| Deep Bayesian Active Learning with Image Data |

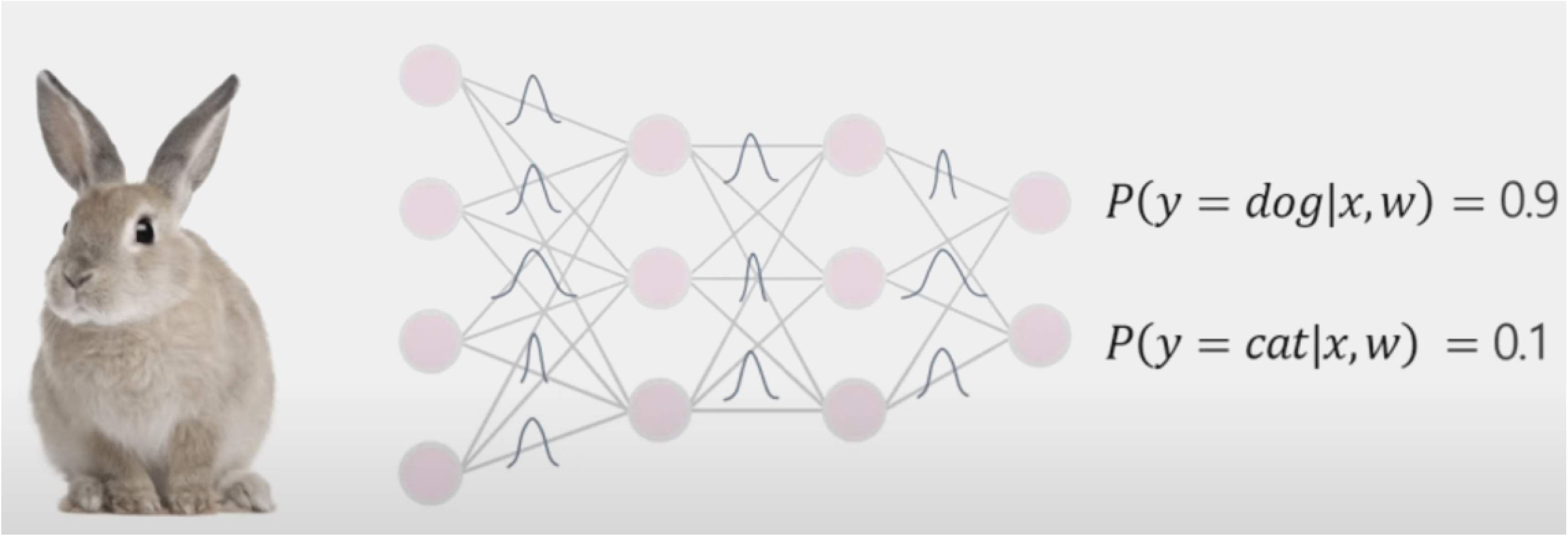
YBIGTA
21기
DA

나준호

2022.05.16

Uncertainty

- Over-confidence 문제
- Uncertainty 정량화 필요성
- Distribution의 variance



Bayesian Neural Network(BNN)

- 저자는 또 다른 논문에서 posterior를 추정을 MC dropout과 L2 regularization로 쉽게 구할 수 있음을 증명함.
- Prior distribution은 보통 Gaussian으로 둠(mean=0,variance=1).
- 최종적인 값은 여러 예측 값의 평균
- Uncertainty는 variance

$$\underbrace{p(W|X,Y)}_{\text{Posterior}} = \frac{\overbrace{p(Y|X,W)}^{\text{Likelihood}}\overbrace{p(w)}^{\text{Prior}}}{\underbrace{p(Y|X)}_{\text{Evidence}}}$$

$$\underbrace{p(Y|X)}_{\text{Evidence}} = \int p(Y|X,W)p(W)dw$$

$$p(y = c|\mathbf{x}, \boldsymbol{\omega}) = \text{softmax}(\mathbf{f}^{\boldsymbol{\omega}}(\mathbf{x}))$$

