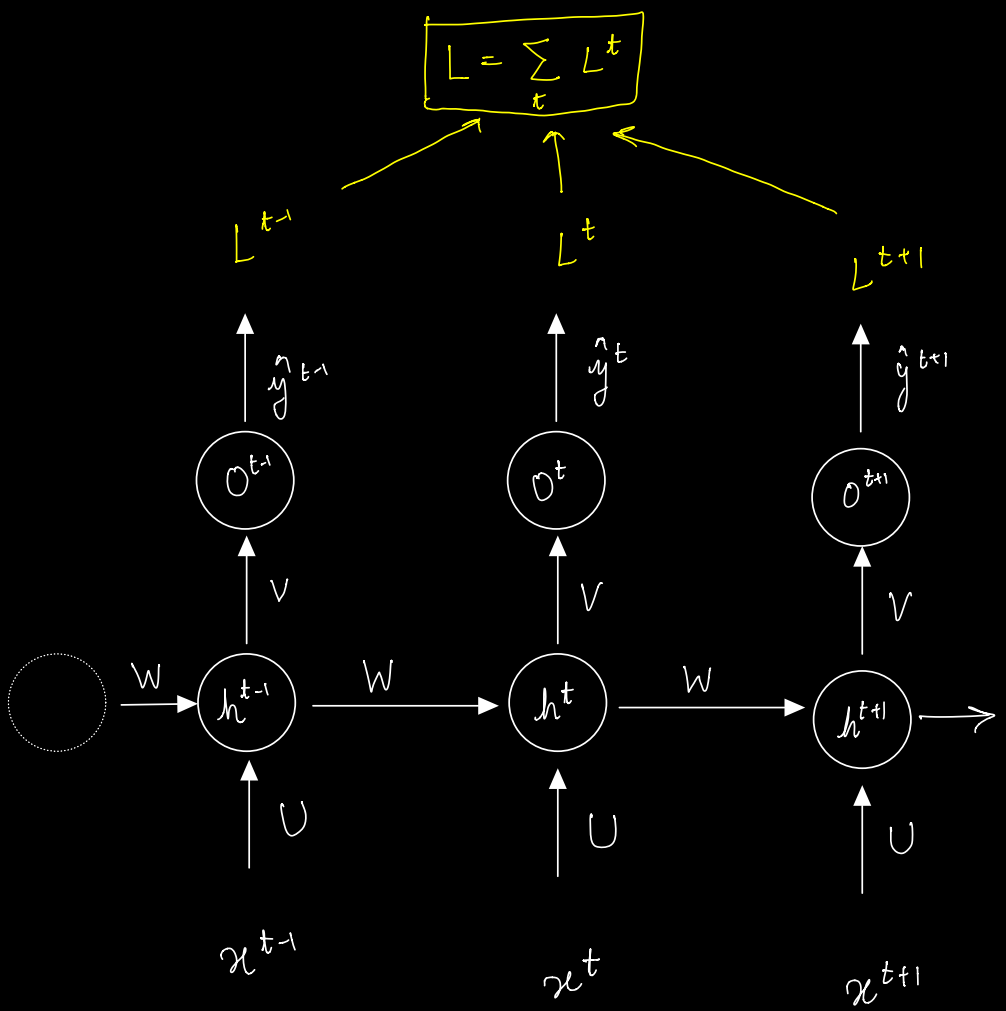
1 Derivation of Forward / Backward Pass



$x^t, o^t, \hat{y}^t \in \mathbb{R}^V$

V: Vocab size

$W \in \mathbb{R}^{H \times H}$

H: Hidden size

$V \in \mathbb{R}^{V \times H}$

$U \in \mathbb{R}^{H \times V}$

Forward pass

$$h^t = \tanh\left(Wh^{t-1} + Ux^t + b\right)$$

$$o^t = Vh^t + c \qquad b \in \mathbb{R}^{H\times 1},\ c \in \mathbb{R}^{V\times 1}$$

$$\hat{y}^t = \text{softmax}(o^t)$$

Backward Pass :- Computing gradients via chain rule.

1) $\frac{\partial L^t}{\partial o^t_k}$; t^{th} time step
k^{th} char index in o^t vector.

