JOHANNES KEPLER
UNIVERSITY LINZ

Optimal Gradient-Based Learning Using Importance Weights (Revisited)

Martin Heusel

30.05.2016

Institute of Bioinformatics
Johannes Kepler University, Linz, Austria

Introduction

JOHANNES KEPLER
UNIVERSITY LINZ

- Already published 2005

K. Obermayer S. Hochreiter. Optimal gradient-based learning using importance weights. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 114–119, 2005.

- For difficult datasets, i.e. highly non-linear, long-term dependencies, unbalanced
- Significant speed up
- Equal or better prediction results
- Record on the “latching benchmark” sequence prediction, extract and exploit dependencies between sites which are 1,000,000 sequence elements apart

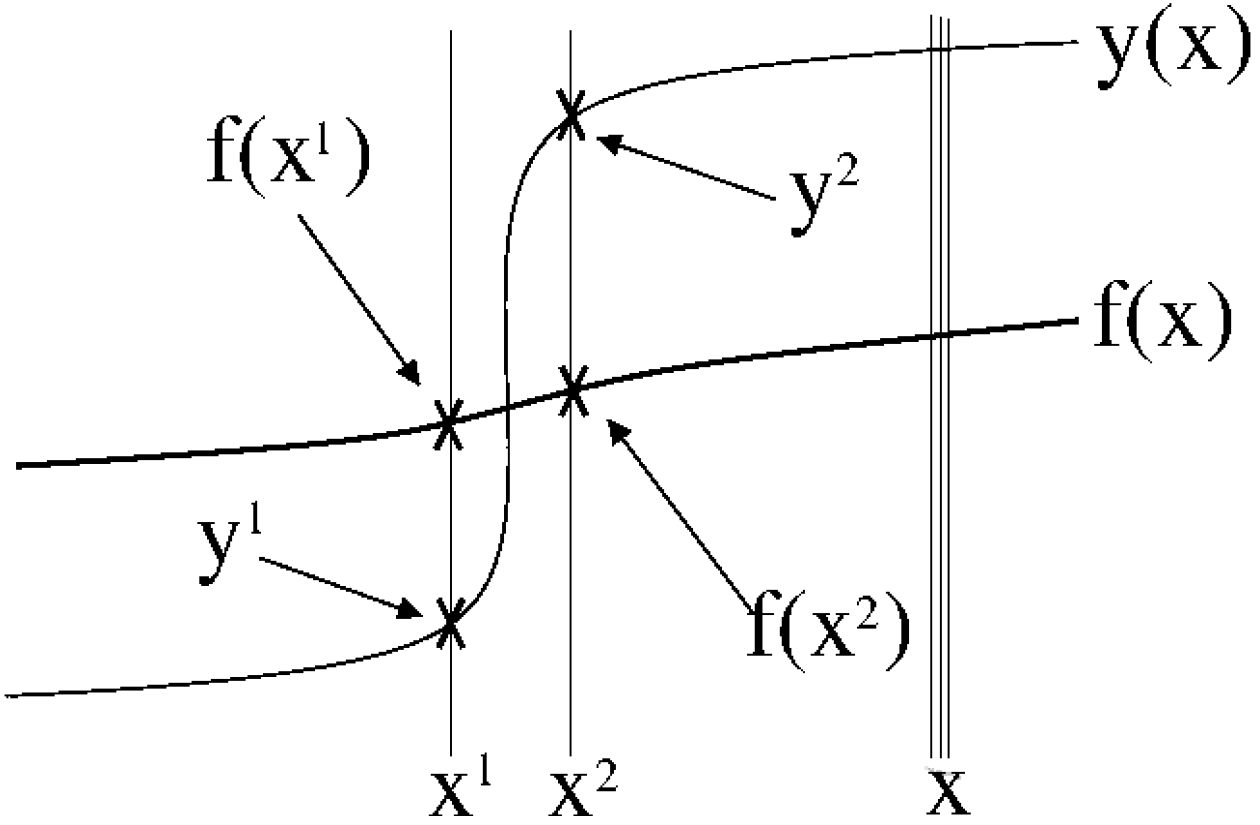
Introduction

JOHANNES KEPLER
UNIVERSITY LINZ

- How Importance Weights behave on today's large datasets, deep nets, minibatch learning, batch normalizing, dropout, multitask, multiclass, etc.?