ELBO $\int q_\theta(\mathbf{w}) \log p(\mathbf{Y}|\mathbf{X}, \mathbf{w}) d\mathbf{w} - \mathrm{KL}(q_\theta(\mathbf{w})||p(\mathbf{w}))$

$$Minimize \; -\frac{1}{M}\sum_{i \in S} ln(p(y_i|f^{g(\theta,\hat{\epsilon})}(x_i))) + \lambda_1 ||M_1||^2 + \lambda_2 ||M_2||^2 + \lambda_3 ||b||^2$$

$$\mathbb{E}[\mathbf{y}^*] \approx \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{y}}_t^*(\mathbf{x}^*)$$

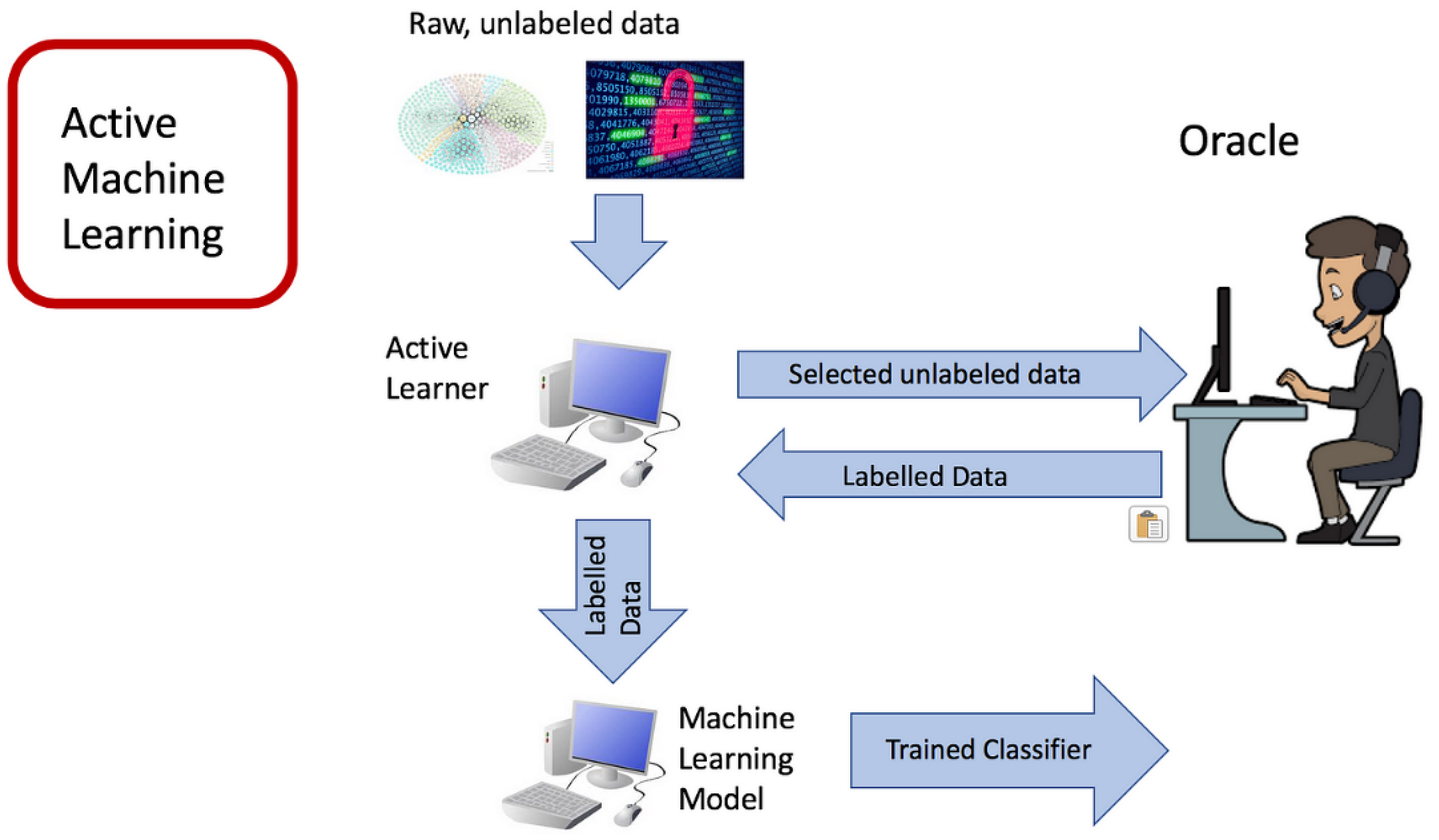
$$\mathrm{Var}[\mathbf{y}^*] \approx \tau^{-1}\mathbf{I}_D + \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{y}}_t^*(\mathbf{x}^*)^T \hat{\mathbf{y}}_t^*(\mathbf{x}^*) - \mathbb{E}[\mathbf{y}^*]^T \mathbb{E}[\mathbf{y}^*].$$

