Проекты центра глубинного обучения и байесовских методов, использующие вычислительные мощности НИУ ВШЭ

Антон Осокин
ЦГУиБМ, ФКН ВШЭ

NATIONAL RESEARCH UNIVERSITY

p(B|A)yesgroup.ru

BayesGroup: https://bayesgroup.ru/

- Founded in 2007
- Currently consists of 20 students, 6 PhD students, 8 researchers, 2 professors
- Industrial partners: Samsung, Sberbank
- Now located at CS HSE, Samsung AI Center, MSU
- Publications on top ML/CV/NLP conferences:

  NeurIPS, ICML, ICLR, CVPR, ICCV, ACL, EMNLP

  > 20 papers over 5 years (conferences from the list above)

Research topics

- Bayesian models of deep neural networks
  - Variational inference
  - Monte-Carlo Markov Chain (MCMC)
  - Network compression
  - Ensembles
  - Uncertainty estimation
- Predicting complex objects (with neural nets)
  - Predicting structured objects
  - Program analysis and synthesis
- Generative models
  - Generative Adversarial Networks (GAN)
  - Variational Auto-Encoder (VAE)
- Stochastic optimization
  - Riemannian optimization
- Applications
  - Computer vision
  - Natural language processing

Projects on the Cluster of HSE

Cluster users (underlined)

- Bayesian methods:
  - **Large scale study of ensemble of DNNs**

    Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, Dmitry Vetrov
- Predicting complex objects
  - **Optimization w.r.t. permutations**

    Artyom Gadetsky, Kirill Struminsky, Christopher Robinson, Novi Quadrianto, Dmitry Vetrov
  - **Cost-sensitive training for autoregressive models**

    Irina Saparina, Anton Osokin
- Generative models
  - **Semi-Conditional Normalizing Flows for Semi-Supervised Learning**

    Andrei Atanov, Alexandra Volokhova, Arsenii Ashukha, Ivan Sosnovik, Dmitry Vetrov
- Applications:
  - **One-shot object detection in natural images**

    Anton Osokin, Denis Sumin, Vasily Lomakin

Project 1: Ensembles of Deep Networks

Large scale study of ensemble of DNNs

- Ensembles of Neural Nets
- Pros:
  - Superior **predictive performance**
  - Superior **estimates of uncertainty** — crucial for risk-sensitive applications e.g., medicine and self-driving vehicles.
- Cons:
  - Computationally expensive
  - No consistent comparison of ensembling methods
  - No consistent protocol for comparing uncertainty estimators

Large scale study of ensemble of DNNs

- Why do we need HPC?
    - Training even a one DNN is time-consuming, the average classification architecture requires the following training time:
        - Small dataset: 6 hours, 1x Tesla v100
        - Large dataset: 35 hours, 4x Tesla v100
    - Ensemble needs 10-100 models just for a single run!
    - Research requires making hundreds of such runs
    (different datasets, architectures, ensembling techniques, ...)
- Results: Submitted a paper to ICLR - one of the strongest conferences in the field (h5-index 150, h5-median 276).
- Future research: more efficient ensembles on both training and inferences

Conference ranking:
https://scholar.google.com/citations?view_op=top_venues

Project 2: Variational Optimization w.r.t. permutations

Optimization w.r.t. permutations

Many combinatorial optimization problems can be casted to optimization over set of permutations:

- *Ranking*: reorder documents to put the relevant ones on top
- *Traveling Salesman Problem*: find the shortest path visiting all cities
- *Causal Structure Learning*: find best DAG which explains causal relations in given data

