Model Analysis and Explanation

모델 분석과 설명

Model analysis level of abstraction

- 한번에 분석하기 어려우므로 단계 별로

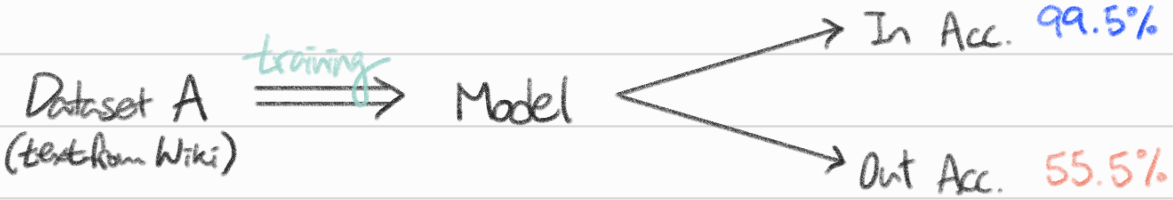
1. As a probability distribution and decision function $P_{model}(y|x)$

2. As a sequence of vector representations in depth and time

   Layer 2
   Layer 1

3. Parameter weights, sepcific mechanisms (attention), dropout, ...

Out-of-domain evaluation sets



In-domain data : Text data from Wikipedia

Out-of-domain data : Except Wikipedia data

task 학습이 아닌
data를 단순히 fitting

NLI (Natural Language Inference)

- 문장 (premise)이 주어졌을 때 다른 문장 (hypothesis)에 의미적으로 수반?

  entail

Entailment, Neutral, Contradiction

같은 의미 무관함 모순

HANS (검증용 dataset)

Influence studies and adversarial examples

Saliency map : model의 prediction에 input이 얼마나 영향 미쳤는지 scoring
=> Simple gradient method

Rubbish example
- Beam search, saliency map 반복 적용해
  가장 낮은 중요도 word 삭제

Add any, Add sent

Analyzing representations

Attention map 이용해 각 layer가 어떤 역할을 하는지 추측

Revisiting model ablations as analysis

제거

Layer를 깊게 vs 얕게 ?

Model accuracy 상승위해 Model tuning
→ 일련의 tuning 과정을
analysis 관점으로 해석 가능