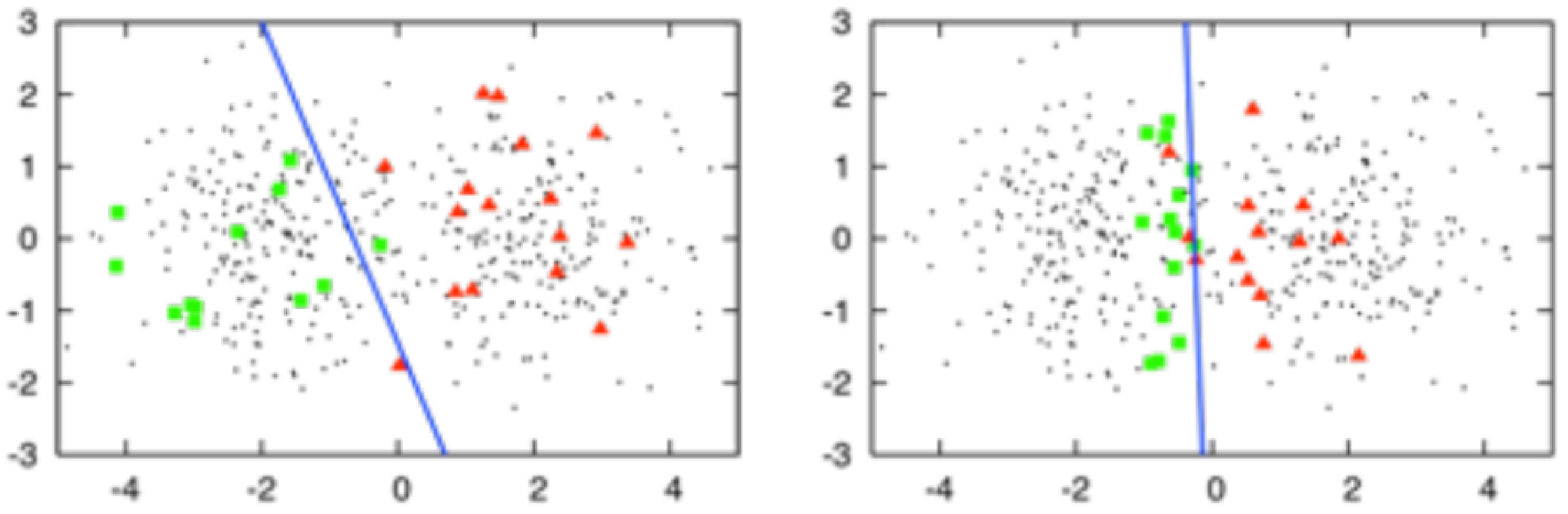
Bayesian Neural Network(BNN)

- Deterministic Neural Network vs. Bayesian Neural Network

Active Learning

- 사람이 모두 라벨링하는 기존의 Passive Learning과 다름
- 모델은 Labeling할 데이터를 제시 (Enrich해줄 수 있는 데이터)
- 하지만, AutoLabeling과 달리 마지막엔 사람이 labeling





random sampling
30 labeled instances
(accuracy=0.7)

active learning
30 labeled instances
(accuracy=0.9)

