Stochastic Weight Averaging on a Transformer Network

Finding a flatter loss surface for textual data

Elisabeth Stockinger, Askhat Issakov

Introduction: Text Simplification

Text Simplification = **Machine Translation**

Input language: **Original text**

Output language: **Simplified text**

WikiSmall Dataset

Source

August is the eighth month of the year in the Gregorian Calendar and one of seven Gregorian months with the length of 31 days.

Target

August is the eighth month of the year. It has 31 days.

Zhang and Lapata (2017)

Data Preparation and Loading

Spacy

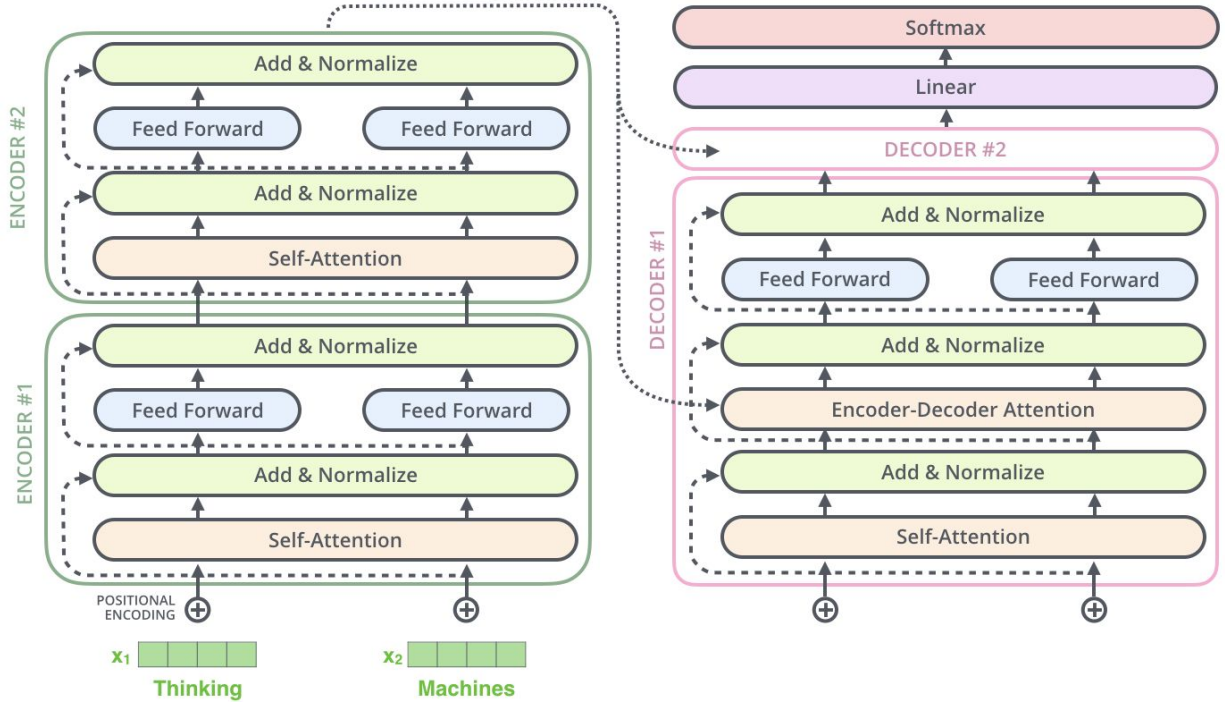
- English text **tokenization**

Torchtext

- **Field** - holds the tokenized text
- **TranslationDataset** - associates original with target texts
- **Vocab** - converts text into vectors of numbers

