04 Improving the accuracy of deep learning

- Larger capacity output function for deep learning -

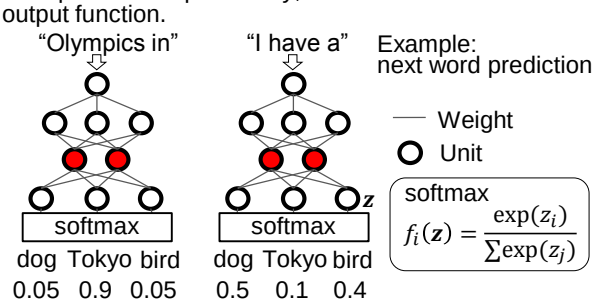
Abstract

Deep learning is used in a lot of applications, e.g., image recognition, speech recognition, and machine translation. In many applications of deep learning, softmax is used as an output activation function for modeling categorical probability distributions. To represent various probabilities, models should output various patterns, i.e., models should have sufficient representation capacity. However, softmax can be a bottleneck of representational capacity (the softmax bottleneck) under a certain condition. In order to break the softmax bottleneck, we propose a novel output activation function: sigsoftmax. To break the softmax bottleneck, sigsoftmax is composed of sigmoid and exponential functions. Sigsoftmax can output more various patterns than softmax without additional parameters and additional computation costs. As a result, the model with sigsoftmax can be more accurate than that with softmax.

Deep Learning

Deep learning is used in a lot of applications.
(e.g., image recognition or machine translation)

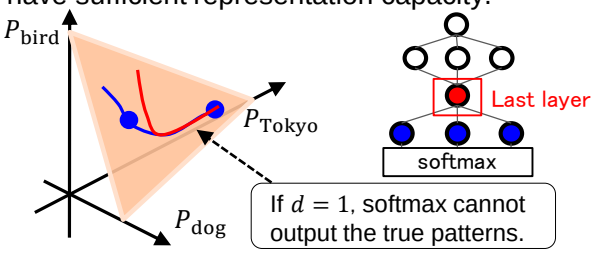
To represent the probability, softmax is used as an output function.



For accurate prediction, models should output a lot of patterns of probabilities, i.e., models require sufficient representation capacity.

Bottleneck of Representation

If the number of units in the last layer d < the number of outputs M -1, softmax does not have sufficient representation capacity.



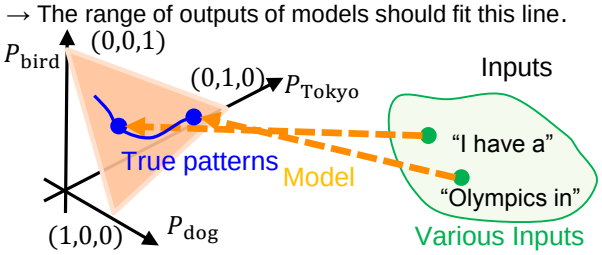
In natural language processing, M = the number of vocabulary is very large, and a lot of parameters are required if we set $d = M$-1.

Problem

We assume the number of outputs M is 3.

Probability of outputs: $(P_{\text{dog}}, P_{\text{Tokyo}}, P_{\text{bird}})$
→ It can be represented as a point on the triangle.
Models connect these points and inputs.

We assume the blue line represents the true patterns of probabilities (output set of various inputs).
→ The range of outputs of models should fit this line.



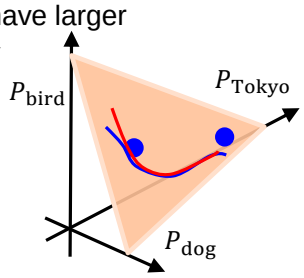
Novel output function: sigsoftmax

Multiplying sigmoid to have larger representation capacity

sigsoftmax

$$f_i(\boldsymbol{z}) = \frac{\sigma(z_i)\exp(z_i)}{\sum \sigma(z_j)\exp(z_j)}$$

$\sigma(z) = \frac{1}{1-\exp(-z)}$: sigmoid
monotonically increasing to 1 from 0



Sigsoftmax has a larger representation capacity than softmax[1] without additional parameters and computation costs.

References

[1] S. Kanai, Y. Fujiwara, Y. Yamanaka, S. Adachi, "Sigsoftmax: Reanalysis of the softmax bottleneck," in Proc. *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Contact

Sekitoshi Kanai Email: cs-liaison-ml at hco.ntt.co.jp
NTT Software Innovation Center

Innovative R&D by NTT
Open House 2019