Active Learning with BNN

- 선행 연구들은 저차원 데이터, kernel method만 대부분 사용하여 Active learning
- 그러나 이 경우엔 고차원 데이터를 다루지 못함.
- 본 논문에서는 BNN을 활용하여 이미지 데이터를 다룸.
- Acquisition Function
  - Variantion Ratios
  - Max Entropy
  - BALD
  - Mean STD

$$x^* = \operatorname{argmax}_{x \in \mathcal{D}_{\text{pool}}} a(x, \mathcal{M}).$$

$$\text{variation-ratio}[\mathbf{x}] := 1 - \max_y p(y|x, \mathcal{D}_{\text{train}})$$

$$\mathbb{H}[y|\mathbf{x}, \mathcal{D}_{\text{train}}] := - \sum_c p(y = c|\mathbf{x}, \mathcal{D}_{\text{train}}) \log p(y = c|\mathbf{x}, \mathcal{D}_{\text{train}}).$$

$$\mathbb{I}[y, \boldsymbol{\omega}|\mathbf{x}, \mathcal{D}_{\text{train}}] = \mathbb{H}[y|\mathbf{x}, \mathcal{D}_{\text{train}}] - \mathbb{E}_{p(\boldsymbol{\omega}|\mathcal{D}_{\text{train}})}\left[\mathbb{H}[y|\mathbf{x}, \boldsymbol{\omega}]\right]$$

$$\sigma_c = \sqrt{\mathbb{E}_{q(\boldsymbol{\omega})}[p(y = c|\mathbf{x}, \boldsymbol{\omega})^2] - \mathbb{E}_{q(\boldsymbol{\omega})}[p(y = c|\mathbf{x}, \boldsymbol{\omega})]^2}$$

$$\sigma(\mathbf{x}) = \frac{1}{C} \sum_c \sigma_c$$

Experiments(MNIST)

- convolution-reluconvolution-relu-max pooling-dropout-dense-relu-dropoutdense-softmax
- 32 convolution kernels, 4x4 kernel size,2x2 pooling, dense layer with 128 units, dropout probabilities0.25 and 0.5
- acquisition until 1000
- Comparison to semi-supervised learning