Optimization over permutations is hard

Variational optimization

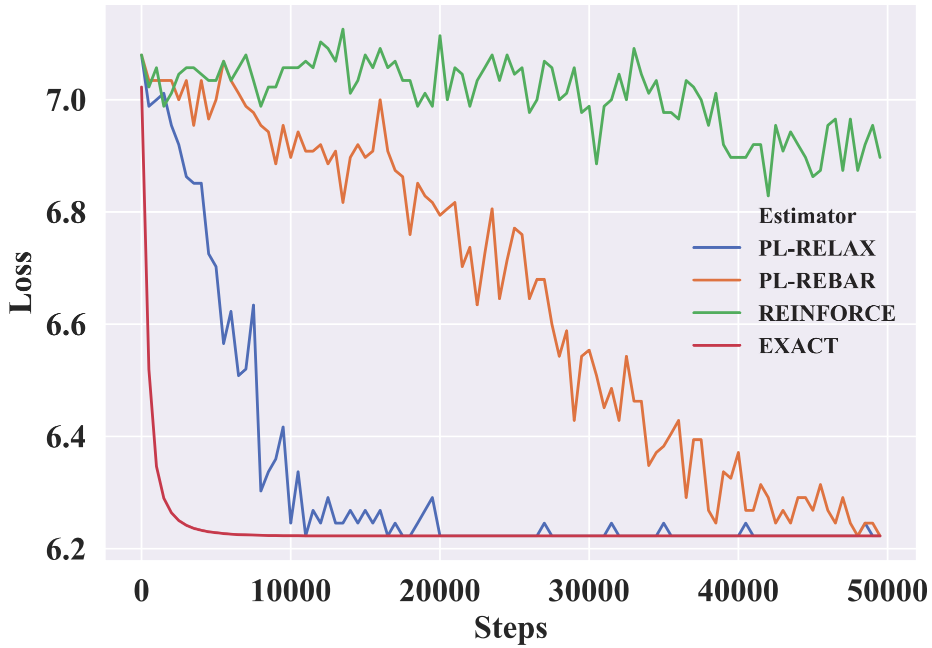
- Instead of optimization w.r.t. permutations apply variational optimization w.r.t. parameters of the distribution on permutations

$$\min_{\pi} f(\pi) \leq \min_{\theta} \mathbb{E}_{p(\pi|\theta)} f(\pi)$$

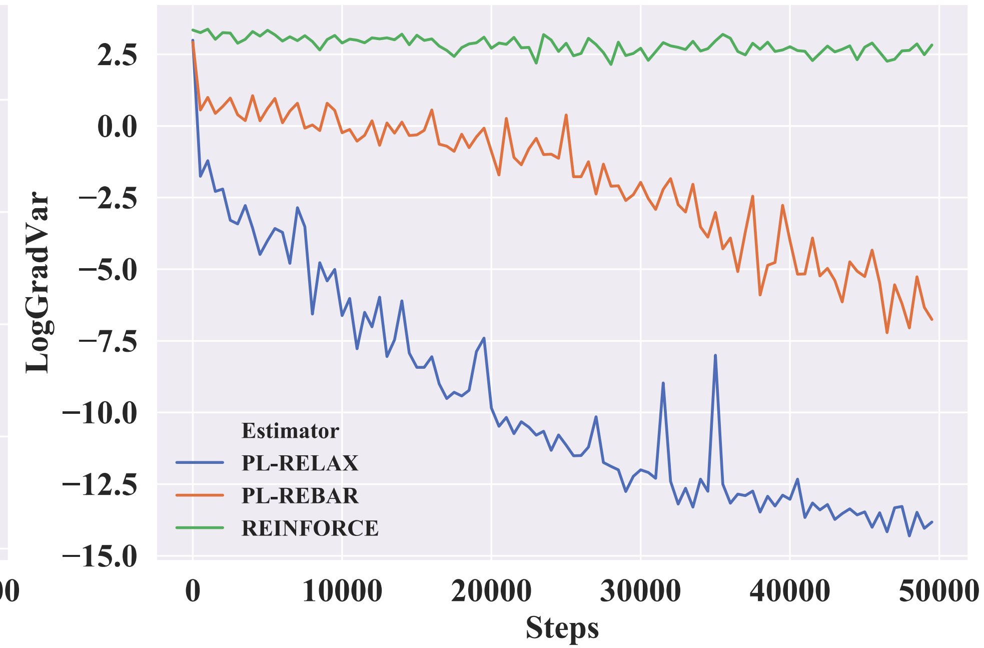
- Use techniques from Reinforcement Learning, RL
- Main difficulty
  - Large variance of gradient estimates
  - Extremely slow convergence!
- This project
  - Control variates for permutations to speed up convergence