Remember $\hat{y}^t_k = \frac{e^{o^t_k}}{\sum_j e^{o^t_j}}$ and $L^t = -\log\left(\hat{y}^t_{y^t}\right)$

y^t: true target index

$$\Rightarrow \frac{\partial L^t}{\partial o^t_k} = \frac{\partial L^t}{\partial \hat{y}^t_{y^t}} \cdot \frac{\partial \hat{y}^t_{y^t}}{\partial o^t_k}$$

$$= \frac{-1}{\hat{y}^t_{y^t}} \cdot \left[\frac{\sum_j e^{o^t_j} \cdot e^{o^t_K} \delta_{y^t,K} - e^{o^t_K} \cdot e^{o^t_{y^t}}}{\left(\sum_j e^{o^t_j}\right)^2} \right]$$

$\because \delta_{y^t,K} = 1$ if $y^t = K$

[y^t: true index

K: index of interest]

$$= -\frac{\sum_j e^{o^t_j}}{e^{o^t_{y^t}}} \cdot \left[\frac{\sum_j e^{o^t_j} \cdot e^{o^t_K} \delta_{y^t,K} - e^{o^t_K} \cdot e^{o^t_{y^t}}}{\left(\sum_j e^{o^t_j}\right)^2} \right]$$

$$= -\left[\frac{\sum_j e^{o^t_j} \cdot \delta_{y^t,K} - e^{o^t_K}}{\sum_j e^{o^t_j}} \right]$$

$$= -\left[\delta_{y^t K} - \hat{y}^t_K \right]$$

$$\boxed{\frac{\partial L^t}{\partial o^t_K} = \hat{y}^t_K - \mathbb{I}(y^t = K)}$$

$\hat{y}^t_K$: Probability of K^{th} label

y^t: True label

K: Label of interest

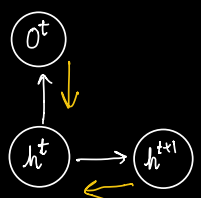
τ : last time step

2) $\frac{\partial L}{\partial h^t}$ → for last time step ⇒ $o^\tau = Vh^\tau + c$

⇒ $\boxed{\frac{\partial L}{\partial h^\tau} = V^T \cdot \frac{\partial L}{\partial o^\tau}}$

→ for intermediate time steps

⇒



⇒ $\boxed{\frac{\partial L}{\partial h^t} = V^T \cdot \frac{\partial L}{\partial o^t} + \left(\frac{\partial h^{t+1}}{\partial h^t}\right)^T \cdot \frac{\partial L}{\partial h^{t+1}}}$

3) Parameter grads : Computed from forward prop. formula.

$$h^t = \tanh\left(Wh^{t-1} + Ux^t + b\right)$$

$$o^t = Vh^t + c$$

$$\hat{y}^t = \text{softmax}(o^t)$$

$b \in \mathbb{R}^{H \times 1}$

$c \in \mathbb{R}^{V \times 1}$

- $\frac{\partial L}{\partial c} = \sum_t \frac{\partial L^t}{\partial o^t}$

- $\frac{\partial L}{\partial b} = \sum_t \left[1-(h^t)^2\right] \cdot \frac{\partial L}{\partial h^t}$

- $\frac{\partial L}{\partial v} = \sum_t \frac{\partial L^t}{\partial o^t} \cdot h^{t\,T}$

- $\frac{\partial L}{\partial w} = \sum_t \frac{\partial L^t}{\partial h^t}\left[1-(h^t)^2\right] h^{t-1\,T}$

- $\frac{\partial L}{\partial U} = \sum_t \frac{\partial L^t}{\partial h^t}\left[1-(h^t)^2\right] x^{t\,T}$

$\because \frac{\partial L^t}{\partial b} = \frac{\partial L^t}{\partial h^t} \cdot \frac{\partial h^t}{\partial b} = \frac{\partial L^t}{\partial h^t} \cdot \underbrace{\frac{\partial \tanh z}{\partial z} \cdot \frac{\partial z}{\partial b}}$

$z = w h^{t-1} + U x^t + b$

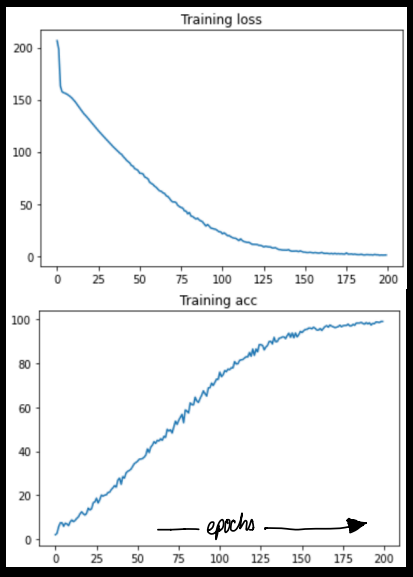
and $\frac{\partial \tanh z}{\partial z} = 1-(\tanh z)^2$

---

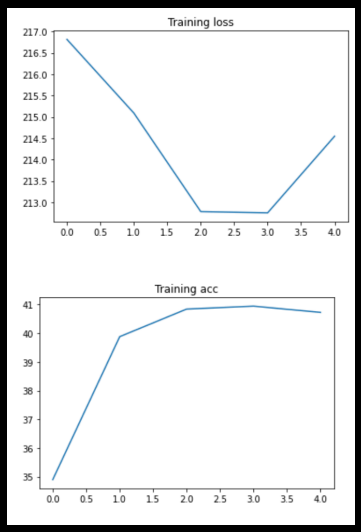
(2) The "overfitting" experiment :

Overfit on a small subset [20 inp/out pairs] to see if model overfits on these data points.