Motivation

JOHANNES KEPLER
UNIVERSITY LINZ



Gradient contributions at x1 and x2 cancel each other out. Gradient contributions can dominate for similar x, large classes.

Motivation

JOHANNES KEPLER
UNIVERSITY LINZ

Optimal learning step for the parameter vector **w** following two goals:

(a) The individual error for every data point decreases by at least a value of p (if possible)

(b) the associated weight change $\Delta\mathbf{w}$ should be as small a possible.

Formulation of the Optimization Problem

JOHANNES KEPLER UNIVERSITY LINZ

Taylor expansion:

$$E(\boldsymbol{x}^i; \boldsymbol{w} + \Delta\boldsymbol{w}) \;=\; E(\boldsymbol{x}^i; \boldsymbol{w}) \;+\; \left\langle \nabla_{\boldsymbol{w}} E(\boldsymbol{x}^i; \boldsymbol{w}), \Delta\boldsymbol{w} \right\rangle \;+\; O\left(\|\Delta\boldsymbol{w}\|^2\right) \;,$$

so that decreasing the error by p leads to $\left\langle -\nabla_{\boldsymbol{w}} E(\boldsymbol{x}^i; \boldsymbol{w}), \Delta\boldsymbol{w} \right\rangle \;\geq\; p.$

Optimization problem:

$$\begin{aligned} min_{\Delta\boldsymbol{w}} \quad & \frac{1}{2}\|\Delta\boldsymbol{w}\|^2 \\ s.t. \quad & \left\langle -\nabla_{\boldsymbol{w}} E(\boldsymbol{x}^i; \boldsymbol{w}) \;,\; \Delta\boldsymbol{w} \right\rangle \;\geq\; p \end{aligned}$$

Convex Optimization Problem with Slack Variables

JOHANNES KEPLER UNIVERSITY LINZ

- No guarantee for error improving by p for each sample i
- Slack variables ξ_i and regularization parameter C

Convex Optimization Problem with Slack Variables

JOHANNES KEPLER UNIVERSITY LINZ

- Convex optimization problem:

$$\min_{\Delta \boldsymbol{w}} \quad \frac{1}{2}\|\Delta \boldsymbol{w}\|^2 \; + \; C \sum_i \xi_i$$

$$s.t. \quad \left\langle -\nabla_{\boldsymbol{w}} E(\boldsymbol{x}^i) \; , \; \Delta \boldsymbol{w} \right\rangle \; \geq \; p \; - \; \xi_i, \quad 0 \leq \xi_i \; .$$

- Dual formulation

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \left\langle \nabla_{\boldsymbol{w}} E(\boldsymbol{x}^i), \nabla_{\boldsymbol{w}} E(\boldsymbol{x}^j) \right\rangle \; - \; p \; \sum_i \alpha_i$$

$$s.t. \quad 0 \leq \alpha_i \leq C \; .$$

Importance Weights

JOHANNES KEPLER UNIVERSITY LINZ

- New learning step using Karush-Kuhn-Tucker conditions:

$$\begin{aligned} \Delta \boldsymbol{w} &= -\sum_i \alpha_i \, \nabla_{\boldsymbol{w}} E(\boldsymbol{x}^i) \quad \text{and} \\ \|\Delta \boldsymbol{w}\|^2 &= p \sum_i \alpha_i + C \sum_{i:\alpha_i = C} \xi_i \; . \end{aligned}$$

- α_i are importance weights for training vectors and individual step sizes for standard gradient update rule
- Importance weights by minimizing contributions of coupling strengths $\left\langle \nabla_{\boldsymbol{w}} E(\boldsymbol{x}^i), \nabla_{\boldsymbol{w}} E(\boldsymbol{x}^j) \right\rangle$