Variance reduction of the stochastic gradient

Function optimization



Variance of gradients



Artyom Gadetsky, Kirill Struminsky, Christopher Robinson, Novi Quadrianto, Dmitry Vetrov
“Low-variance Black-box Gradient Estimates for the Plackett-Luce Distribution”, accepted to AAAI 2020
https://arxiv.org/abs/1911.10036

Project 3: Cost-sensitive training for autoregressive models

Autoregressive models

Input

Output sequence

$$\mathbf{x} \Longrightarrow (y_1, \dots, y_{t-1}, y_t, \dots, y_T)$$

autoregressive dependency

Examples:

• machine translation	text	⇒	text
• speech generation	text	⇒	audio wave
• image captioning	image	⇒	text

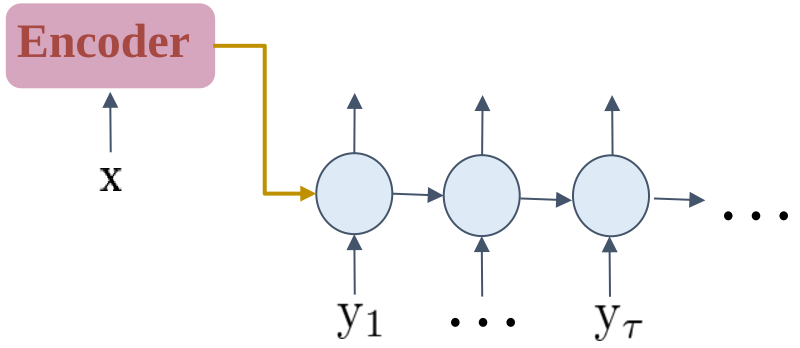
Standard training for autoregressive models

Maximum likelihood estimation (MLE):

$$\log p(\mathbf{y}|\mathbf{x},\theta) = \sum_{t=0}^{\tau-1} \log p(y^{t+1}|y^1,\ldots,y^t,\mathbf{x},\theta) \longrightarrow \max_\theta$$

Problems:

- model never sees its errors during training
- MLE does not depend on test metrics



Decoder: RNN/ Transformer

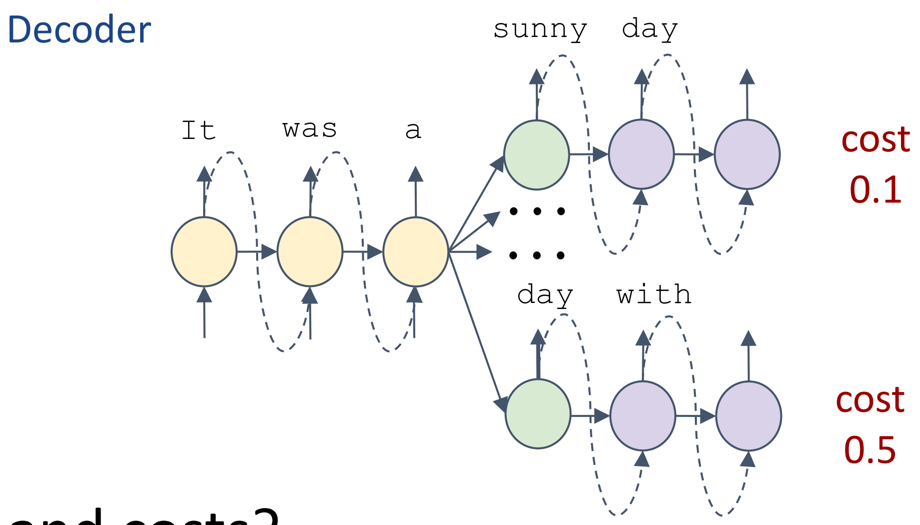
Learning-to-Search approach (SeaRNN)

(Daume III et al., 2009)
(Leblond et al., 2018)

Reduction to cost-sensitive classification

SeaRNN at t-th step:

1. Construct prefix
2. Choose k words to try
3. Complete all prefixes + tries
4. Compute the costs



Our project:

- How should we define reference policy and costs?
- Which loss is better to use?
- How important are the values of costs?

