Survey of Papers on Learned and Random Features, Kernels, Transfer Learning, Uncertainty.

Summary Notes by Max Guo

July 18, 2022

(?) - denotes a lack of familiarity or understanding of a particular concept at time of reading

(?) - denotes a confusion as to why the authors included this

List of papers gone through in this overleaf:

1. Why do better Loss Functions Lead to Less Transferable Features?
2. Conservative Uncertainty Estimation by Fitting Prior Networks
3. On Kernel-Target Alignment
4. An Empirical Study on The Properties of Random Bases for Kernel Methods

1 Paper 1: Why do better Loss Functions Lead to Less Transferable Features?

1.1 Information

Year: 2021.
Conference: NIPS.
Authors: Simon Kornblith, Ting Chen, Honglak Lee, Mohammad Norouzi.
Institutions: Google Research, University of Michigan.

1.2 Research Question

How does the choice of training objective affect the transfer ability of the CNN hidden representations?

1.3 Findings

Setting: ImageNet, classification

- Choice of loss has little effect when networks are fine-tuned on new task.
- Differences among loss functions is only apparent in the last few layers of the network.
- Class separation levels are sensitive to different objectives and hyperparameters.
- **Higher class separation levels → higher accuracy, but then these representations are not**

good for downstream tasks.

2 Paper 2: Conservative Uncertainty Estimation by Fitting Prior Networks

2.1 Information

Year: 2020.
Conference: ICLR.
Authors: Kamil Ciosek, Vincent Fortuin, Ryota Tomioka, Katja Hofmann, Richard Turner
Institutions: Microsoft Research Cambridge, ETH Zurich, University of Cambridge

2.2 Research Problem

How can we estimate high-quality uncertainties from deep neural networks?

(Authors provide new method)

2.3 Criticisms of Related Works

- non-Bayesian inference can be suboptimal, so uncertainty should be related to Bayesian inference (?)
- Deep Ensembles (popular method), in practice, cannot be related to Bayesian inference (no theoretical justification) and may given overconfidence uncertainties.
- Monte Carlo dropout can be viewed as Bayesian inference, but requires some math tricks. Also gives overconfident predictions.
- BNNs: sacrifices some accuracy, training and tuning is hard, good posterior approximation is hard.

2.4 Findings

- **New Method**:
  - Obtain B priors from randomly initialized neural networks
  - Fit B predictor networks independently to each to the priors
  - For uncertainty, take the (average) error between the prior and the fitted predictor networks.
- Note that, like GP inference, uncertainty estimation does not depend on regression label.
- Proof: These estimates are conservative (are greater than a corresponding GP).
- Proof: Uncertainties become smaller with more and more data.

3 Paper 3: On Kernel-Target Alignment

3.1 Information

Year: 2001.
Conference: NIPS.

Authors: Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, Jaz Kandola
Institutions: BIOwulf Technologies, University of London.

3.2 Research Question

How to tell the similarity between two kernels?

3.3 Contributions

The authors introduce kernel-alignment, a metric for similarity between two kernels or between kernel-target.

Kernel Alignment. Given: unlabeled dataset $X = \{x_1, \ldots, x_n\}$, inner product between Gram matrices:

$$\langle K_1, K_2 \rangle_F = \sum_{i,j=1}^{m} K_1(x_i, x_j) K_2(x_i, x_j),$$

then the empirical alignment of a kernel k_1 with kernel k_2 with respect to X is:

$$\hat{A}(X, k_1, k_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}}$$

(also viewed as the cosine of the angles between K_1 and K_2).

Theoretical Results.

- (Concentration) Alignment not too dependent on training set S
- Combining two kernels that are both aligned to the target results in a kernel more aligned to the target.
- (?) Didn't fully understand the Concentration or Algorithm portions.

4 Paper 4: An Empirical Study on The Properties of Random Bases for Kernel Methods

4.1 Information

Year: 2017.
Conference: NIPS.
Authors: Maximilian Alber, Pieter-Jan Kindermans, Kristof T. Schutt, Klaus-Robert Muller, Fei Sha
Institutions: Technische Universitat Berlin

4.2 Research Question

How do random features of approximated kernel machines differ from learned features of neural networks?

4.3 Contributions

Setting: Classification

- Analyze four cases:
  - Random Basis: approximated kernel machine, standard form

  - Unsupervised Adapted Basis: learn basis functions to approximate kernel (no target information)
  - Supervised Adapted Basis: learn basis functions to match labels via kernel-target alignment (yes target information)
  - Discriminatively Adapted Basis: End-to-end trained Neural Network
- SAB and DAB don't approximate a given kernel well (as expected - why should it?), but UAB does better at approximating the kernel than RB
- SAB and DAB generally have better classification accuracies than UAB or RB
- All trends generally decrease in importance as the number of features increases
- For transfer learning (MNIST), $SAB > UAB > DAB > RB$, but trend also decreases as the number of features increases.