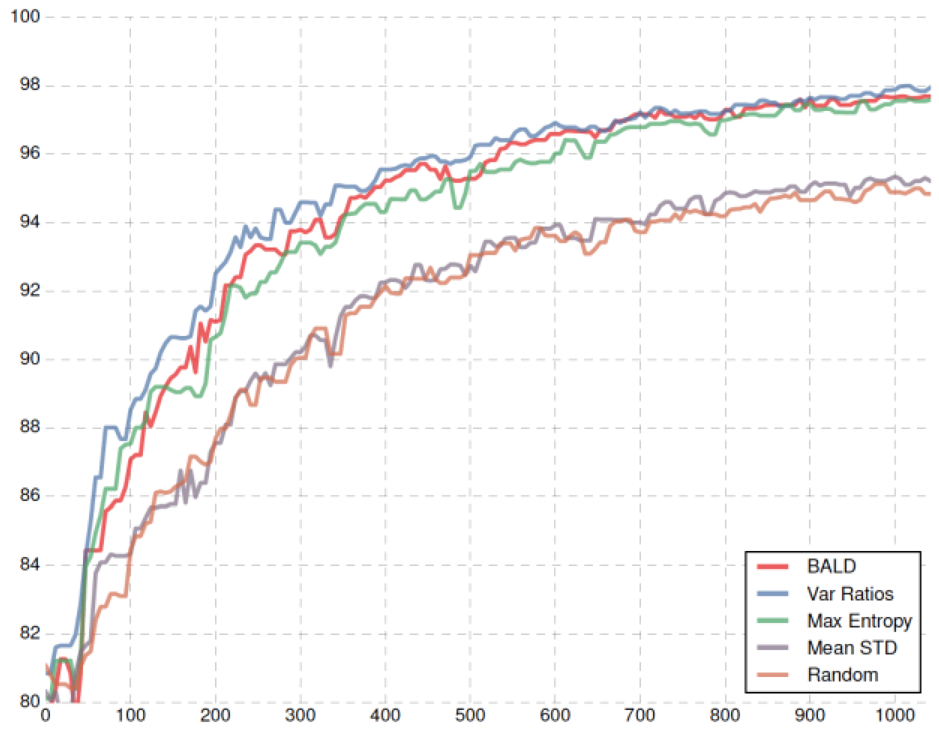
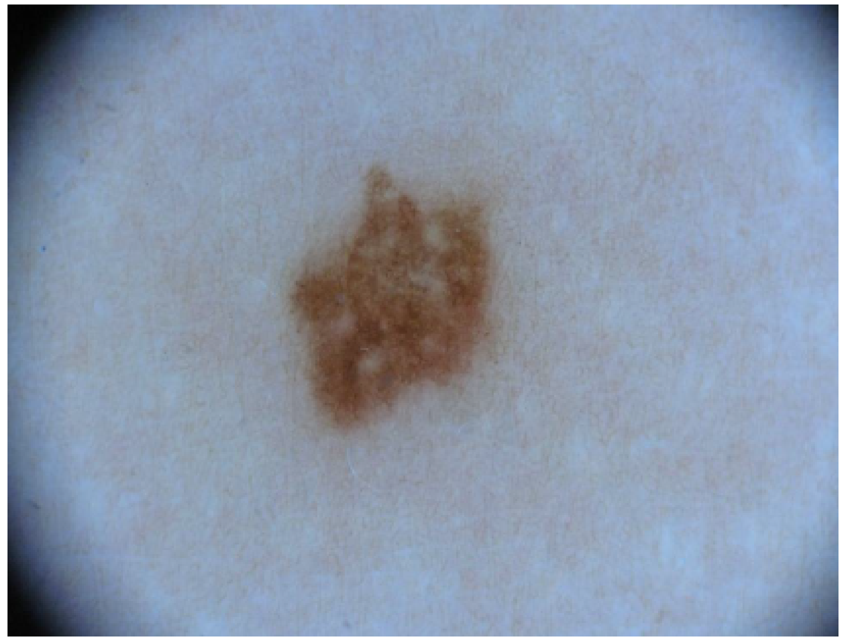


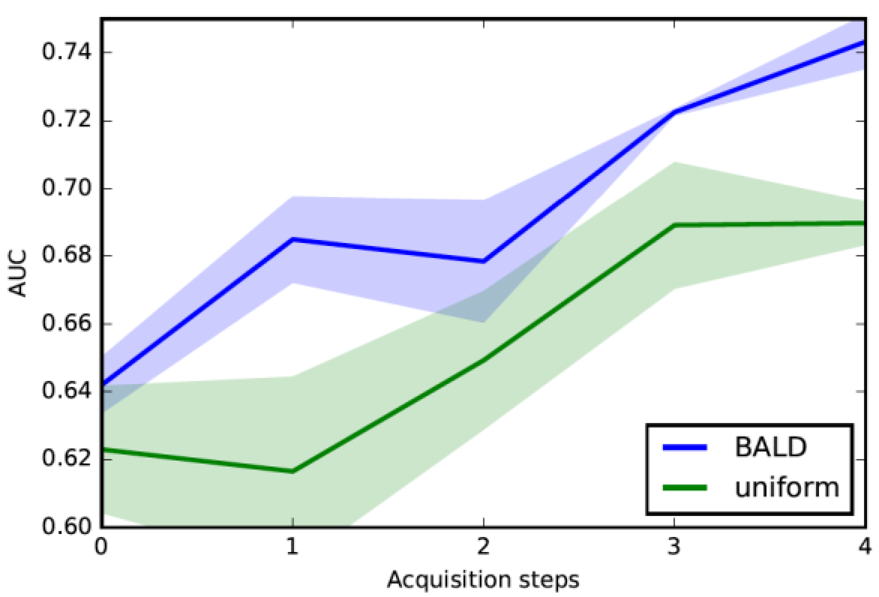
Figure 1. **MNIST test accuracy as a function of number of acquired images from the pool set** (up to 1000 images, using validation set size 100, and averaged over 3 repetitions). Four acquisition functions (*BALD*, *Variation Ratios*, *Max Entropy*, and *Mean STD*) are evaluated and compared to a *Random* acquisition function.

