Deep Kernel Learning for Information Extraction from Cancer Pathology Reports

Devanshu Agrawal
Abhishek Dubey
Georgia Tourassi
Jacob Hinkle

November 17, 2019

Objective

- National cancer surveillance
  - Important for cancer research, funding, and legislation
- The Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute (NCI)
  - Goal: To curate a database of all cancers diagnosed in the US
- Cancer pathology reports
  - Contain tumor information such as location.
  - Manual extraction is costly
- Build an automated information extraction pipeline for pathology reports

Breast ICD-0-3 topographical sites

Primary Site

C500 Nipple (areolar)
Paget disease without underlying tumor

C501 Central portion of breast (subareolar) area extending 1 cm around areolar complex
Retroareolar
Infraareolar
Next to areola, NOS
Behind, beneath, under, underneath, next to, above, cephalad to, or below nipple
Paget disease with underlying tumor

C502 Upper inner quadrant (UIQ) of breast
Superior medial
Upper medial
Superior inner

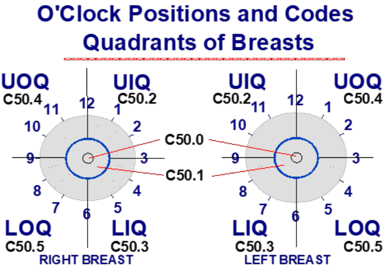
C503 Lower inner quadrant (LIQ) of breast
Inferior medial
Lower medial
Inferior inner

C504 Upper outer quadrant (UOQ) of breast
Superior lateral
Superior outer
Upper lateral

C505 Lower outer quadrant (LOQ) of breast
Inferior lateral
Inferior outer
Lower lateral

C506 Axillary tail of breast
Tail of breast, NOS
Tail of Spence

C508 Overlapping lesion of breast
Inferior breast, NOS
Inner breast, NOS
Lateral breast, NOS
Lower breast, NOS
Medial breast, NOS
Midline breast NOS
Outer breast NOS
Superior breast, NOS
Upper breast, NOS
3:00, 6:00, 9:00, 12:00 o'clock



Source: `https://seer.cancer.gov/manuals/2018/AppendixC/Coding_Guidelines_Breast_2018.pdf`

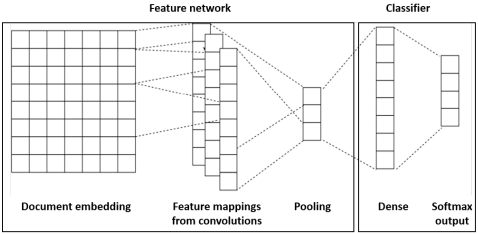
Information extraction methods

- Rule-based methods[1]
- Methods with manually crafted features (machine learning)[2]
  - Cast extraction problem as text classification
  - Logistic regression
  - Support vector machines
- Methods with automated feature extraction (deep learning)[3]
  - Convolutional neural networks (CNN)

[1] Nguyen et al. JAMIA (2010).
[2] Li and Martinez. ALTA (2010).
[3] Qiu et al. JBHI (2017).

Convolutional neural networks

- Shallow-wide architecture[4]
  - Word embedding layer
  - Parallel convolutional filter banks
  - Rectified linear unit (ReLU) activation
  - Global max pooling
  - Softmax classifier



Modified from

https://ieeexplore.ieee.org/abstract/document/7918552

[4] Kim. EMNLP (2014).

Key challenges

- Limitations of CNN:
  - Limited performance when training data is scarce
    - Heavy class imbalance
  - Limited uncertainty quantification
- Solution: Deep kernel learning (DKL)
  - Composition of a CNN feature network with a Gaussian process (GP)
  - Bayesian but scalable and expressive
  - Has been applied in computer vision[5]
  - First application to text classification

[5] Wilson et al. NIPS (2016).

