison of constrained hyper-parameter optimization techniques for decoder hyper-parameters for speech recognition

Chandrashekaran, Akshay, et al."Automated optimization of decoder hyper-parameters for online LVCSR." IEEE Workshop on Spoken Language Technologies (2016).

Electrical & Computer ENGINEERING

Outline

Electrical & Computer ENGINEERING

Recall

Need to perform efficient search

- Over the hyper-parameter search space
- Reduce the wall-time for optimization

Electrical & Computer ENGINEERING

Outline

Electrical & Computer ENGINEERING

Motivation

- Too many hyper-parameters in training a Deep Learning system.
- Optimizing all simultaneously will result in slow convergence.
- Hyper-parameters can be *chunked* based on where they are utilized.
  - Training Hyper-parameters.
  - Inference Hyper-parameters.

Electrical & Computer ENGINEERING

Motivation

- Too many hyper-parameters in training a Deep Learning system.
- Optimizing all simultaneously will result in slow convergence.
- Hyper-parameters can be *chunked* based on where they are utilized.
  - Training Hyper-parameters.
  - Inference Hyper-parameters.

Electrical & Computer ENGINEERING

Motivation

- Too many hyper-parameters in training a Deep Learning system.
- Optimizing all simultaneously will result in slow convergence.
- Hyper-parameters can be *chunked* based on where they are utilized.
  - Training Hyper-parameters.
  - Inference Hyper-parameters.