Fast SMO Version

JOHANNES KEPLER
UNIVERSITY LINZ

- Using old solutions for the α_i leads to faster new solutions if only a few $\alpha_i \neq 0$
- Finding α_i , α_j which yield largest update
- Choice of α and search is efficient because no equality constraint

Sparse Solution / New LOO Bound

JOHANNES KEPLER UNIVERSITY LINZ

- $\alpha_i = 0$ for pairs of data points with similar gradient information which is common in today's large datasets
- α_i large for antiparallel gradients
- Controlled by parameters p and C
- Data points never learned, LOO almost unbiased $\rightarrow$
- New bound on generalization error using LOO estimators analog to bounds for SVMs

Experiments

JOHANNES KEPLER
UNIVERSITY LINZ

- Hemoglobin protein sequences, three species
- Protein remote homology fold prediction on SCOP data, 318 fold classes

Hemoglobin Protein Sequences

JOHANNES KEPLER UNIVERSITY LINZ

- 736 Hemoglobin A protein sequences from the organisms Mammials, Fish and Sauria from Uniprot
- Class size ratios 384:93:259 respectively
- Randomly allocate sequences into training and test, ratio 4:1
- (j)LSTM with SGD, Adadelta and Importance Weights
- 20 Memory Cells, sliding window 5
- Learning rate fixed to 0.01, online learning
- Parameters p and C for Importance Weights fixed to 0.1 each
- Softmax, Cross-Entropy Loss