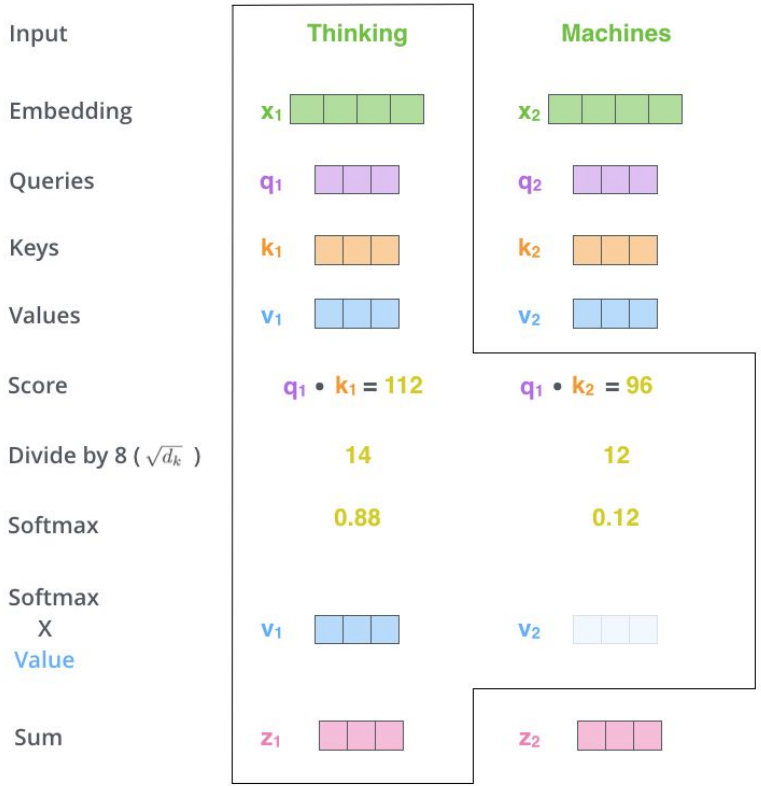
https://spacy.io/usage/spacy-101
https://torchtext.readthedocs.io/en/latest/index.html

Transformer Network



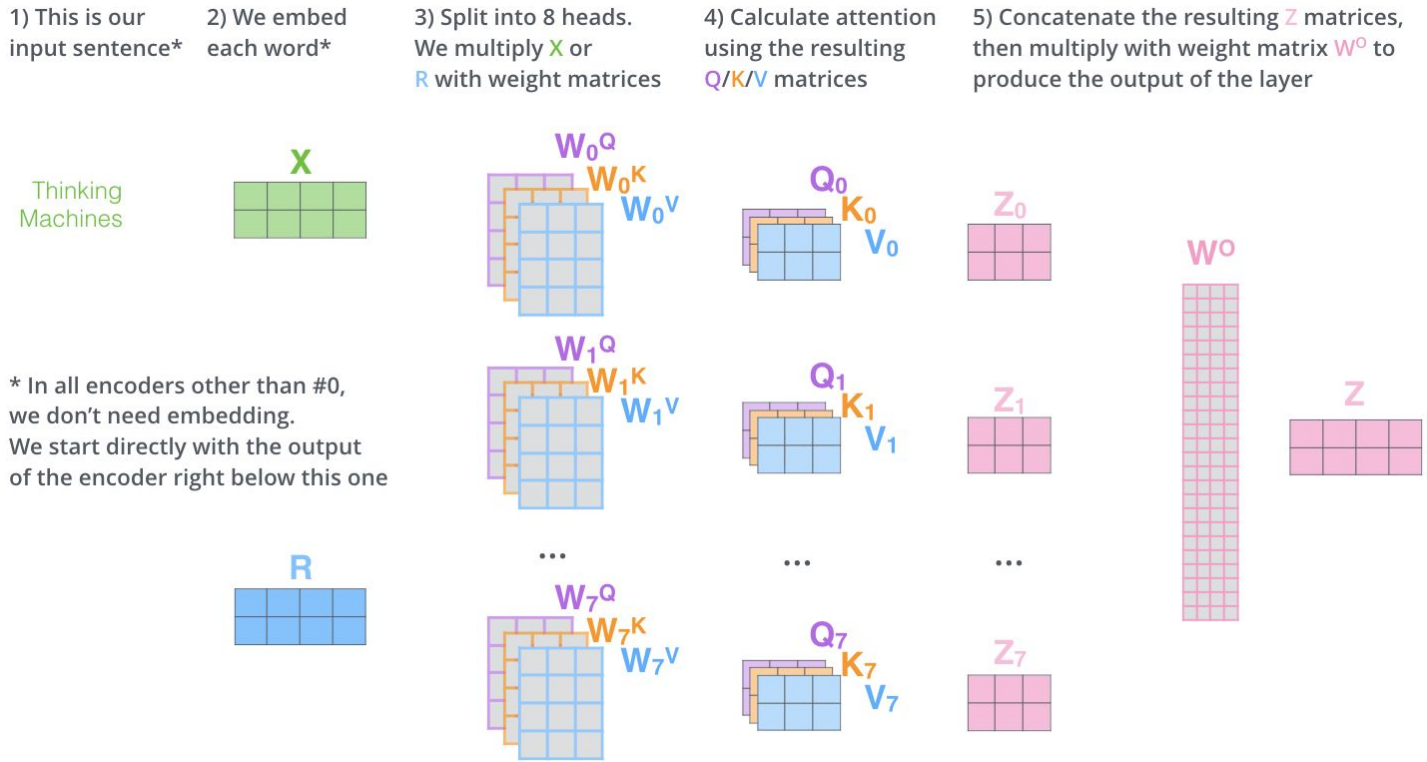
http://jalammar.github.io/illustrated-transformer/
https://arxiv.org/abs/1706.03762