Electrical & Computer ENGINEERING

Current Approach

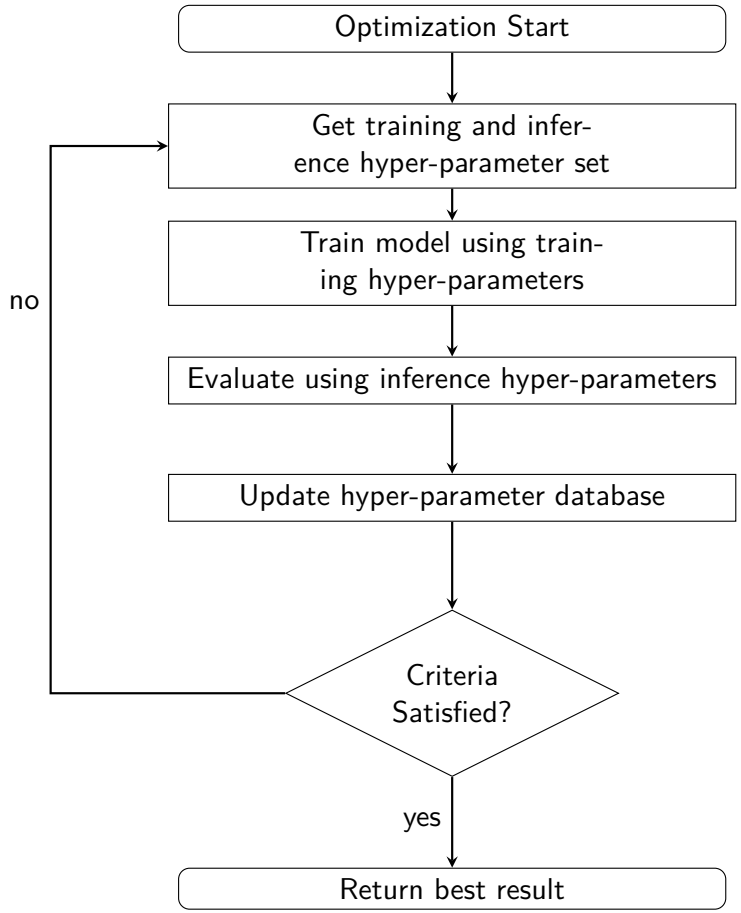


Figure: Flow chart for existing hyper-parameter optimization techniques

Electrical & Computer ENGINEERING

Proposed Approach: Hierarchical Optimization

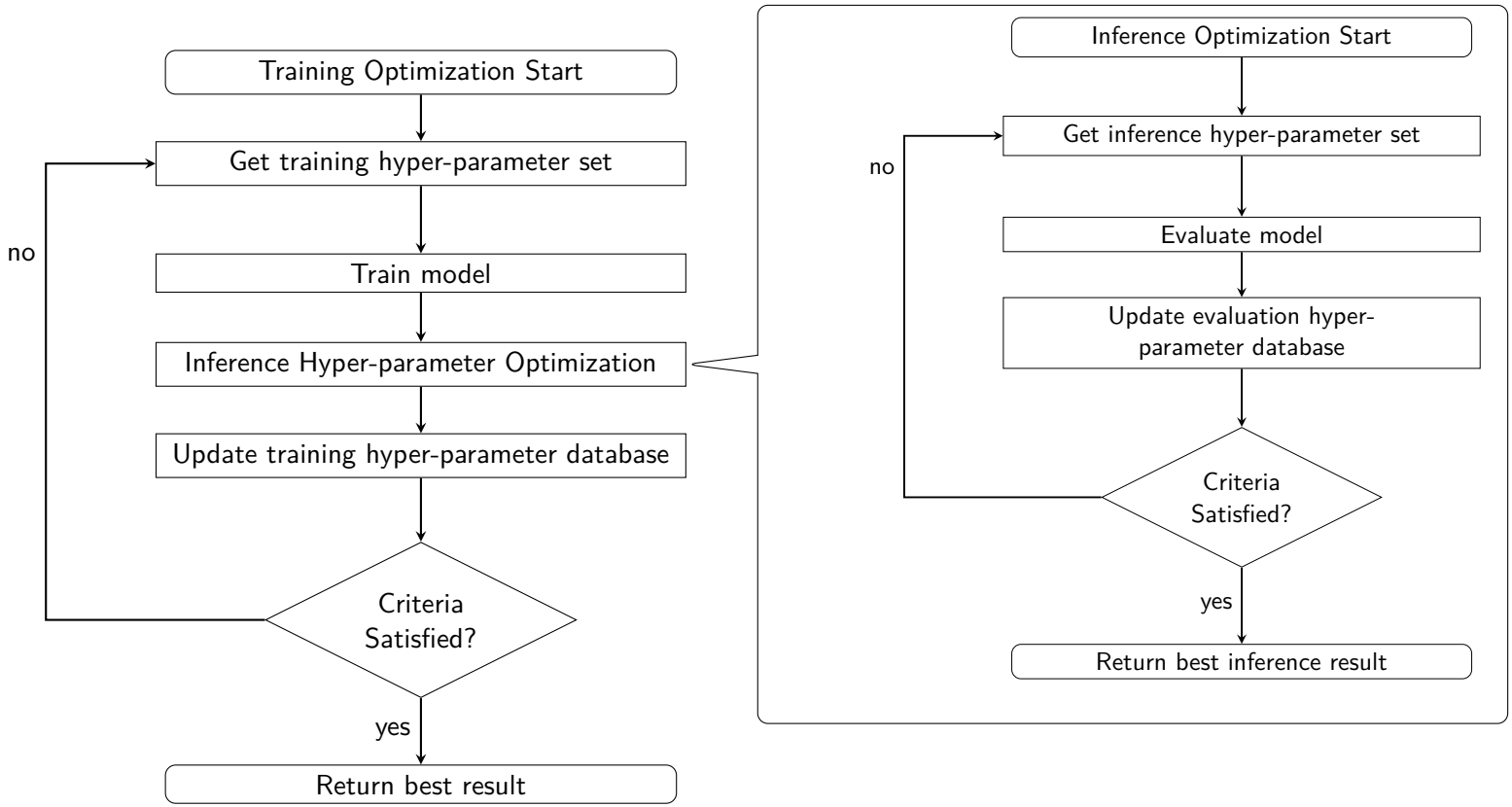


Figure: Flow chart for proposed hierarchical optimization technique

Electrical & Computer ENGINEERING

Hierarchical Constrained Bayesian Optimization

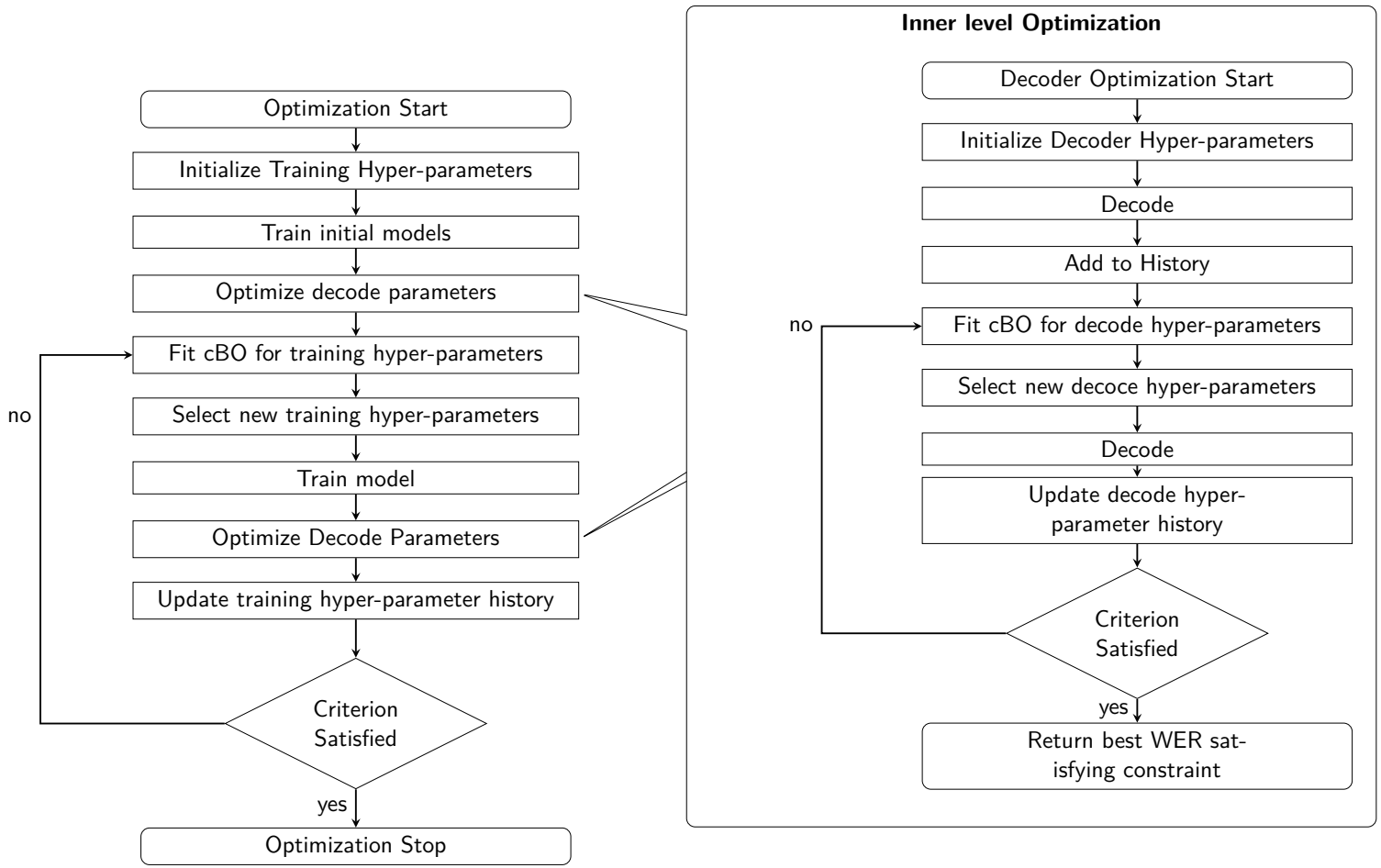


Figure: Hierarchical Constrained Bayesian Optimization Flow Chart

Electrical & Computer ENGINEERING

Outline

Electrical & Computer ENGINEERING

Motivation

- Neural Network training takes a large number of epochs to converge.
- Training poor networks till convergence $\rightarrow$ Waste of time.
- Find a way to terminate poor models.

Electrical & Computer ENGINEERING

Motivation

- Neural Network training takes a large number of epochs to converge.
- Training poor networks till convergence $\rightarrow$ Waste of time.
- Find a way to terminate poor models.

Electrical & Computer ENGINEERING

Motivation

- Neural Network training takes a large number of epochs to converge.
- Training poor networks till convergence $\rightarrow$ Waste of time.
- Find a way to terminate poor models.

