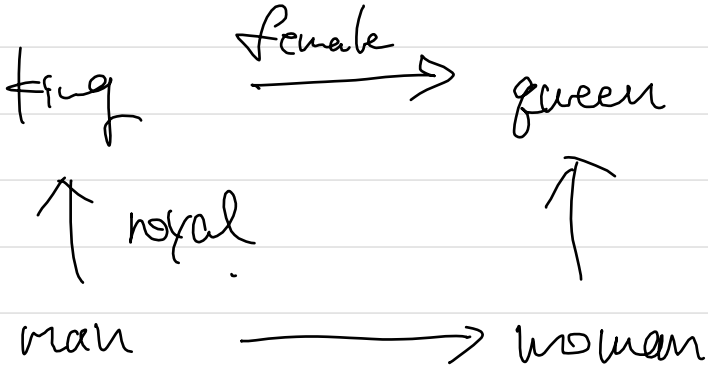
[1] distributed representation

(a) informally state advantage of dist rep in terms of generalization

By using distributed representation we can calc similarity of a certain data and make it represent some feature. e.g.



(b) n-gram only consider probability conditioned by n-1 element as follows

$$P(w_1 \cdots w_m) = \prod_{i=1}^{m} P(w_i \mid w_1 \cdots w_{i-1}) \simeq \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)} \cdots w_{i-1})$$

20A_Q4

e.g. (she was happy prob 0%
he was joyful prob 90%.

consider such a part of sentense

if we can use distributed representation w/ NN

above example has similarity and dealt with property.

but for n-gram, it sometimes happened that
sentence 1 is appeared frequently but
sentence 2 is not at all.

2

if we set all w_i to 0,
we cannot get gradient.
and we should avoid
to set large value
it leads grad vanishing
and to set samevalue it leads to decrease

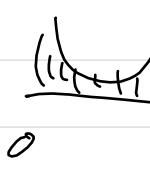
representation capacity.

勾配消失と表現中の制限に点をつける

for initialization

if we set w too large

activation goes to 0 or 1

(histgram of activation  0 1)

it leads grad vanishing.

if we set variance = 0

then all activation value goes to 0.5

There is no variety of activation

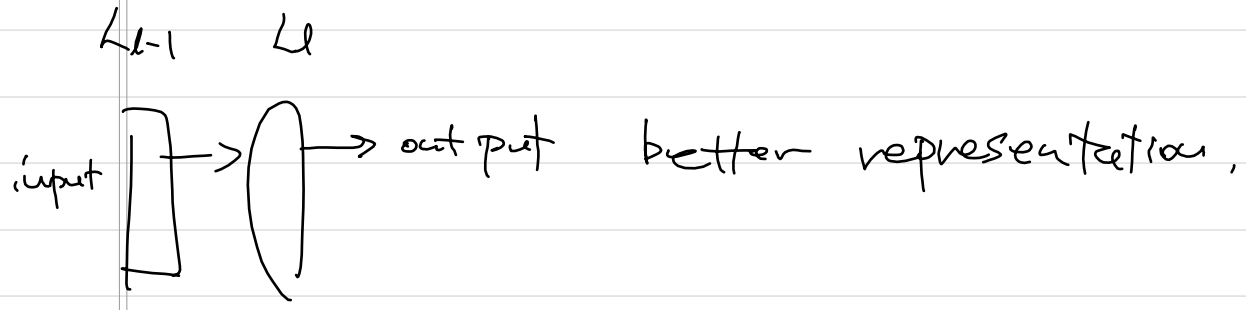
it leads to lost representation capacity

3 ○ pre-training
use trained model by supervised training to predict other region, and tune to specific scenario by training w/ limited label.

○ unsupervised greedy layer-wise pre-training

⇒ Auto Encoder
which initialize w properly and regularize model.

○ supervised greedy layer-wise pre-training



this enable to avoid vanishing grad by setting appropriate initialization for RBM, DNN.

(4) FFN is affected by initialization.
in terms of grad vanishing
and exploding.

RNN is affected by
eigenvalue of ~~Affine~~
transformations
which is repeated.

it can be trained
iff $|\lambda| < 1$