Technique	Test error
Semi-supervised:	
Semi-sup. Embedding (Weston et al., 2012)	5.73%
Transductive SVM (Weston et al., 2012)	5.38%
MTC (Rifai et al., 2011)	3.64%
Pseudo-label (Lee, 2013)	3.46%
AtlasRBF (Pitelis et al., 2014)	3.68%
DGN (Kingma et al., 2014)	2.40%
Virtual Adversarial (Miyato et al., 2015)	1.32%
Ladder Network (Γ-model) (Rasmus et al., 2015)	1.53%
Ladder Network (full) (Rasmus et al., 2015)	0.84%
Active learning with various acquisitions:	
Random	4.66%
BALD	1.80%
Max Entropy	1.74%
Var Ratios	1.64%

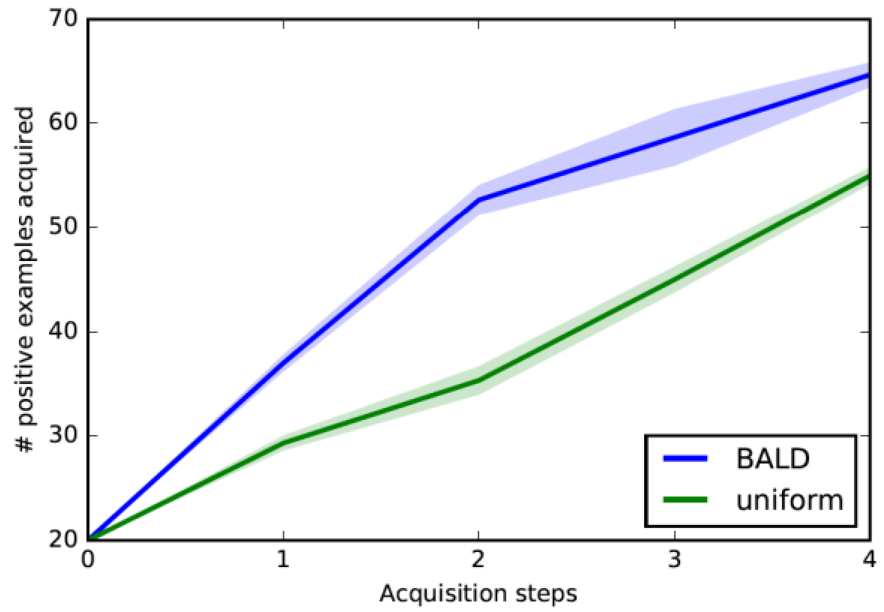
Experiments(ISIC 2016 melanoma diagnosis dataset)

- Model : VGG16 (Simonyan & Zisserman,2015) / 2 classification (malignant/benign)
- Small and imbalanced dataset
- initial dataset : 80 negative / 20 positive
- acquisition until all pool sets have been exhausted