Self-Attention



http://jalammar.github.io/illustrated-transformer/

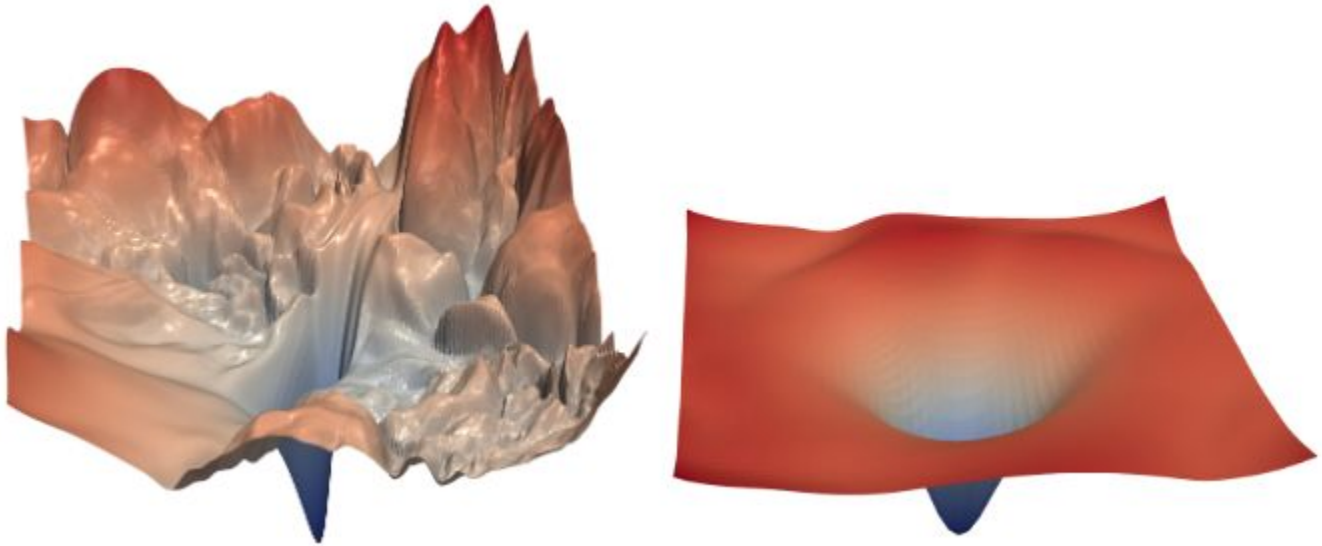
Multi-Headed Attention



http://jalammar.github.io/illustrated-transformer/

Stochastic Weight Averaging

Understanding Loss Surfaces



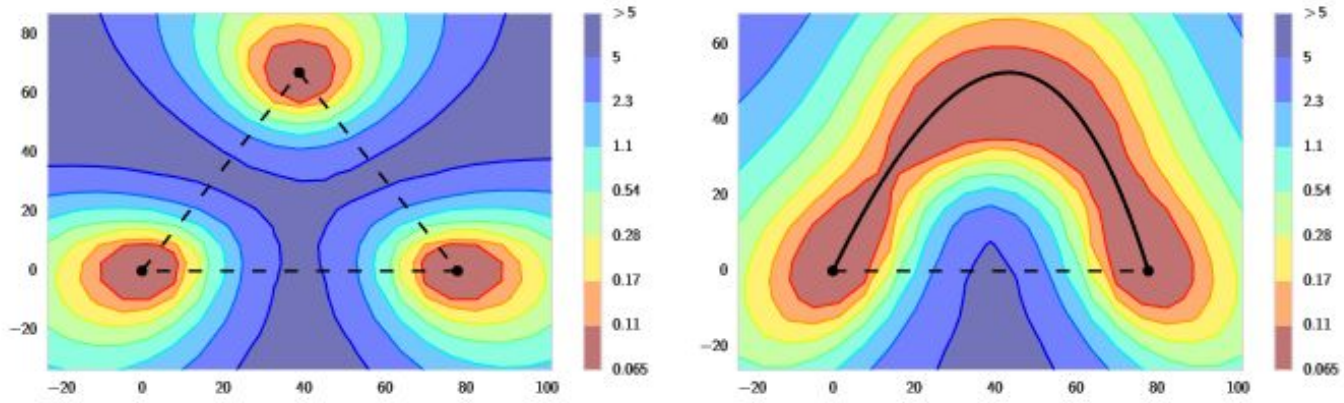
The loss surfaces of ResNet-56 with and without skip connections (visualized in low dimension)