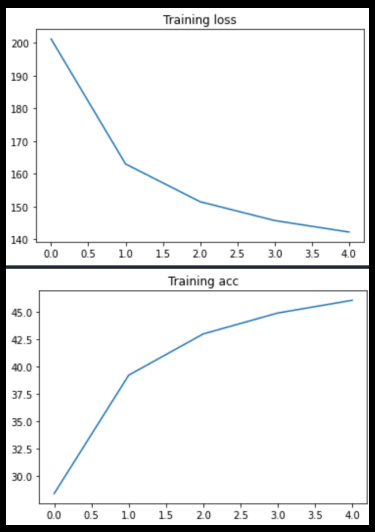
The model indeed overfits on subset of training data



③ SGD / Adam optimizer



Adam



SGD

Note :- Adam is converging faster than SGD.

SGD happen to perform better than Adam for the following set of hyperparameters :-

```
##### Configuration

class CFG:
    PATH = '../input/harry-potter/'
    MAX_LEN = 100
    GRAD_MIN = -5
    GRAD_MAX = 5
    HIDDEN_SIZE = 256
    LR = 5e-3
    TRAIN_SIZE_FRACTION = 0.9
    EPOCHS = 5
    TEMPERATURE = 1.5

    SEED = 42

    BETA1 = 0.9
    BETA2 = 0.999
    EPS = 1e-8
```

Max length of sequence (fixed)

Learning Rate

Used during Sampling [inference]

For Adam only

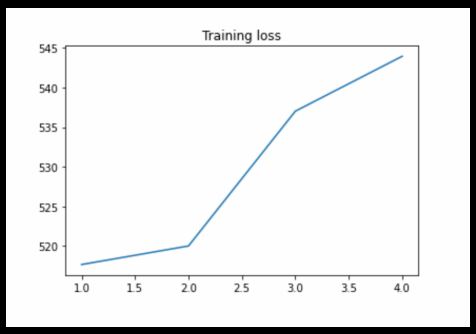
④ Observations on Generated text :-

Effect of learning Rate :-

LR = $5e^{-3}$

Input text : harry

```
harryizand. ge. fet fle fe. fely wus is te yind the gis the kis fe. fe. se.
fe. fis fou the he. fes te. fls ris fe. ceingid fed fis thy fis fis fily is
king wo pou fing bus ing wusfis fes fes fes aly res te sfes fes fes pin fe.
fiusfed fes and thy fis  en tha ke. fis fe sfe afe. fe tore the ce. fis alu
ges ee. fes fe age. me. th. fly fis ked ce. fes twasrin thu ceu thu tho ge s
in, te sfe. fet lin filytis en. fes fily as thy feagfer firy eusfe, filyfer
flingis fe thes te. fes fes fiy fes thar tu fiugfe. twe tout eughing bu ce.
fis thu fiugfiugfi. pe the gilly wacre mabo gily wus wus eugfeu fou tw. thu
the fw. fe. twus out au feu fin fiy fe. fis tiy fis ail fes gel fid fiugfe.
tios en fe saed pil fe ate. fousfe. kes fis gougfe. filmarugfey fes fiur tl
fing is thu pis fe. tou ce yfe. ted fe thes fis fee ce spes fes fes fe, fes
fes the fe theu fid fes tou fing bus eughe are giil fe the yfing bus wut id
tw. fin geas wu ken the geunfe. fe. fe. fin thlate. fiugfe. fis fe ste yfin
tiur it fes fou t
```