Electrical & Computer ENGINEERING

Extrapolation of Learning Curves

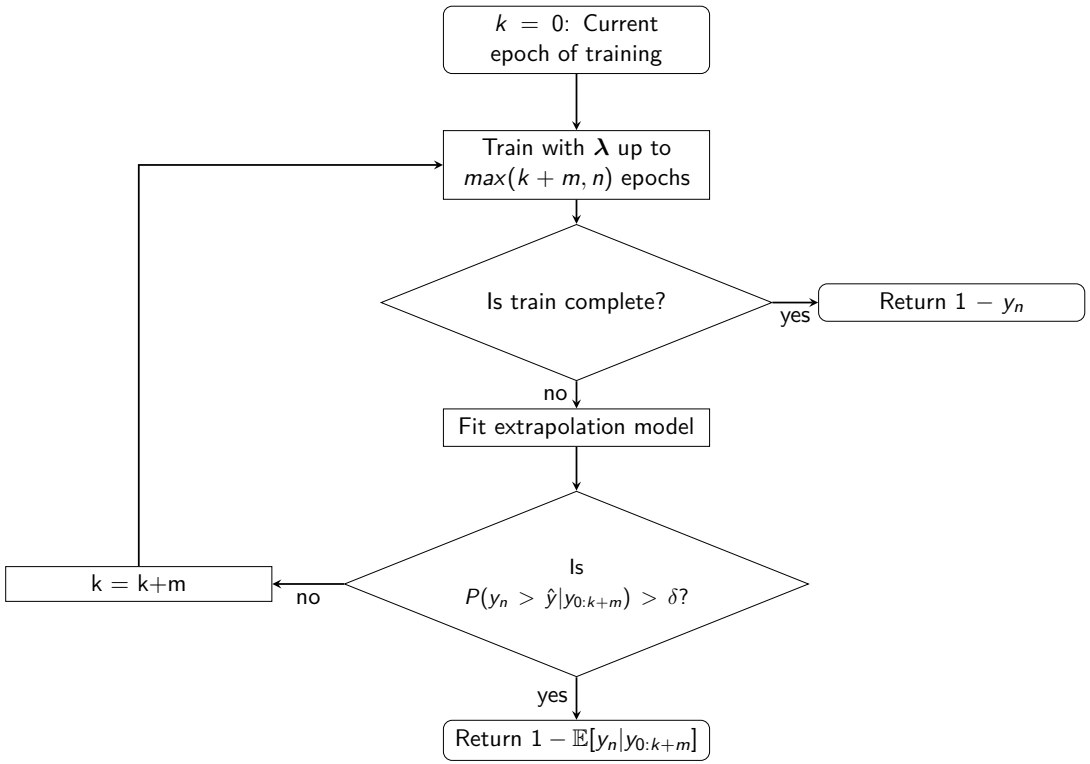


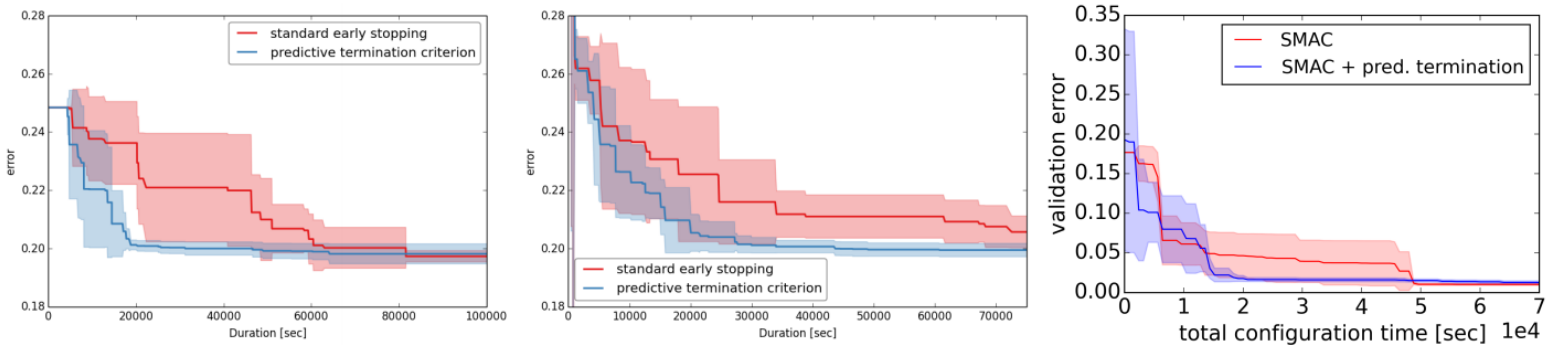
Figure: Predictive termination by extrapolation of learning curves

Domhan, Tobias, et al. "Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves." Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI). 2015.

- $\hat{y}$: Best accuracy seen so far
- λ: Selected hyper-parameter set
- n: Maximum number of epochs
- δ: Predictive termination threshold

Electrical & Computer ENGINEERING

Experimental Results



(a) SMAC on k-means CIFAR-10 (b) TPE on k-means CIFAR-10 (c) SMAC on MNIST

Figure: Results on CIFAR-10 and MNIST image classification tasks using DNNs

Domhan, Tobias, et al. "Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves." Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI). 2015.

Electrical & Computer ENGINEERING

Research Questions

- Current approach uses only expected value from extrapolation model
  - Incorporating extrapolation uncertainty?
- Current approach has been shown to work in only single objective optimization
  - How does the approach extend to multiple objectives?
  - Can the approach be applied to hierarchical constrained Bayesian optimization?

Electrical & Computer ENGINEERING

Incorporating Uncertainty into Bayesian Optimization

Proposed Solution

- Sample from extrapolation model
- Compute the standard deviation of the samples
- Incorporate into the covariance matrix of Bayesian Optimization

Benefits:

- Does not completely disregard visiting previously discontinued models.

Electrical & Computer ENGINEERING

Outline

Electrical & Computer ENGINEERING

Hyper-parameters for Speech Recognition

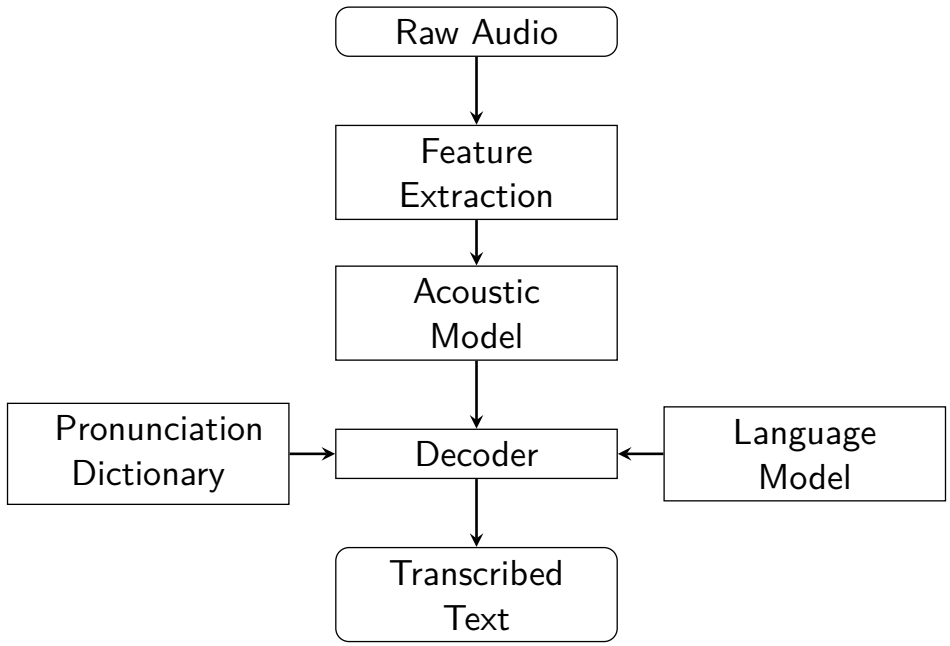


Figure: Typical Speech Recognition System Pipeline

Electrical & Computer ENGINEERING

Hyper-parameters in an ASR system

Feature hyper-parameters

- Log-Mel Filterbank Features

Davis, Steven, and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." IEEE transactions on acoustics, speech, and signal processing 28.4 (1980): 357-366.

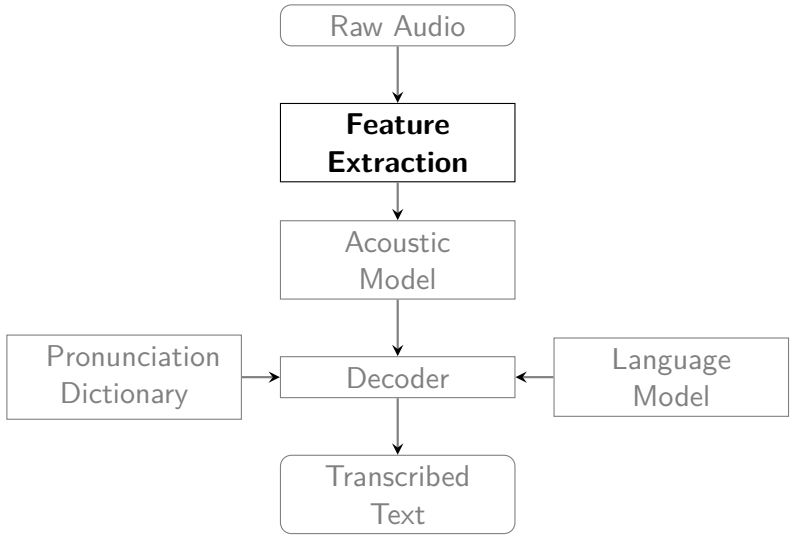


Figure: Typical Speech Recognition System Pipeline

Electrical & Computer ENGINEERING

Hyper-parameters in an ASR system

Feature hyper-parameters

- Log-Mel Filterbank Features
  - Window Size
  - Window Shift
  - Amount of Pre-emphasis
  - Number of Frequency banks
  - Number of Mel-scale Filterbanks

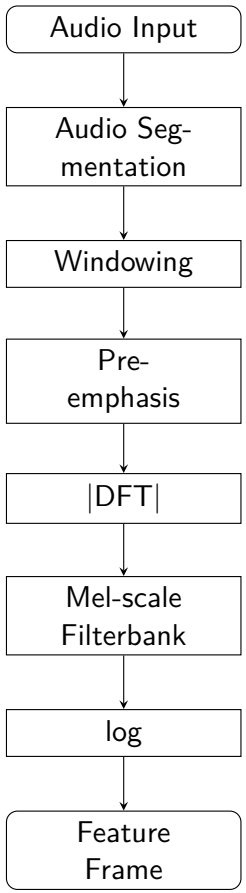


Figure: Log-Mel feature extraction pipeline

Electrical & Computer ENGINEERING

Hyper-parameters in an ASR system

Feature hyper-parameters

- Log-Mel Filterbank Features
  - **Window Size**
  - **Window Shift**
  - Amount of Pre-emphasis
  - Number of Frequency banks
  - **Number of Mel-scale Filterbanks**