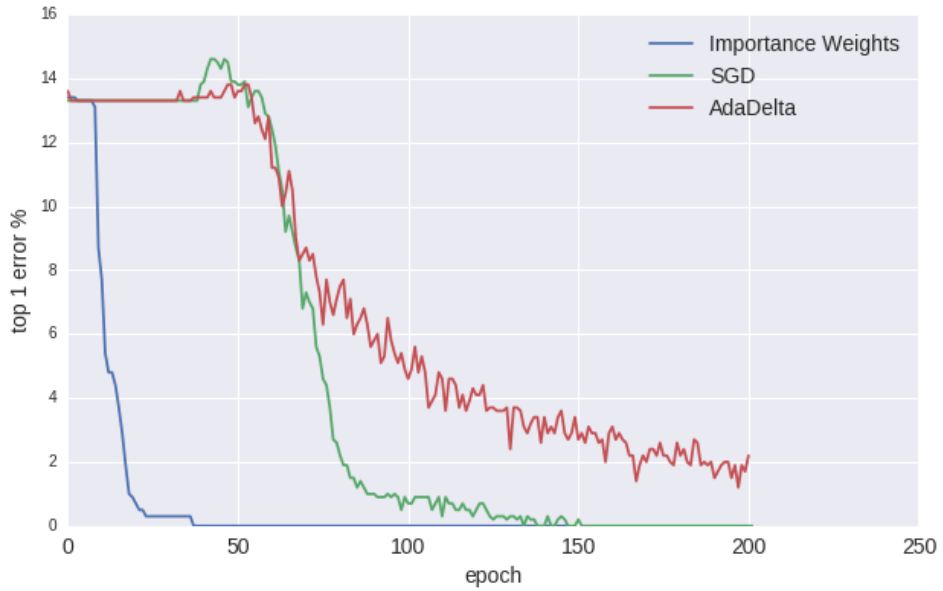
Hemoglobin Protein Sequences

JOHANNES KEPLER UNIVERSITY LINZ

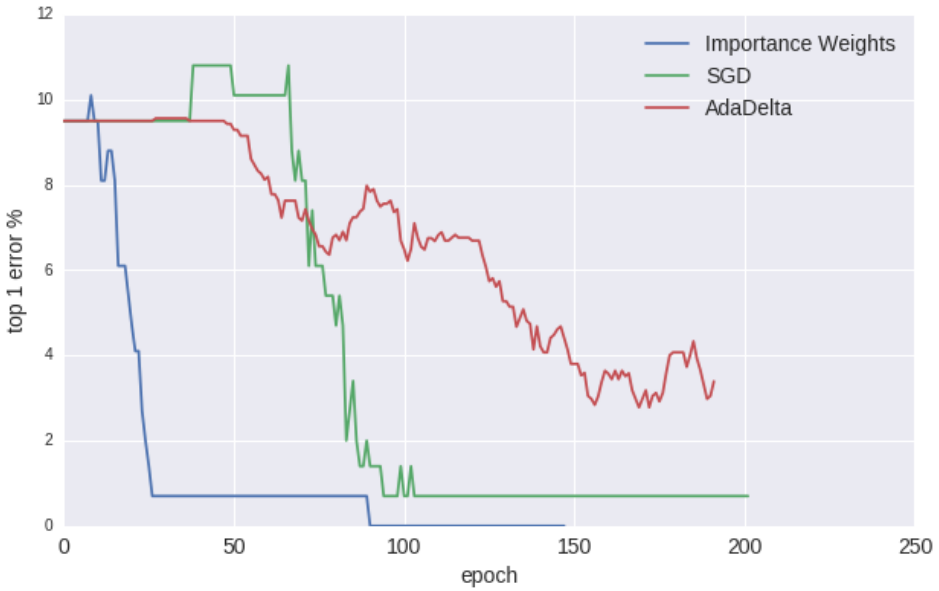
```
CLUSTAL O(1.2.1) multiple sequence alignment


fish        -SLSAKDKANVKAIWGKILPKSDEIGEQALSRMLVVYPQTKAYFSHWASVAPGSAPVKKH
bird        MVLSAADKTNVKGVFSKIGGHADDYGAETLERMFIAYPQTKTYFPHFD-LQHGSAQIKAH
human       MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFD-LSHGSAQVKGH
              **  **:***. :.*:  :: : * ::*.**:: :* **:** *:  :  *** :* *

fish        GITIMNQIDDCVGHMDDLFGFLTKLSELHATKLRVDPTNFKILAHNLIVVIAAYFPAEFT
bird        GKKVAAALVEAVNHIDDIAGALSKLSDLHAQKLRVDPVNFKFLGHCFLVVVAIHHPXALT
human       GKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFT
            * .:   : :.* *:**:   *: **:*** ******.***:*.* ::*.:* : *  :*

fish        PEIHLSVDKFLQQLALALAEKYR
bird        PEVHASLDKFMCAVGAVLTAKYR
human       PAVHASLDKFLASVSTVLTSKYR
            * :* *:***:  :. .*: ***
```

Hemoglobin 3 Class Prediction
Top 1 Error

JOHANNES KEPLER
UNIVERSITY LINZ



Training



Test

SCOP Protein Remote Homology Fold Prediction

JOHANNES KEPLER UNIVERSITY LINZ

- SCOP (Structural Classification of Proteins)
- Database of experimentally measured Protein folds
- Coordinate data mostly derived from Protein Data Bank
- Tree hierarchy of class, fold, superfamily, family, sequence
- SCOP 2.06: 77439 PDB entries, 7 classes, 1221 folds, 2008 superfamilies, 4851 families