Gaussian processes

- GP classifier:

$$\mathbf{y} = \mathbf{g}(\mathbf{f}(\mathbf{x})). \qquad \text{(GP classifier)}$$

- Latent GP defines prior over functions
- $f \sim \mathcal{GP}(\mu, k)$ iff $[f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]$ follows a joint normal distribution for any n inputs,

$$\mathrm{E}[f(x)] = \mu(x) \text{ and } \mathrm{Cov}[f(x), f(x')] = k(x, x').$$

- Inverse-link functions $\mathbf{g}$:

$$\mathrm{softmax}(\mathbf{z})|_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}.$$

$$\mathrm{robustmax}(\mathbf{z})|_i = \begin{cases} 1-\varepsilon & \text{if } i = \arg\max(\mathbf{z}) \\ \frac{\varepsilon}{C-1} & \text{otherwise.} \end{cases}$$

- Kernels:

$$k_{\mathrm{rbf}}(\mathbf{x}, \mathbf{x}') = \sigma^2 e^{-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\lambda^2}}.$$

$$k_{\mathrm{lin}}(\mathbf{x}, \mathbf{x}') = \sigma^2 \mathbf{x}^\top \mathbf{x}'.$$

Sparse variational Gaussian processes

- Challenges:
  - Exact GP inference not possible with classification inverse-link functions
  - GP inference has complexity $O(N^3)$ for N training points
- Solution: Sparse variational GP (SVGP)[6]
  - Variational inference
  - Inducing points
- Maximize evidence lower bound (ELBO) with respect to:
  - Inducing points
  - Variational parameters (values at inducing points)
  - Kernel hyperparameters
- ELBO naturally includes regularization

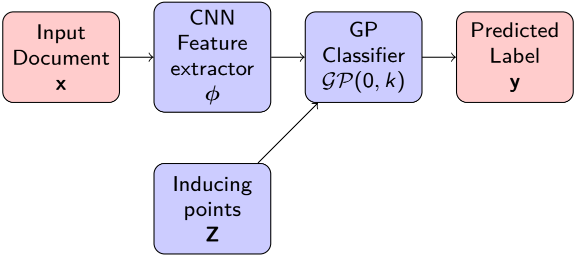
[6] Hensman et al. JMLR (2015).

Deep kernel learning

- Deep kernel:

$$k_{\text{deep}}(\mathbf{x}, \mathbf{x}') = k(\phi(\mathbf{x}; \omega), \phi(\mathbf{x}'; \omega)).$$

- Inducing points live in feature space, not input space
  - Necessary for text input



Datasets

- Primary tumor site extraction from de-identified electronic pathology reports (EPR)
  - Gathered from five SEER cancer registries (CT, HI, KY, NM, Seattle)
  - Three breast and three lung tumor sites
  - Used 10-fold cross validation for EPR

Dataset	Classes	Training points per class	Test points per class
EPR	6	123.75	13.75
20News-22[7]	20	124.45	376.6
IMDB-1[8]	2	125.0	12 500.0
20News-100	20	565.7	376.6
IMDB-5	2	625.0	12 500.0
IMDB-100	2	12 500.0	12 500.0

[7] Dua and Graff. (2017).
[8] Maas et al. ACL (2011).

Models

For the last two models, "Dense layers" and "Kernel" apply to the classifier used at test time, not during training.

Model	Dense layers	Kernel	Options for pretraining	Fixed Features
CNN-1	1		CNN-1, best DKL	
CNN-2	2		CNN-2, best DKL	
DKL-lin		Linear	CNN-1	
DKL-RBF		RBF	CNN-1	
CNN-SVGP		Linear		CNN-1
DKL-LSC	1			best DKL

Generalization error

Mean test F_{micro} scores (as percentages) with standard deviations across 20 random seeds.

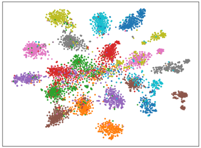
Dataset	CNN-1 (scratch)	CNN-1 (pre. CNN)	DKL-lin (scratch)	DKL-lin (pre. CNN)
EPR	86.2 ± 0.6	86.5 ± 0.6	83.8 ± 0.7	86.6 ± 0.4
20News-22	68.9 ± 1.0	67.9 ± 0.8	75.7 ± 0.8	70.8 ± 0.7
IMDB-1	72.8 ± 1.7	72.1 ± 1.7	73.5 ± 3.2	73.8 ± 1.9
20News-100	79.0 ± 0.4	78.7 ± 0.5	83.4 ± 0.5	82.6 ± 0.5
IMDB-5	79.2 ± 0.8	77.6 ± 0.7	82.9 ± 0.5	77.1 ± 1.0
IMDB-100	88.7 ± 0.3	88.3 ± 0.4	89.1 ± 0.3	88.9 ± 0.3

DKL extracts better features

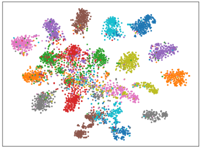
Mean test F_{micro} scores (as percentages) with standard deviations across 20 random seeds.