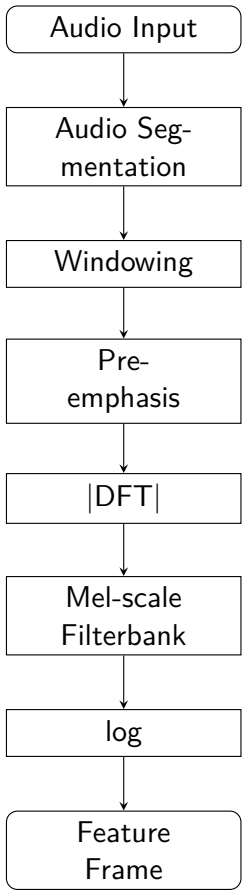


Figure: Log-Mel feature extraction pipeline

Electrical & Computer ENGINEERING

Hyper-parameters in an ASR system

Acoustic Model hyper-parameters

- Feed Forward Deep Neural Network

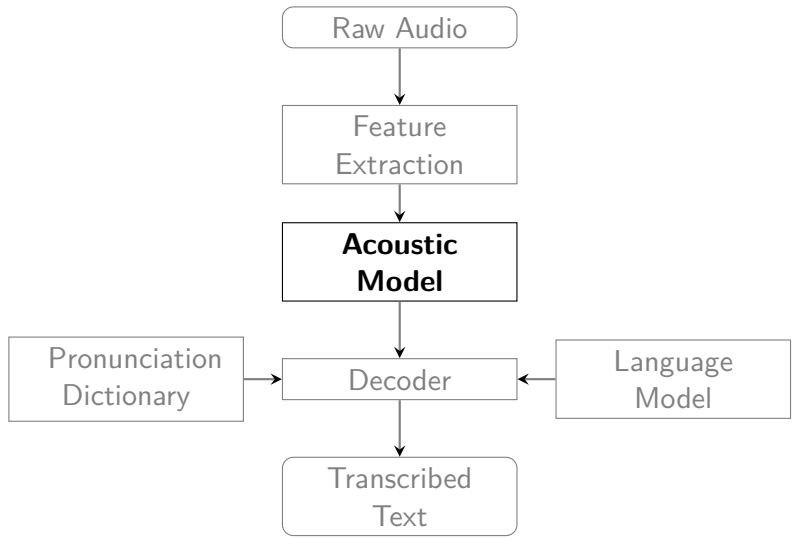


Figure: Typical Speech Recognition System Pipeline

Electrical & Computer ENGINEERING

Hyper-parameters in an ASR system

Acoustic Model hyper-parameters

- Feed Forward Deep Neural Network
  - Amount of input frame splicing
  - Number of hidden layers
  - Number of neurons per hidden layer
  - Type of activation for neurons

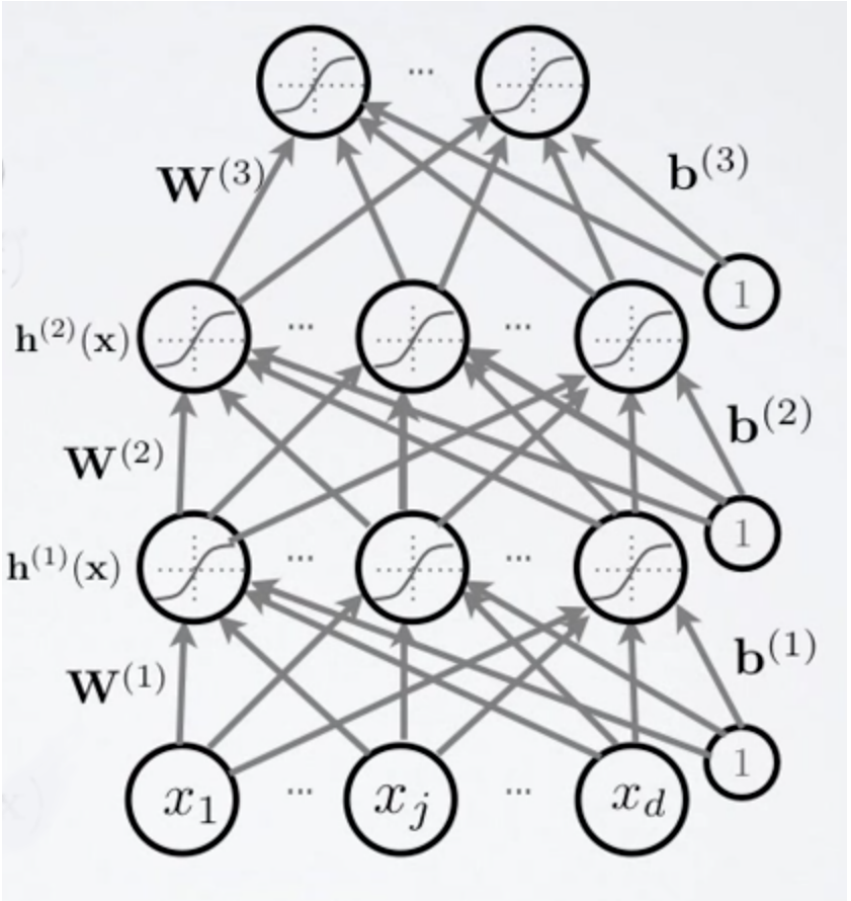


Figure: Feed-forward DNN architecture

Electrical & Computer ENGINEERING

Hyper-parameters in an ASR system

Acoustic Model hyper-parameters

- Feed Forward Deep Neural Network
  - **Amount of input frame splicing**
  - **Number of hidden layers**
  - **Number of neurons per hidden layer**
  - Type of activation for neurons

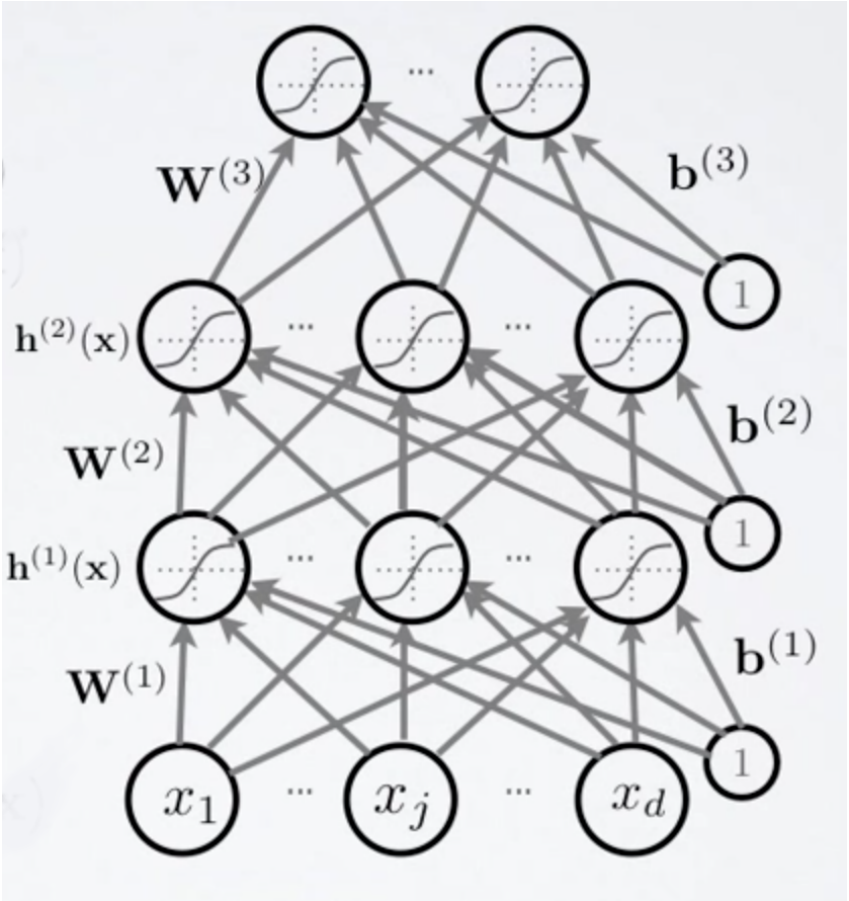


Figure: Feed-forward DNN architecture

Electrical & Computer ENGINEERING

Hyper-parameters in an ASR system

Language Model hyper-parameters

- N-gram Language model
  - N
  - Pruning Threshold
- RNN Language model
  - Number of hidden layers
  - Number of neurons per layer