SCOP FASTA Sequences

JOHANNES KEPLER UNIVERSITY LINZ

>d1rfya_ a.2.13.1 (A:) Transcriptional repressor TraM {Agrobacterium tumefaciens [TaxId: 358]}
KKVELRPLIGLTRGLPPTDLETITIDAIRTHRRLVEKADELFQALPETYKTGQACGGPQHIRYIEASIEMHAQMSALNTLISILGFIPK
>d1us6b_ a.2.13.1 (B:) Transcriptional repressor TraM {Agrobacterium tumefaciens [TaxId: 358]}
MELEDANVTKKVELRPLIGLTRGLPPTDLETITIDAIRTHRRLVEKADELFQALPETYKTGQACGGPQHIRYIEASIEMHAQMSALNTLYSILGFIPKVV
>d2hjda_ a.2.13.1 (A:) automated matches {Agrobacterium tumefaciens [TaxId: 358]}
FELRPVIGLTRGLSSADIETLTANAIRLHRQLLEKADQLFQVLPDDIKIGTAAGGEQHLEYIEAMIEMHAQMSAVNTLVGLLGFIPKVS
>d2qalr1 a.4.8.1 (R:19-73) Ribosomal protein S18 {Escherichia coli [TaxId: 562]}
EIDYKDIATLKNYITESGKIVPSRITGTRAKYQRQLARAIKRARYLSLLPYTDRH
>d2fnaa1 a.4.5.11 (A:284-356) Hypothetical protein SSO1545, C-terminal domain {Sulfolobus solfataricus [TaxId: 2287]}
REIARKRYLNIMRTLSKCGKWSDVKRALELEEGIEISDSEIYNYLTQLTKHSWIIKEGEKYCPSEPLISLAFS
>d2foka1 a.4.5.12 (A:5-143) Restriction endonuclease FokI, N-terminal (recognition) domain {Flavobacterium okeanokoites [TaxId: 244]}
IRTFGWVQNPGKFENLKRVVQVFDRNSKVHNEVKNIKIPTLVKESKIQKELVAIMNQHDLIYTYKELVGTGTSIRSEAPCDAIIQATIADQGNKKGYIDNW...
>d4kphl1 b.1.1.0 (L:1-107) automated matches {Mouse (Mus musculus) [TaxId: 10090]}
EIVLTQSPAIMSASLGEEITLTCSASSSVNYMHWYQQKSGTSPKLLIYTTSNLASGVPSRFSGSGSGTFYSLTISSVEAEDAADYYCHQWSSYPTFGGGTKLEIK
>d4p46a1 b.1.1.0 (A:2-111) automated matches {Mouse (Mus musculus) [TaxId: 10090]}
QVRQSPQSLTVWEGETAILNCSYENSAFDAFPWYQQFPGEGPALLIAIRSVSDKKEDGRFTIFFNKREKKLSLHITDSQPGDSATYFCAASKGADRLTFGKGTQLIIQP
>d5buva1 b.82.1.0 (A:2-173) automated matches {Yersinia enterocolitica [TaxId: 393305]}
IINITELNISGCYLIESPIFSDERGEFVKTHHQEIFKNFGLEIPSAEEYYSRSKNNVIRGMHFQQYPDDHNKLVFCPEGEVLDVFLDIRKDSNTYGQFMSFILNPHNRRSI...
>d1ep0a_ b.82.1.1 (A:) dTDP-4-dehydrorhamnose 3,5-epimerase RmlC {Methanobacterium thermoautotrophicum [TaxId: 145262]}
EFRFIKTSLDGAIIIEPEVYTDERGYFMETFNEAIFQENGLEVRFVQDNESMSVRGVLRGLHFQREKPQGKLVRVIRGEIFDVAVDLRKNSDTYGEWTGVRLSDENRREFFI...
>d2xwsa1 c.92.1.3 (A:1-125) automated matches {Archaeoglobus fulgidus [TaxId: 2234]}
MRRGLVIVGHGSQLNHYREVMELHRKRIEESGAFDEVKIAFAARKRRPMPDEAIREMNCDIIYVVPLFISYGLHVTEDLPDLLGFPRGRGIKEGEFEGKKVVICEPIGE...
>d3pdia_ c.92.2.0 (A:) automated matches {Azotobacter vinelandii [TaxId: 322710]}
GCAKPKPGATDGGCSFDGAQIALLPVADVAHIVHGPIACAGSSWDNRGTRSSGPDLYRIGMTTDLTENDVIMGRAEKRLFHAIRQAVESYSPPAVFVYNTCVPALIGD...
>d1jdqa1 d.68.3.3 (A:20-98) Hypothetical protein TM0983 {Thermotoga maritima [TaxId: 2336]}
MAKYQVTKTLDVRGEVCPVPDVETKRALQNMKPGEILEVWIDYPMSKERIPETVKKLGHEVLEIEEVGPSEWKIYIKVK
>d1j7ha_ d.79.1.1 (A:) Conserved 'hypothetical' protein YjgF {Haemophilus influenzae [TaxId: 727]}
MMTQIIHTEKAPAAIGPYVQAVDLGNLVLTSGQIPVNPATGEVPADIVAQARQSLENVKAIIEKAGLTAADIVKTTVFVKDLNDFAAVNAEYERFFKENNHPNFP...
>d4y28j_ f.23.18.0 (J:) automated matches {Pea (Pisum sativum) [TaxId: 3888]}
RDLKTYLSVAPVASTLWFAALAGLLIEINRLFPDALTFPFF
>d1jb0j_ f.23.18.1 (J:) Subunit IX of photosystem I reaction centre, PsaJ {Synechococcus elongatus [TaxId: 32046]}
MKHFLTYLSTAPVLAAIWMTITAGILIEFNRFYPDLLFHPL
>d4kt0j_ f.23.18.1 (J:) automated matches {Synechocystis sp. [TaxId: 1148]}
MDGLKSFLSTAPVMIMALLTFTAGILIEFNRFYPDLLFHP