Results: small tasks, small scale NMT (NeurIPS workshop, 2019)

Challenge: scaling to large datasets!

Project 4: Semi-Conditional Normalizing Flows for Semi-Supervised Learning

Semi-Supervised Learning, SSL

- Deep Learning relies on large datasets
- Labelling data is expensive
- There is a lot of unlabelled data
- The idea is to use unlabelled data to learn conditional generative model.

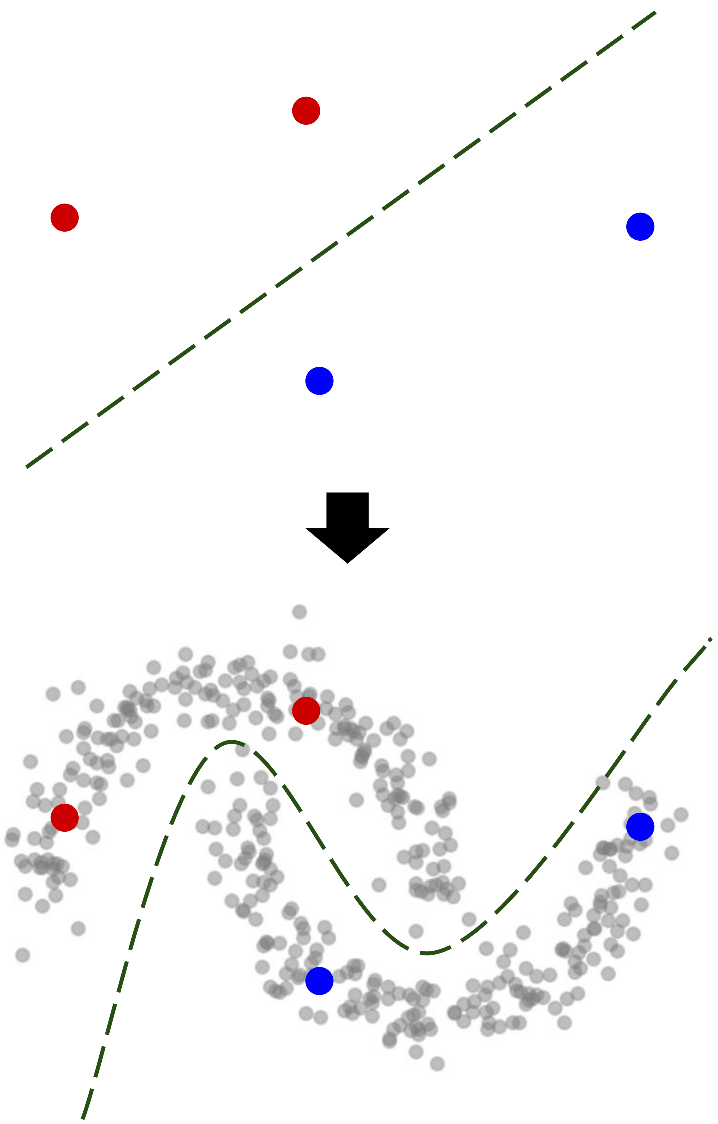
Algorithm

1. Training: learn the conditional model f

data $$L(\theta) = \sum_{(x_i,y_i)\in\mathcal{L}} \log p_\theta(x_i,y_i) + \sum_{x_j\in\mathcal{U}} \log p_\theta(x_j)$$

Labeled data Unlabeled data

2. Prediction: $$p(y|x) = \frac{p_\theta(x|y)p(y)}{\sum_{l=1}^{K} p_\theta(x|l)p(l)}$$



Deep Generative Models for SSL

- How to model distribution $p_\theta(x|y)$
  over high-dimension support (e.g. images and texts)?
- Flexible family of generative models: normalizing flows
- This project:
  - Semi-Conditional Normalizing Flow
  - Results on a MNIST
- To fit the state-of-the-art model on CIFAR we need **5 days of 8 GPUs**

- Will be presented at a workshop of NeurIPS 2019:
  https://arxiv.org/abs/1905.00505

Project 5: One-shot object detection in natural images

Object Detection

- Task: find and label objects
- Neural nets do very well
- High-quality open-source implementations