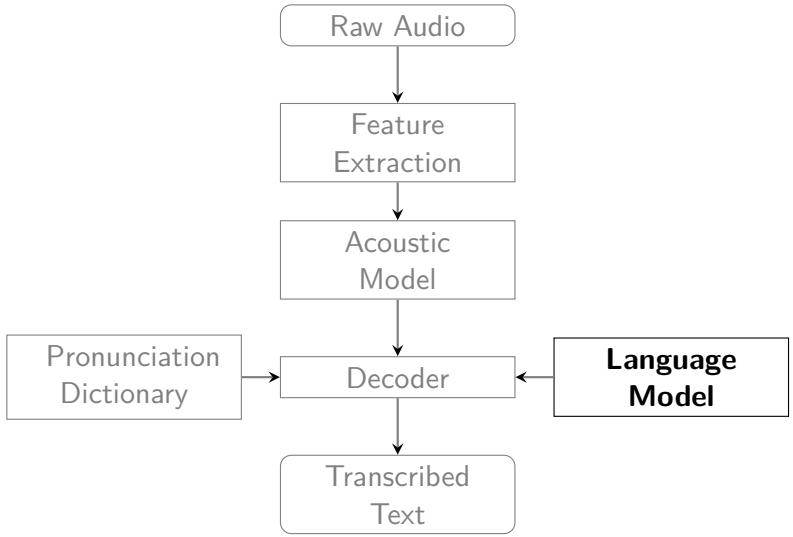


Figure: Typical Speech Recognition System Pipeline

Electrical & Computer ENGINEERING

Hyper-parameters in an ASR system

Decoder hyper-parameters

- Lattice-based Viterbi Decoder
  - Acoustic scale
  - Decoding beam
  - Minimum number of active states per frame
  - Maximum number of active states per frame
  - Lattice pruning beam
  - Lattice pruning interval

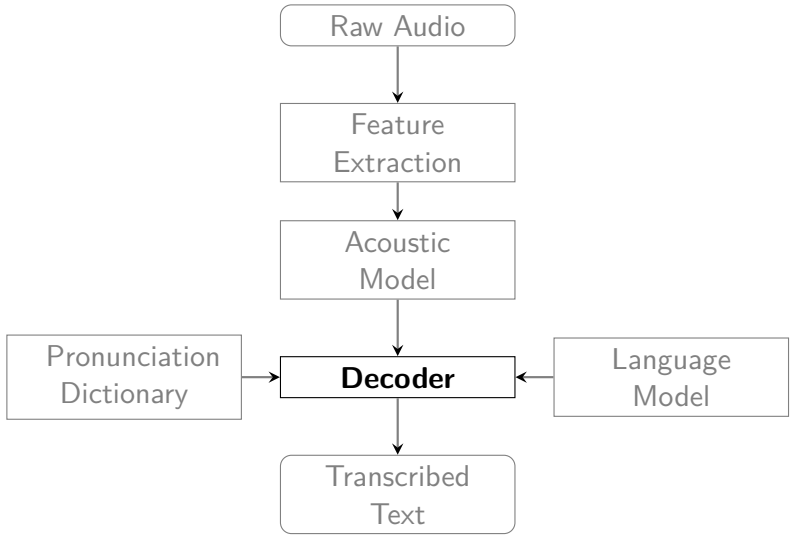


Figure: Typical Speech Recognition System Pipeline

Electrical & Computer ENGINEERING

Hyper-parameters in an ASR system

Decoder hyper-parameters

- Lattice-based Viterbi Decoder
  - **Acoustic scale**
  - **Decoding beam**
  - **Minimum number of active states per frame**
  - **Maximum number of active states per frame**
  - **Lattice pruning beam**
  - **Lattice pruning interval**

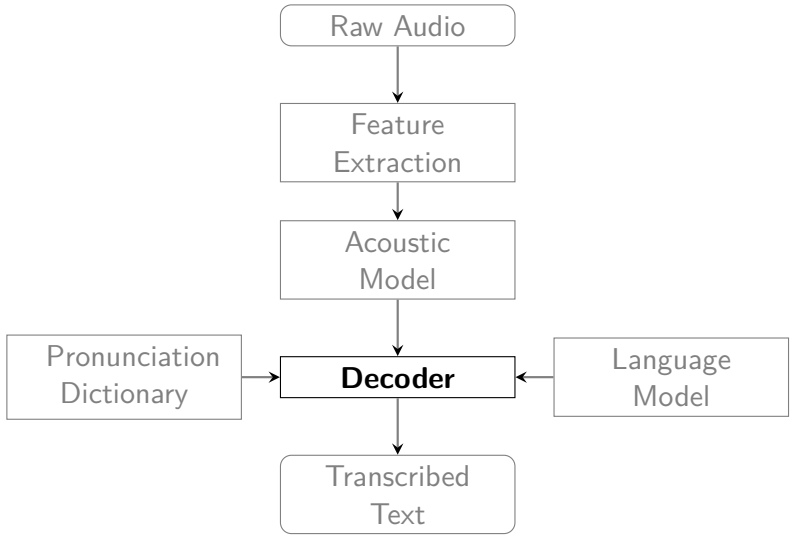


Figure: Typical Speech Recognition System Pipeline

Electrical & Computer ENGINEERING

Hyper Parameters

Stage	Name	Type	Values
Training	Frame Size	Integer	5:5:50
	Frame Shift	Integer	10:5:20
	Number of Mel bins	Integer	5:5:50
	Amount of Splicing	Integer	0:1:9
	Number of Hidden Layers	Integer	1:1:6
	Neurons per Hidden Layer	Integer	512:256:2816
Inference	Acoustic Scale	Real	0.05 - 0.15
	Decoder Beam	Real	10.0 - 18.0
	Max Active States	Integer	3000:500:8000
	Min Active States	Integer	50:50:300
	Lattice Pruning Beam	Real	4.0 - 10.0
	Lattice Pruning Interval	Integer	5:5:50

Electrical & Computer ENGINEERING

Outline

Electrical & Computer ENGINEERING

Optimization Metrics

- Method of evaluating a given algorithm.
- Types:
  - Performance Metrics: Word Error Rate (WER)
  - Computational Metrics: Real Time Factor (RTF)

Electrical & Computer ENGINEERING

Word Error Rate (WER)

- Ratio of number of substitutions, insertions and deletions required to match a hypothesis to reference to the number of words in the reference.

$$WER = \frac{S + D + I}{N}$$

- Lower Bound: 0

Electrical & Computer ENGINEERING

Real Time Factor (RTF)

- Ratio of time taken to decode audio to total duration of audio

$$RTF = \frac{\sum_{i=0}^{N} t_i(decode)}{\sum_{i=0}^{N} t_i(audio)}$$

- Lower Bound: 0^+

Electrical & Computer ENGINEERING

Experimental Setup

- Data:
  - Wall Street Journal Corpus
    - Train: 284 hrs
    - Dev: 1.08 hrs
    - Eval: 0.72 hrs
- Edge Hardware:
  - NVidia Jetson TX1
    - GPU: 256-core with NVIDIA Maxwell Architecture
    - CPU: 64-bit ARM A57 CPUs (4 cores)
    - RAM: 4GB (Unified Memory)

Training happens on server. Computational metric computation happens on edge hardware.