Astral Database

JOHANNES KEPLER
UNIVERSITY LINZ

- Astral database maintains sequence similarity filtered sequences from SCOP
- 28010 sequences with max 95% similarity

Dataset construction

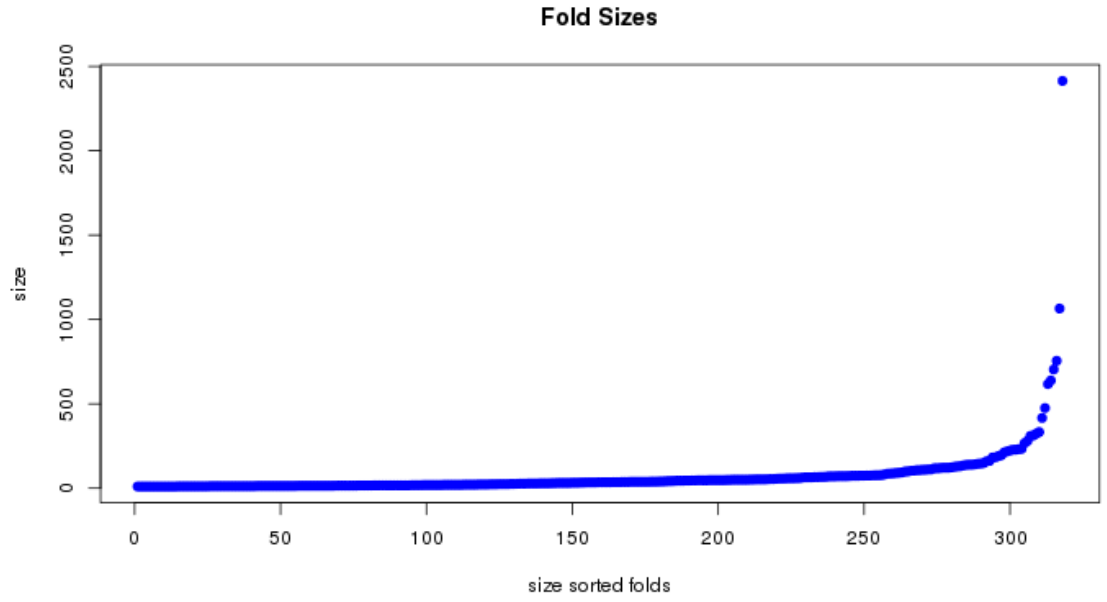
JOHANNES KEPLER
UNIVERSITY LINZ

- Astral 95%
- Discarded 2 classes: Multidomain proteins, small proteins
- 25943 sequences left, 318 folds
- Remote homology detection (divergent evolution)
- Superfamilies have low inter similarities
- Difficult prediction, not possible with simple alignment methods e.g. PSI-BLAST ROC 0.7 ROC50 0.26
- For each fold with > 4 superfamilies, use superfamily with median size within this fold as left out test sequences