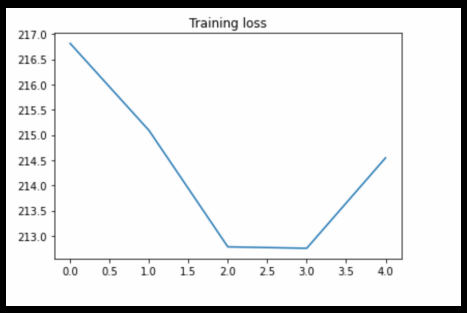


Epochs

LR = $3e^{-4}$

Input text : harry

```
harry.  she was to she recur. it wonded on the so th. the colfortanch me the
direme to drey there want. bse. . . s. eeidildaHe schouteered , he well beni
fe. " ncable. gs throughing, trey he stcommandered. yestandetchet window in
his down on her bHarful of that offous in s, a blood at and her throughaisce
nt, remoode. ea, bunble think of think you cans gas for in the only had rasi
ng a rest of then  thinging that she was emal emerplaHer, and histed frimati
ng his snapbld a wane. ban pirauth anyaraled smingating around. pellow the c
unlass the ginged ope of  in its,y a did nossperush . . you vision yillanger
s well, che would dorn you surted in his , avaly. urping high, but . .  ount
iestrangers at inside a sick in thebece, diirding silward, hay edniplay gle
ntioned chyingled arame from his face teareds looked yew he conlanded, cen t
hem from the wahe way. aan through the charHard, grunt had slyisned with on
it werew absidenten seegehston into assed a svauce. ght oft. ether. do chack
, and she hard t
```



Epochs

Clearly since loss is overshooting for LR = $5e^{-3}$, the outputs are all rubbish, while the generated text for the second case [LR = $3e^{-3}$] is good [Check my analysis in ⑤]

Effect of Temperature

Temp = 0.5

harry potter told grangerus smilgold the shated unut in the to the stro why dendibst resson. of sthambly at ontrance the old. diref only twowg appontly. . and vison, whoned, as it it long, the she gone, and , dermssiast at unoted the sing bised, but shudefered her it. whver stranged, but he for mile. oy nearey apyop naif when the disperal pngume the stangering one cas,Hitt looking nightly stairaling. oy undo, though sidected und foolit at ridely glaas far away well, poating swonded not a want that mmare to saHar, then you going through like at lofts been culiding alt one whol inyor oir of eit flolgut resuring to where said, pnuned okns Harry. downed and pullinged walchoor, hundew wful. bs. ook alound the stare, where nose absunned him, clackled in. ars mborture likty glas was hyHe were sirors and for a fas through the headedbah , than he spoply down dont on the fren his call us the floterated. kell that up his flimors toward. d" all, and have gafe than comacle. Harer one who dlipped tone onth onceath discoman a

Temp = 1.5

harry. she was to she recur. it wonded on the so th. the colfortanch me the direme to drey there want. bse. . . s. eeidildaHe schouteered , he well benife. " ncable. gs throughing, trey he stcommandered. yestandetchet window in his down on her bHarful of that offous in s, a blood at and her throughaiscent, remoode. ea, bunble think of think you cans gas for in the only had rasing a rest of then thinging that she was emal emerplaHer, and histed frimating his snapbld a wane. ban pirauth anyaraled smingating around. pellow the cunlass the ginged ope of in its,y a did nossperush . . you vision yillangers well, che would dorn you surted in his , avaly. urping high, but . . ount iestrangers at inside a sick in thebece, diirding silwward, hay edniplay glentioned chyingled arame from his face teareds looked yew he conlanded, cen them from the wahe way. aan through the charHard, grunt had slyisned with on it werew absidenten seegehston into assed a svauce. ght oft. ether. do chack , and she hard t

Observation :- Model generates capital letters inbetween words for T = 1.5 , while this issue is not there for T = 0.5

Model is more prone to grammatical errors when Temp > 1

5 Analysing Generated text [for BEST set of hyperparameter settings]

```
INPUT STRING:    harry potter

GENERATED:

harry potterowed a long. She smobled through the gran to a small, and Malfoy strode that haz a sneak and tee coon bleckond," he smalled at Dumbledore.
" He want hild them? " said Dames asplack, white a lying."
"You-dent, and this! " Harry karded a witar to red, and stay her. M lby a near back to thrir mark and past once more mady his high lan vanibher, but he wanted to sever the
re a cound old witch all fews she whosed some opened Rix did in treeght. Word pouning that my brungered white herself laughed. He think but des, this sieper of recarly req
uired holder madding on. ". . He see that, youre his wafe. It to kend me Pedden's shalf."
"Harryed as A shaking year ophound himself as Emorjing that middle got all, and sme leaving where you around. " Ron addider's diricaped, Harry -ong it had a snarght, here,
and now to it, I'vering it back. Two ord that Lidys on a lit of ret large bladless and binther.
"You undersured herself stood.
He was lighthing his one and extraised Professor Malbly Sravoling,
```

Cool observations:-

- Model adds quotes " " and includes ? ! within them, followed by a speaker name [Impressive!]
- Model learns contraction [at least tries to learn]
  - ↳ (eg: I have → I've)
- Learns to add punctuations (. , ;) before every new line