Electrical & Computer ENGINEERING

Optimization Setup

- Number of model builds: 25
- Objective function: WER
- Constraint: $RTF \leq 0.5$
- Hierarchical Constrained Bayesian Optimization:
  - Number of decoder evaluations per model: 20
- Hyper-parameter Optimization done on dev set.

Electrical & Computer ENGINEERING

Outline

Electrical & Computer ENGINEERING

Hierarchical Constrained Bayesian Optimization

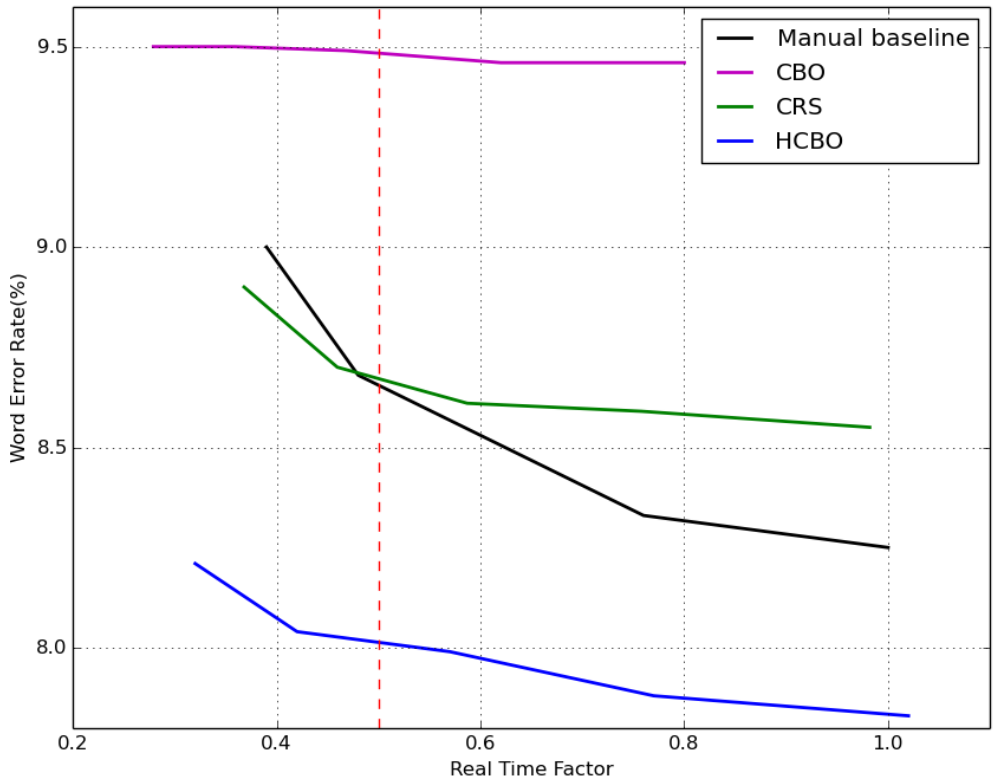


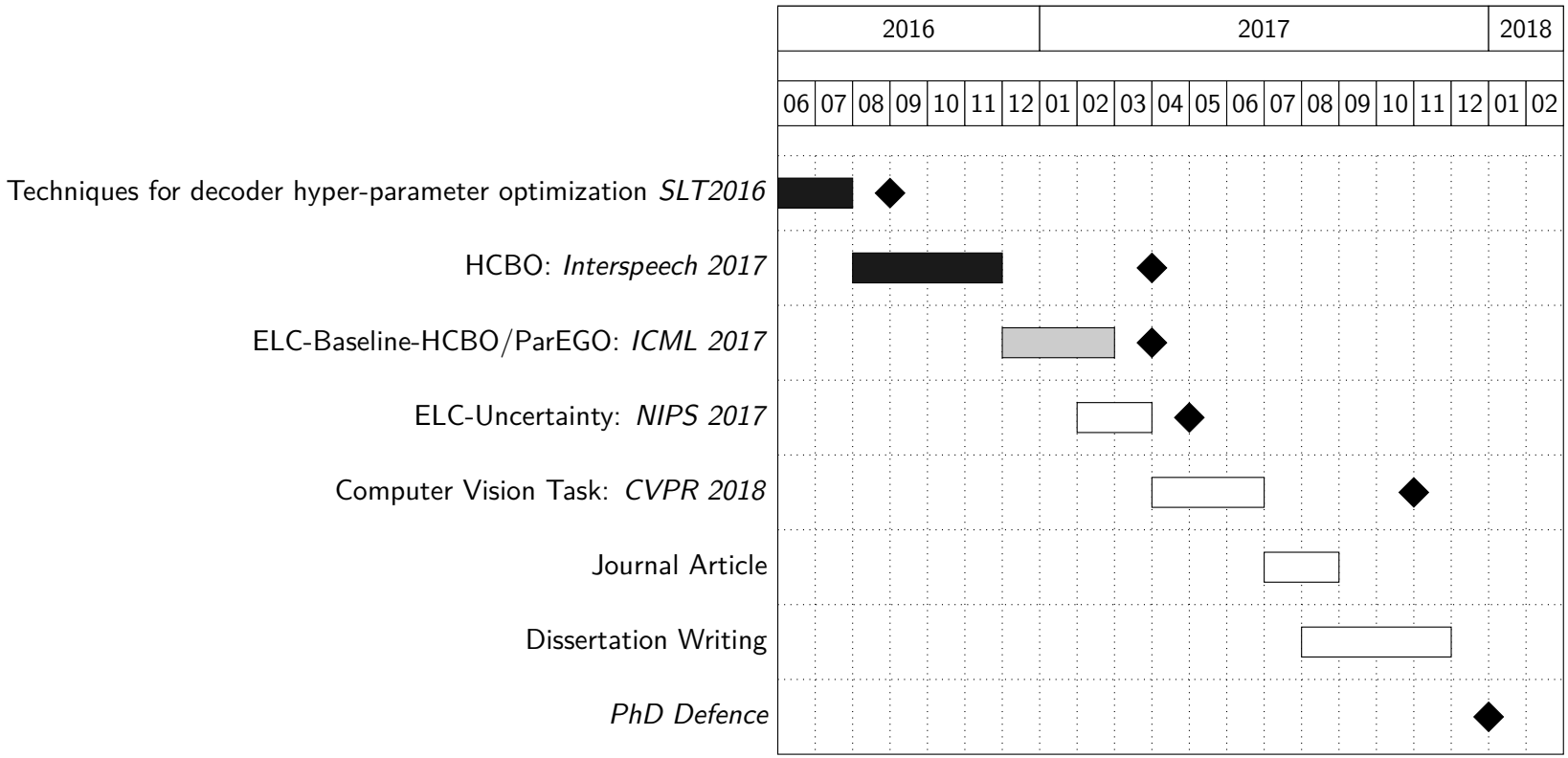
Figure: Comparison of best models built with various constraint based hyper-parameter optimization techniques

Electrical & Computer ENGINEERING

Outline

Electrical & Computer ENGINEERING

Timeline



Electrical & Computer ENGINEERING

Outline

Electrical & Computer ENGINEERING

Contributions

"We'll start to see edge computing come online in a big way within the next five years."
- Peter Levine, 2016

- Growing interest in model structure optimization for deep learning.
  - But little to no focus on optimizing towards edge computing.