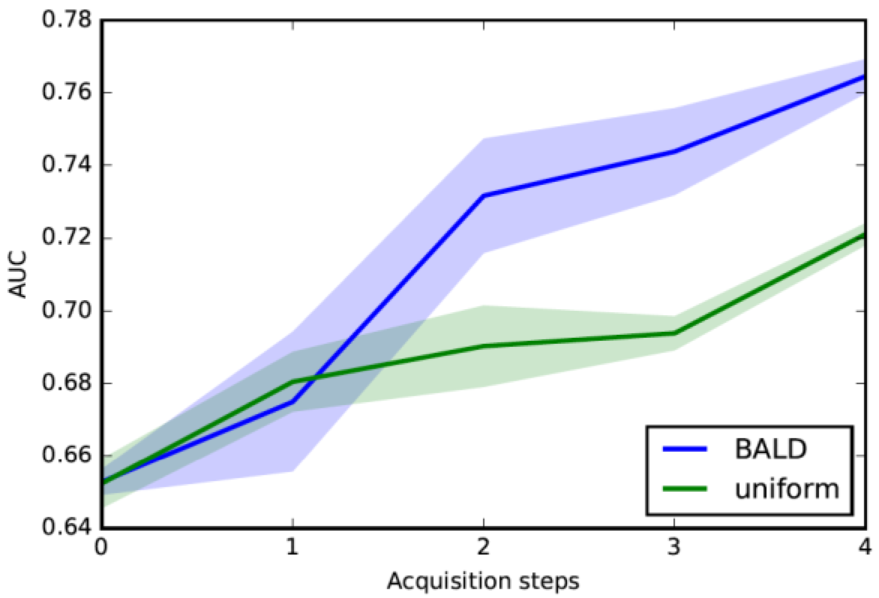




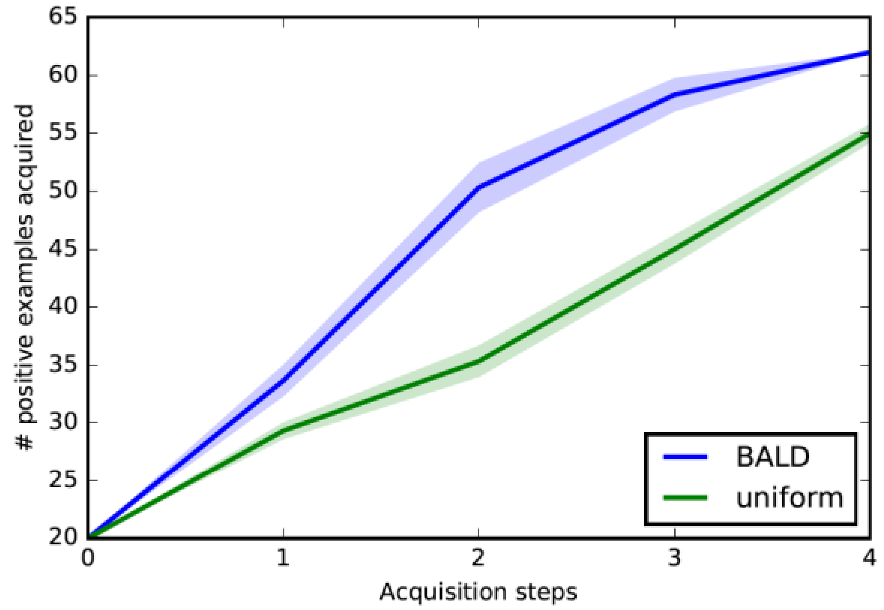
(a) AUC as a function of acquisition step, first test split



(b) # of positive examples as a function of acquisition step, first test split



(c) AUC as a function of acquisition step, second test split



(d) # of positive examples as a function of acquisition step, second test split

Discussion

- 두 번째 Real dataset에서 Acquisition function 간의 성능 차이를 보여주기보다는 Active learning을 통해서 학습했을 때와 전체 데이터셋을 사용해서 학습했을 때의 성능 차이를 보여주는 게 더 낫지 않았을까
- 성능 차이가 그렇게 크지 않다면 Semi-supervised learning보다 확실히 나은 점이 무엇일까

Reference...

- Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning
  - Yarin Gal,2016
- Active Learning for (Bayesian) Neural Networks
  - Lukas Erlenbach, https://towardsdatascience.com/active-learning-for-bayesian-neural-networks-b8471212850f
- [Open DMQA Seminar] Uncertainty Quantification in Deep Learning
  - https://www.youtube.com/watch?v=oCVEFv5P088&t=3421s
- Bayesian Deep Learning Course
  - Seong-Jun Choi, https://www.edwith.org/bayesiandeeplearning