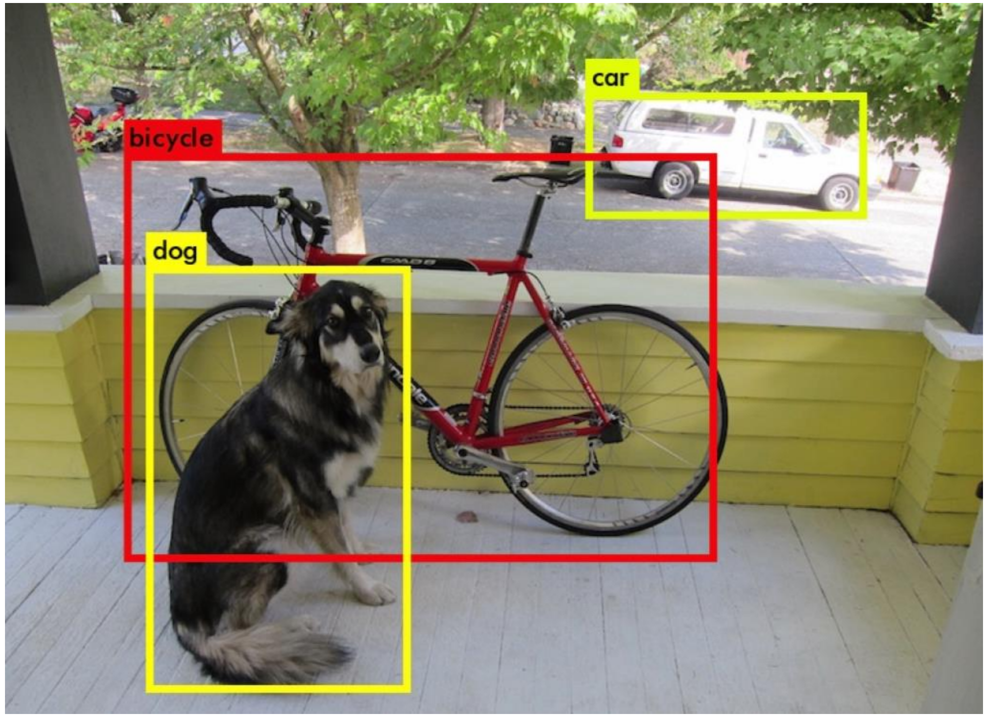


image credit: Joseph Redmon

- Drawback: need a lot of training data:
  - COCO dataset: 80 classes (18k instances per class)
  - 1k images: performance drops to unusable level: to 9 mAP from 36 mAP [Gupta et al., 2019]

One-Shot Object Detection

Detect this:

Here:

Results:

Our project

- SOTA in one-shot detection: RepMet [Karlinsky et al., 2019]
- Two stage method:
  - Detector of all objects
  - Recognition with metric learning

- Ours: One-Stage One-Shot Detector
  - Reiterating old ideas from computer vision, but with deep networks
- Key signal: matching local features

Qualitative example

9

10

Baseline: detector + metric learning

Our method:

Our project

- SOTA in one-shot detection: RepMet [Karlinsky et al., 2019]
- Two stage method:
  - Detector of all objects
  - Recognition with metric learning

- Ours: One-Stage One-Shot Detector
  - Reiterating old ideas from computer vision
- Key signal: matching local features
- Results one three datasets
- Under review for CVPR 2020

Thank you!

Cluster users (underlined)

- Bayesian methods:
  - **Large scale study of ensemble of DNNs**

    Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, Dmitry Vetrov
- Predicting complex objects:
  - **Optimization w.r.t. permutations**

    Artyom Gadetsky, Kirill Struminsky, Christopher Robinson, Novi Quadrianto, Dmitry Vetrov
  - **Cost-sensitive training for autoregressive models**

    Irina Saparina, Anton Osokin
- Generative models
  - **Semi-conditional normalizing flows for semi-supervised learning**

    Andrei Atanov, Alexandra Volokhova, Arsenii Ashukha, Ivan Sosnovik, Dmitry Vetrov
- Applications:
  - **One-shot object detection in natural images**

    Anton Osokin, Denis Sumin, Vasily Lomakin