Józefowicz, Rafal et al. "An Empirical Exploration of Recurrent Network Architectures." ICML (2015).
McGraw, Ian et al. "Personalized speech recognition on mobile devices" ICASSP 2016.
Data Center on Wheels:
http://www.businessinsider.com/edge-computing-is-the-next-multi-billion-tech-market-2016-12

Electrical & Computer ENGINEERING

Contributions

"We'll start to see edge computing come online in a big way within the next five years."
- Peter Levine, 2016

- Growing interest in model structure optimization for deep learning.
  - But little to no focus on optimizing towards edge computing.

Józefowicz, Rafal et al. "An Empirical Exploration of Recurrent Network Architectures." ICML (2015).
McGraw, Ian et al. "Personalized speech recognition on mobile devices" ICASSP 2016.
Data Center on Wheels:
http://www.businessinsider.com/edge-computing-is-the-next-multi-billion-tech-market-2016-12

Electrical & Computer ENGINEERING

Contributions

- Current Automated Techniques outperform manual optimization
  - But, slower convergence as number of hyper-parameters increase.
  - Do not exploit structure within hyper-parameters themselves.

Deb, Kalyanmoy, et al. "A fast and elitist multiobjective genetic algorithm: NSGA-II." IEEE transactions on evolutionary computation 6.2 (2002): 182-197.
Snoek, Jasper, et al. "Practical bayesian optimization of machine learning algorithms." Advances in neural information processing systems. 2012.
Eggensperger, Katharina, et al. "Towards an empirical foundation for assessing bayesian optimization of hyperparameters." NIPS workshop on Bayesian Optimization in Theory and Practice. 2013.

Electrical & Computer ENGINEERING

Contributions

- Current Automated Techniques outperform manual optimization
  - But, slower convergence as number of hyper-parameters increase.
  - Do not exploit structure within hyper-parameters themselves.

Deb, Kalyanmoy, et al. "A fast and elitist multiobjective genetic algorithm: NSGA-II." IEEE transactions on evolutionary computation 6.2 (2002): 182-197.
Snoek, Jasper, et al. "Practical bayesian optimization of machine learning algorithms." Advances in neural information processing systems. 2012.
Eggensperger, Katharina, et al. "Towards an empirical foundation for assessing bayesian optimization of hyperparameters." NIPS workshop on Bayesian Optimization in Theory and Practice. 2013.

Electrical & Computer ENGINEERING

Contributions

- Current Automated Techniques outperform manual optimization
  - But, slower convergence as number of hyper-parameters increase.
  - Do not exploit structure within hyper-parameters themselves.

Deb, Kalyanmoy, et al. "A fast and elitist multiobjective genetic algorithm: NSGA-II." IEEE transactions on evolutionary computation 6.2 (2002): 182-197.
Snoek, Jasper, et al. "Practical bayesian optimization of machine learning algorithms." Advances in neural information processing systems. 2012.
Eggensperger, Katharina, et al. "Towards an empirical foundation for assessing bayesian optimization of hyperparameters." NIPS workshop on Bayesian Optimization in Theory and Practice. 2013.

Electrical & Computer ENGINEERING

Conclusion

- Computation metric from Edge Device
  - Allows for hardware-aware optimization.
- Hierarchical Constrained Bayesian Optimization.
  - Exploits hierarchy in hyper-parameters to increase search efficiency.
- Incorporating uncertainty in Extrapolation of Learning Curves.
  - Lesser wall-time required for explorative search (ParEGO).
  - Ensures models not discarded due to erroneous extrapolation.

Electrical & Computer ENGINEERING

Conclusion

- Computation metric from Edge Device
  - Allows for hardware-aware optimization.
- Hierarchical Constrained Bayesian Optimization.
  - Exploits hierarchy in hyper-parameters to increase search efficiency.
- Incorporating uncertainty in Extrapolation of Learning Curves.
  - Lesser wall-time required for explorative search (ParEGO).
  - Ensures models not discarded due to erroneous extrapolation.

Electrical & Computer ENGINEERING

Conclusion

- Computation metric from Edge Device
  - Allows for hardware-aware optimization.
- Hierarchical Constrained Bayesian Optimization.
  - Exploits hierarchy in hyper-parameters to increase search efficiency.
- Incorporating uncertainty in Extrapolation of Learning Curves.
  - Lesser wall-time required for explorative search (ParEGO).
  - Ensures models not discarded due to erroneous extrapolation.

Electrical & Computer ENGINEERING

Conclusion

- Computation metric from Edge Device
  - Allows for hardware-aware optimization.
- Hierarchical Constrained Bayesian Optimization.
  - Exploits hierarchy in hyper-parameters to increase search efficiency.
- Incorporating uncertainty in Extrapolation of Learning Curves.
  - Lesser wall-time required for explorative search (ParEGO).
  - Ensures models not discarded due to erroneous extrapolation.

A structured, efficient search approach to produce DL models with performance comparable to Cloud on an Edge Device

Electrical & Computer ENGINEERING

Questions?

Electrical & Computer ENGINEERING

Deep Speech 2 AM flop computation

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z)$$
$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r)$$
$$h_t = (1 - z_t) \circ \sigma_h(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h) + z_t \circ h_{t-1}$$

Where

x_t :Input vector of size m

h_t :Output vector of size n

z_t :Update gate vector

r_t :Reset gate vector

W, U, b :parameter matrices and bias vectors

σ_g :sigmoid

σ_h :tanh

Hence,

GRU layer flops / frame $= 6mn + 6nn$

Electrical & Computer ENGINEERING

Seq2Seq flop computation

For an LSTM Layer

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f)$$
$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o)$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c)$$
$$h_t = o_t \circ \sigma_h(c_t)$$

Where

x_t :input vector of size m

o_t :Output vector of size n

c_t :Cell state vector

W, U, b :Parameter matrices and bias vectors

f_t, i_t, o_t :Gate vectors

σ_g :sigmoid

σ_c :tanh

σ_h :tanh

Hence

LSTM layerflops/word $= 8mn + 8nn$

Electrical & Computer ENGINEERING

Seq2Seq flop computation

For output word softmax layer:

Vocabulary size :k

embedding size :n

Hence

Output softmax layer flops/word :$2nk$

Electrical & Computer ENGINEERING

Single Objective Optimization

- Random Sampling
- Sequential Model-based Global Optimization
  - Gaussian Process based Bayesian Optimization
  - Tree of Parzen Estimators
- Evolutionary Algorithms

Electrical & Computer ENGINEERING

Single Objective Optimization

- Random Sampling
- Sequential Model-based Global Optimization
  - **Gaussian Process based Bayesian Optimization**
  - Tree of Parzen Estimators
- Evolutionary Algorithms

Electrical & Computer ENGINEERING

Bayesian Optimization

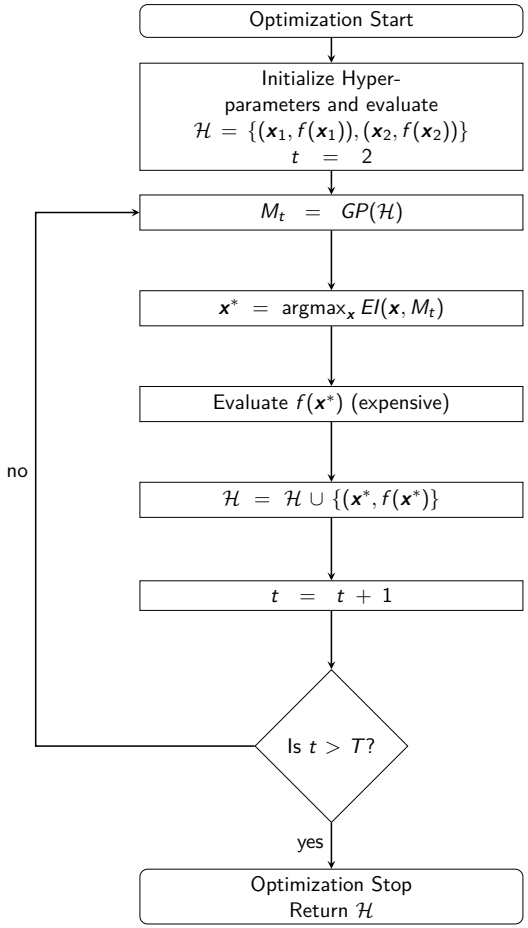


Figure: Procedure for Bayesian Optimization

Where

- $\mathcal{H}$: set of all evaluated points
- $GP(.)$: Gaussian Process with given prior mean and covariance function
- EI: Expected improvement at a given point using given model

Shahriari, Bobak, et al. "Taking the human out of the loop: A review of bayesian optimization." Proceedings of the IEEE 104.1 (2016): 148-175.