Dataset

JOHANNES KEPLER
UNIVERSITY LINZ

- 23435 sequences for training, 318 folds
- 184 sequences for test, 67 folds



Model

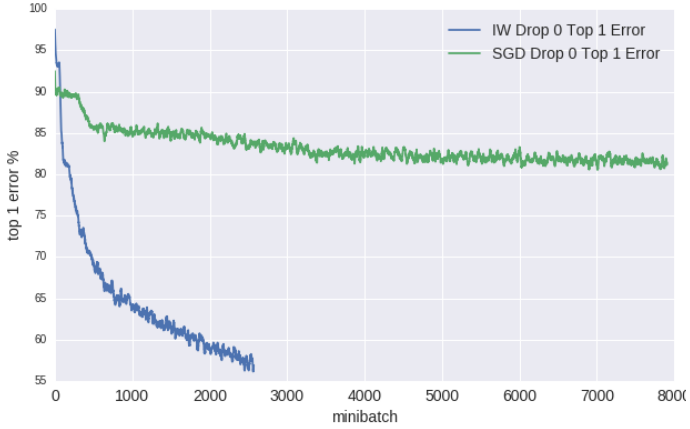
JOHANNES KEPLER
UNIVERSITY LINZ

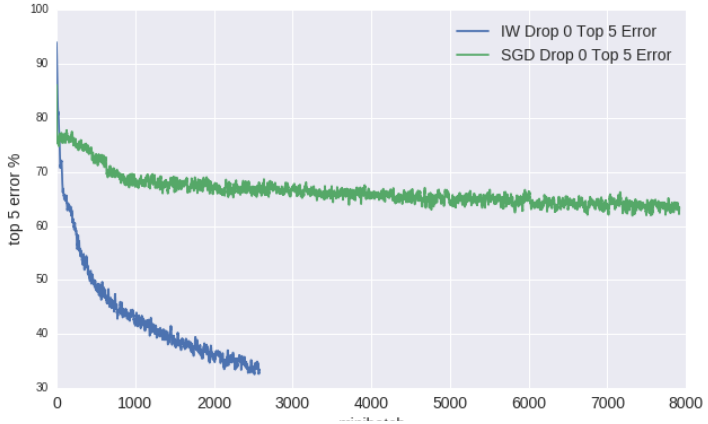
- (j)LSTM
- 200 Memory Cells
- Sliding window size 5
- Minibatch 500
- Learning rate for SGD 0.01, 4 threads
- p 0.1, C 1.0 for Importance Weights, jOptimizer, GPU, 1 thread, p * loss
- dW cut 1.0
- no dropout, no batch normalization

