Extra Articles:

DM-10: Neural Variational Inference and Learning in Belief Networks

- VI for neural-nets, **has a FF net to sample from variational posterior**, jointly trained with model
- No state storage, no mixing issues → better than MCMC • **sample net** → gradient estimates for model and net params
- **Variance reduction techniques:** whitening, CI assumptions for layer • **model + net:** encoder + decoder autoencoder pair

DM-14: Deep Autoregressive Networks

- **Deep stochastic hidden layers** are autoregressed → quick exact ancestral sampling
- **MDL principle** • SGD • **AR:** connections from preceding same-layer and preceding-layer units
  is backpropped

DM-16: Semi-supervised learning with deep generative models

- Generative model is a deep NN • **variational Q** is inference/recognition model • **joint optimization** stochastic backprop
- **Reparameterization trick** of expectations + MC approx • **optimization:** SGD, RMSProp, Adagrad
- Not cheaper than autoregress, but more versatile • **translation-invariant MNIST:** recognizes digits + style
- Posterior on latents used to train classifiers • **M1:** Gaussian • **M2:** Gaussian multinomial

DM-17: Recurrent Models of Visual Attention

- **Attention** formalized as control task • **RNN** selects location to attend to • End-to-end training. Backprop for RNN, policy gradient for RL/control

model
- Glimpse sensor + glimpse network • core network • location network, action network → action (sensor), env. action → label
  next glimpse
- **Reinforce** algorithms

DM-18: Neural Turing Machines

- E2E differentiable • **R/W ops** blurry via attention mechanism • **addressing:** content (cosine similarity) + location
- **Optimization:** RMSProp • working memory system • **tested on:** copy/recall/sort tasks. !!

DM-29 Weight Uncertainty in Neural Networks

Bayes by Backprop

Weights are probability distributions. Posterior of weights given data

Reparameterization trick. Uses vanilla bprop gradients because they work

Prior chosen with cross-validation with Thompson sampling

DM-33b Learning to Transduce with Unbounded Memory

It's LSTM controller + fully diffable stack/queue/dequeue structures

RMSProp training. Gradient clipping. ~~use~~ NLP tasks

DM-39 Spatial Transformer Networks

Diffable, learnable module/layer for nets → convnets

Transformations on feature maps. Localization network + sampling grid + img sampling

DM-45 Grid Long Short Term Memory

Multidimensional LSTM Grid cells connected between layers as well as data space/time dimensions

- **1Dim** grid LSTM is FFNN using LSTM for standard transfer functions. **Relation to highway net**
- 2Dim grid LSTM is stacked LSTM • 3+ Dim grid LSTM is multidim LSTM but is LSTM in all dims, not just depth

DM-48 An Empirical Exploration of RNN Architectures

It's where they 'evolve' RNN cell structures. LSTM forget gate w/ initial bias = 1 → **performs like a GRU**

They find better arch

DM-49 Delving Deep into Rectifiers: Surpassing Human Level Performance on ImageNet Classification

They invent PReLU. Adaptively learns rectifier params. ReLU → PReLU **kicks ass!**

Learned via backprop and momentum SGD

DM-50 ZERO-BIAS AUTOENCODERS AND THE BENEFITS OF COADAPTING FEATURES

NEGATIVE BIASES IN AE ARE BAD. NOT PROBABILISTICALLY JUSTIFIED, BUT WORKS

TRUNCATED RECTIFIER. ON UNREGULARIZED AES (NOT CONTRACTIVE, NOT DENOISING)

TRAINED WITH TRUNCATION, TESTED WITHOUT. PERFORMS ON PAR/BETTER THAN OTHER AES

MOAR-02 A RECURRENT LATENT VARIABLE MODEL FOR SEQUENTIAL DATA

VARIATIONAL RNN: LATENT RV INCLUDED IN DYNAMIC RNN HIDDEN STATE.

HERE VRNN CONTAINS A VAE AT EVERY TIMESTEP, CONDITIONED ON RNN h_{t-1}

TIMESTEP-WISE VARIATIONAL LB. JOINT TRAINING OF GENERATIVE AND INFERENCE MODEL WITH VARIATIONAL LB WRT PARAMS.

SPEECH MODELING, HANDWRITING TASKS

MOAR-03|04 A TUTORIAL ON DEEP LEARNING - PART I, PART II

VANILLA INTRO. NONLINEAR CLASSIFIERS. BACKPROP. AUTOENCODERS. CONVNETS. RNN

MOAR-07 **AUTO-ENCODING VARIATIONAL BAYES**

THE KINGMA OF AUTOENCODERS. **SGVB**

REPARAMETRIZATION TRICK

MOAR-09 **BATCH NORMALIZATION** - ACCELERATING DEEP NN TRAINING BY REDUCING INTERNAL COVARIATE SHIFT

ALLOWS MUCH HIGHER LR → FASTER CONVERGENCE?

MINIBATCH NORMALIZATION $\left(\frac{x-\mu}{\sigma^2}\right)$ · BN TRANSFORM (SCALE + SHIFT) USED AT EVAL/INFERENCE TIME TO FIX VALUES

REDUCES NEED FOR DROPOUT, L2, L1 REGULARIZATION. TESTED ON IMAGENET. 30X SPEEDUPS

- IS IN THE NETWORK, BACKPROP THROUGH NORMALIZATION PARAMS

MOAR-10 BAYESIAN DARK KNOWLEDGE

IS VARIATIONAL AE. SVGD. REPARAM TRICK. DEEP ANN AS APPROXIMATION OF POSTERIOR PREDICTIVE

MCMC + MODEL DISTILLATION; STUDENT NETWORK + TEACHER NETWORK (ENSEMBLE OF)

STUDENT APPROXES DISTRIBUTION OF TEACHERS. GOOD FOR AN MCMC METHOD.

MOAR 17- GATED FEEDBACK RECURRENT NEURAL NETWORKS

GF-RNN, LSTM OR GRU CELLS. FOR LEARNING MULTIPLE TIMESCALES

GATED FEEDBACK CONNECTIONS FROM UPPER LAYERS TO LOWER LAYERS

MOAR-18: GENERATIVE IMAGE MODELING USING SPATIAL LSTM

MULTI/M LSTM, N-PRECEDING STATES CONNECTIVITY. RECURRENCIES EVERYWHERE.

MOAR-19 **HIGHWAY NETWORKS**

GATING UNITS TO REGULATE INFORMATION FLOW THROUGH NETWORK, 100S OF LAYERS.

$y = H(x, W_H) \cdot T(x, W_T) + x \cdot C(x, W_C)$ TRANSFORM GATE, CARRY GATE. BIAS INIT < 0 TO PROMOTE CARRY BEHAVIOR

MOAR-22 MEAN FIELD THEORY FOR SIGMOID BELIEF NETWORKS.

ANCIENT. **PROTO-VARIATIONAL**

MOAR-24 THE WAKE-SLEEP ALGORITHM FOR UNSUPERVISED NEURAL NETWORKS.

GENERAL W/S FORMULATION. ANCIENT

MOAR-30 VISUALIZING AND UNDERSTANDING RECURRENT NEURAL NETWORKS.

DIAGNOSTICS. ANALYSIS OF LSTM AND GRU INTERNAL BEHAVIOR.