Electrical & Computer ENGINEERING

Bayesian Optimization

$$f(\boldsymbol{x}) \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$$

$$m(\boldsymbol{x}) : \text{Prior mean function of GP}$$

$$k(\boldsymbol{x}, \boldsymbol{x}') : \text{Covariance function for GP}$$

$$f(\boldsymbol{x})|\boldsymbol{x}, \mathcal{H} \sim \mathcal{N}\left(\mu_t(\boldsymbol{x}), \sigma_t^2(\mathbf{x})\right)$$

$$\mu_t(\mathbf{x}) = \mathbf{k}^T \mathbf{K}^{-1} \mathbf{f}_{1:t}$$

$$\sigma_t^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}^T \mathbf{K}^{-1} \mathbf{k}$$

$$\mathbf{k} = \begin{bmatrix} k(\mathbf{x}, \mathbf{x}_1) \\ k(\mathbf{x}, \mathbf{x}_2) \\ \vdots \\ k(\mathbf{x}, \mathbf{x}_t) \end{bmatrix} \mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \dots & k(\mathbf{x}_1, \mathbf{x}_t) \\ k(\mathbf{x}_2, \mathbf{x}_1) & \dots & k(\mathbf{x}_2, \mathbf{x}_t) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_t, \mathbf{x}_1) & \dots & k(\mathbf{x}_t, \mathbf{x}_t) \end{bmatrix}$$

Electrical & Computer ENGINEERING

Expected Improvement

$$EI(\mathbf{x}) = \mathbb{E}[f(\mathbf{x}) - f(\mathbf{x}^+)]$$

$$EI(\mathbf{x}) = \begin{cases} (\mu_t(\mathbf{x}) - f(\mathbf{x}^+))\Phi(Z) \\ +\sigma_t(\mathbf{x})\phi(Z) & \text{if } \sigma_t(x) > 0 \\ 0 & \text{else} \end{cases}$$

$$Z = \frac{\mu_t(\mathbf{x}) - f(\mathbf{x}^+)}{\sigma_t(\mathbf{x})}$$

where

$\mathbf{x}^+$: best hyper-parameter so far

ϕ : Normal PDF

Φ : Normal CDF

Electrical & Computer ENGINEERING

Uncertainty in Bayesian Optimization

$$y = f(\boldsymbol{x}) + \epsilon$$

where

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

then

$$\mu_t(\mathbf{x}) = \mathbf{k}^T(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}_{1:t}$$

$$\sigma_t^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}^T(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{k}$$

Electrical & Computer ENGINEERING

Augmented Tchebyscheff Function

- Combines multiple normalized objective function values using a scalarizing vector.

$$ATF(\boldsymbol{x}) = \max_j \left(w_j \hat{f}_j(\boldsymbol{x})\right) + \rho \sum_{k=1}^{M} w_k \hat{f}_k(\boldsymbol{x})$$

where

$$w_k \geq 0 \ \forall k$$

$$|\boldsymbol{w}|_2 = 1$$

Dächert, Kerstin, Jochen Gorski, and Kathrin Klamroth. "An augmented weighted Tchebycheff method with adaptively chosen parameters for discrete bicriteria optimization problems." Computers and Operations Research 39.12 (2012): 2929-2943.

Electrical & Computer ENGINEERING

Experimental Results

Experiment	WER (%)		RTF
	dev93	eval92	
Manual	8.73	5.09	0.43
CRS			
CBO*	9.65	5.43	0.46
HCBO	8.06	4.32	0.49

Table: Performance of the HCBO on the TX1 platform

Electrical & Computer ENGINEERING

Experimental Results

TODO: Add graph showing tradeoff curves using each technique.

Electrical & Computer ENGINEERING

What further?

- Can decoder hyper-parameter settings be transferred across different instances of training hyper-parameters?

Electrical & Computer ENGINEERING

KWS

Table: Comparison of the performance of the optimization techniques against the baseline. The number in the bracket is the relative percentage gain over the baseline.

Optimization Method	Optimization Metric	KwAcc (%improvement)	AUC (%improvement)
Baseline	-NA-	82%	0.161
Random Sampling	KwAcc	**89% (8.5%)**	0.188 (-16.7%)
	AUC	88% (7.3%)	0.137 (14.9%)
Bayesian Optimization	KwAcc	87% (6.1%)	0.139 (13.6%)
	AUC	86% (4.9%)	**0.125 (22.3%)**

Electrical & Computer ENGINEERING

CHiME1 VAD

Optimization Type	Fscore	RTF
Manual Optimization	0.9443 (0.9448)	0.0074
Random Sampling[1]	0.9384 (0.9390)	0.0268
Bayesian Optimization	0.9551 (0.9555)	0.0931

Table: F-Scores and RTFs of best model for audio only VAD. All numbers within brackets denote the results for the test data

Electrical & Computer ENGINEERING

CHiME1 VAD

RTF Thresholds	Constrained Random Sampling[1]	Constrained Bayesian Optimization
0.0001	0.8024 (0.7985)	0.8030 (0.8031)
0.0002	0.8167 (0.8150)	0.8417 (0.8429)
0.0005	0.8732 (0.8705)	0.8960 (0.8964)
0.001	0.9009 (0.9005)	0.9237 (0.9329)
0.002	0.9160 (0.9162)	0.9473 (0.9478)
0.005	0.9278 (0.9298)	0.9509 (0.9514)
0.01	0.9335 (0.9334)	0.9519 (0.9523)

Table: F-Scores for best models by constrained optimization at different RTF thresholds for audio only VAD

Electrical & Computer ENGINEERING

CHiME1 ASR

Optimization Type	WER(%)	RTF
Manual Optimization	23.32 (24.87)	0.5729
Random Sampling[1]	22.94 (24.05)	0.4943
Bayesian Optimization	22.61 (24.01)	0.4044

Table: WERs and RTFs of best model for audio only ASR

ASR Type	Constrained Random Sampling	Constrained Bayesian Optimization
Audio only	24.07 (25.19) [1]	22.56 (23.92)

Table: Comparison of model WERs for ASR with RTF threshold 0.25

Electrical & Computer ENGINEERING

WSJ

Optimization	WER (%)		RTF
	dev93	eval92	
Manual	8.09	4.75	0.19
CBO	9.44	5.53	0.07
HCBO	8.04	4.94	0.16

Table: Performance of optimization techniques on server platform

	Manual			HCBO		
	WER (%)		RTF	WER (%)		RTF
	dev93	eval92		dev93	eval92	
TK1	9.52	5.72	0.45	8.26	4.98	0.44
TX1	8.73	5.09	0.43	8.06	4.32	0.49

Table: Performance of the optimization techniques on the Embedded platforms

Electrical & Computer ENGINEERING