Dataset	DKL-lin (best)	DKL-LSC	CNN-1 (pre. DKL)	CNN-SVGP
EPR	86.6 ± 0.4	85.3 ± 0.6	86.6 ± 0.7	72.0 ± 2.2
20News-22	75.7 ± 0.8	74.8 ± 1.0	70.0 ± 0.9	69.0 ± 0.9
IMDB-1	73.8 ± 1.9	73.9 ± 2.0	69.9 ± 1.9	57.0 ± 4.6
20News-100	83.4 ± 0.5	83.1 ± 0.5	79.0 ± 0.5	78.6 ± 0.7
IMDB-5	82.9 ± 0.5	83.2 ± 0.5	77.5 ± 0.7	79.0 ± 1.5
IMDB-100	89.1 ± 0.3	89.2 ± 0.2	88.8 ± 0.2	88.5 ± 0.3

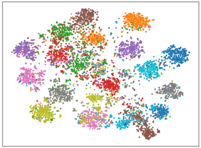
DKL extracts better features: A visualization



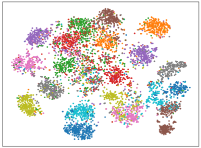
(a) DKL-lin



(b) DKL-lin (pre. CNN)



(c) CNN-1 (pre. DKL)

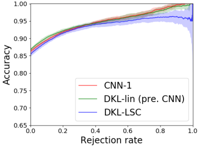


(d) CNN-1

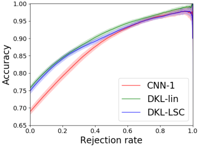
t-SNE visualizations of the 20-Newsgroups test set passed through the feature networks of four different models trained on 20-Newsgroups-100, for the seed giving the biggest performance gap between the CNN-1 and DKL-lin models.

Uncertainty quantification

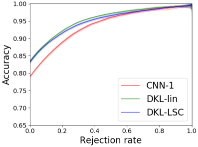
- Confidence score: Probability of predicted class



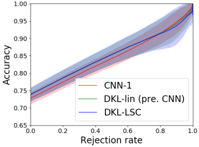
(a) EPR



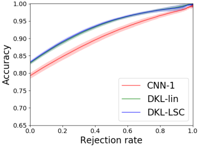
(b) 20Newsgroups-22



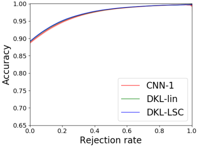
(c) 20Newsgroups-100



(d) IMDB-1



(e) IMDB-5



(f) IMDB-100

Accuracy-rejection curves (ARCs) averaged vertically over 20 random seeds.

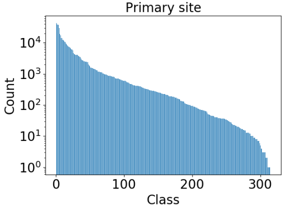
Conclusions and future work

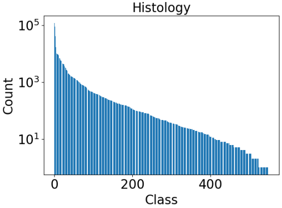
- DKL can be beneficial for text classification
  - Potential for information extraction from pathology reports
- DKL improves feature extraction
- Uncertainty quantification
  - Anomaly detection
- DKL on larger datasets
  - Less beneficial
  - Could remain relevant if there is heavy class imbalance

Future work on bigger data

- 546,981 pathology reports from KY and LA SEER cancer registries
- Multitask with extreme class imbalance
- Modeling class-level and task-level covariance with DKL
- Using Summit

Task	Num. Classes
Primary site	314
Laterality	7
Grade	9
Histology	547
Behavior	4





Funding disclosure

This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health.
This work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DEAC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725.

Evidence lower bound

$$\mathrm{ELBO}(\boldsymbol{\theta}, \mathbf{Z}, \boldsymbol{gamma}) = \sum_{i=1}^{N} \int \ln[p(\mathbf{y}_i \mid \mathbf{f}_i)] \; q(\mathbf{f}_i; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{f}_i - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{U}; \boldsymbol{\theta}) \mid p(\mathbf{U})).$$