SCOP Fold Prediction
Top 1/5 Error Training/Test

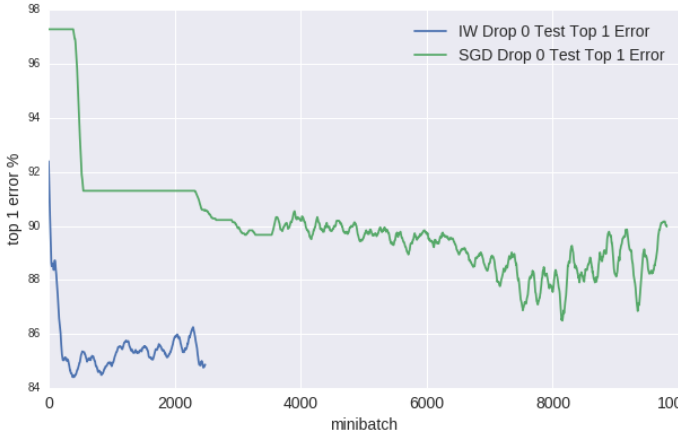
JOHANNES KEPLER UNIVERSITY LINZ

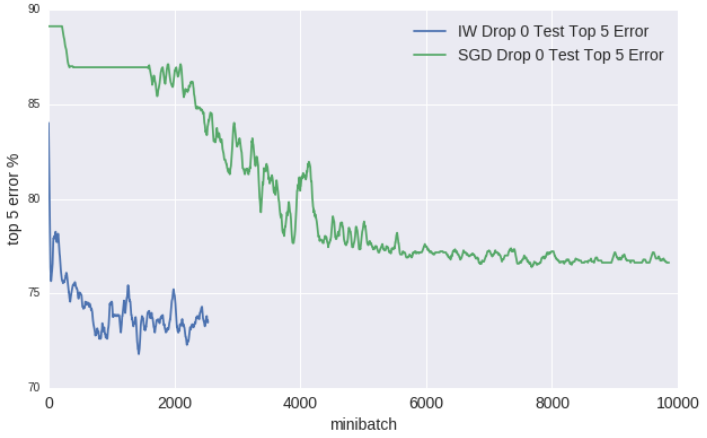
Training





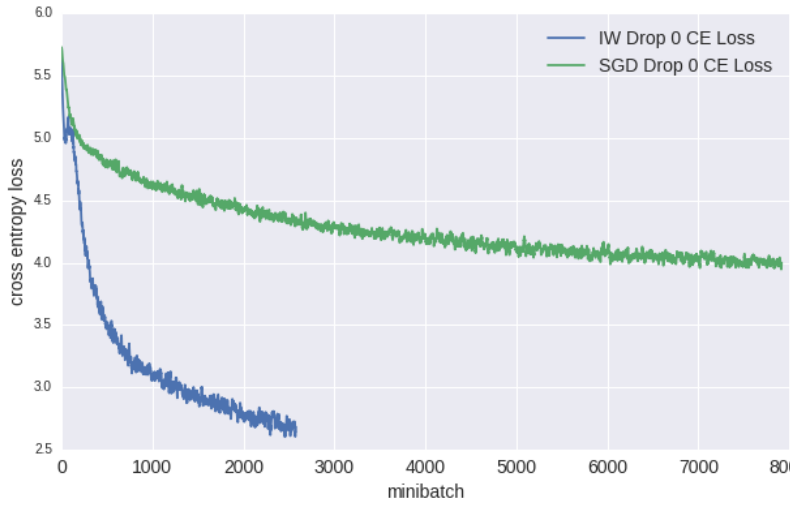
Test



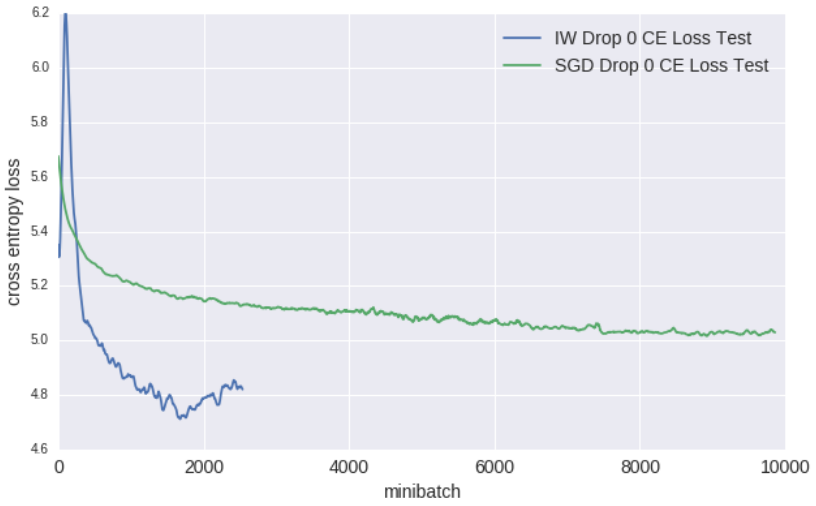


SCOP Fold Prediction
Cross-Entropy Loss Training/Test

JOHANNES KEPLER
UNIVERSITY LINZ



Training



Test

Class Test Accuracies Sorted By Class Size

JOHANNES KEPLER UNIVERSITY LINZ