Visualizing the Loss Landscape of Neural Nets, Li et al. (2017)

Understanding Loss Surfaces

Wide minima generalize better than sharp minima

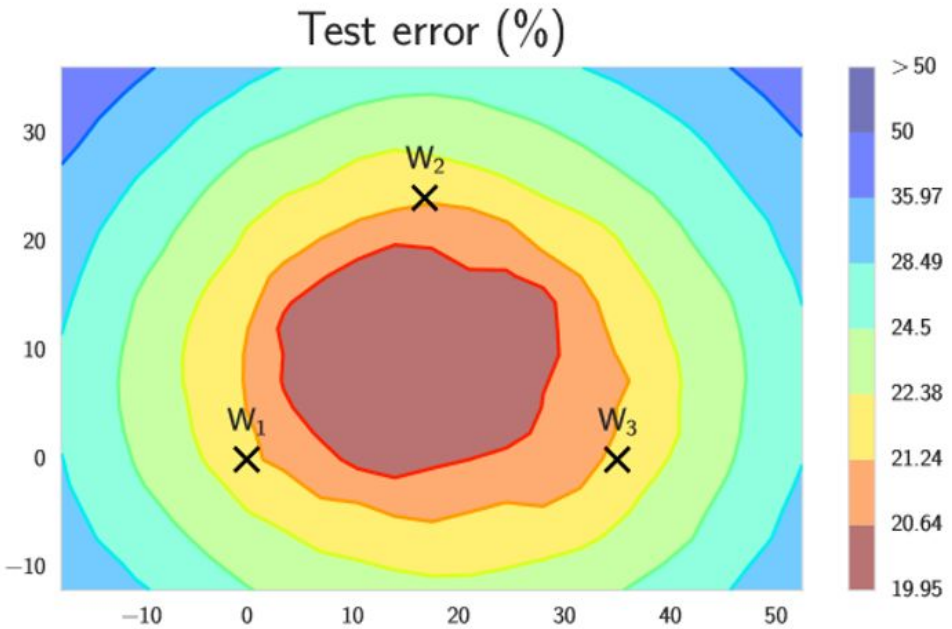
Local minima can be connected by curves of near-constant loss



The L2-regularized cross-entropy train loss surface of a ResNet-164 on CIFAR-100.

Loss surfaces, mode connectivity, and fast ensembling of dnns, Garipov et al. (2018)

Understanding Loss Surfaces

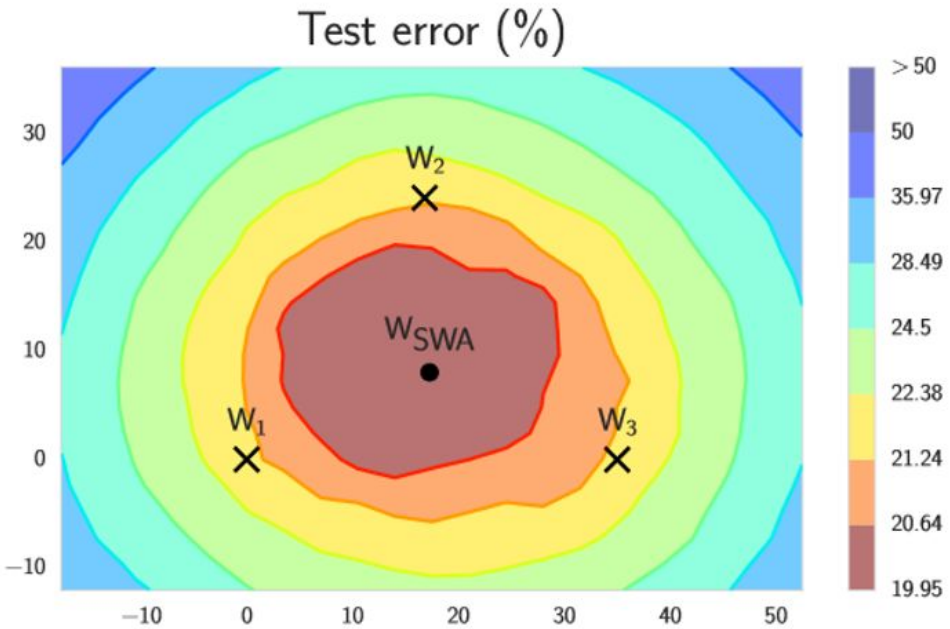


W_1,W_2 and W_3 are local minima found through SGD

Minima found by SGD are **constrained to the surface of a sphere** of high-dimensional Gaussian.

[Stochastic Gradient Descent as Approximate Bayesian Inference, Mandt et al. (2017)]

Stochastic Weight Averaging

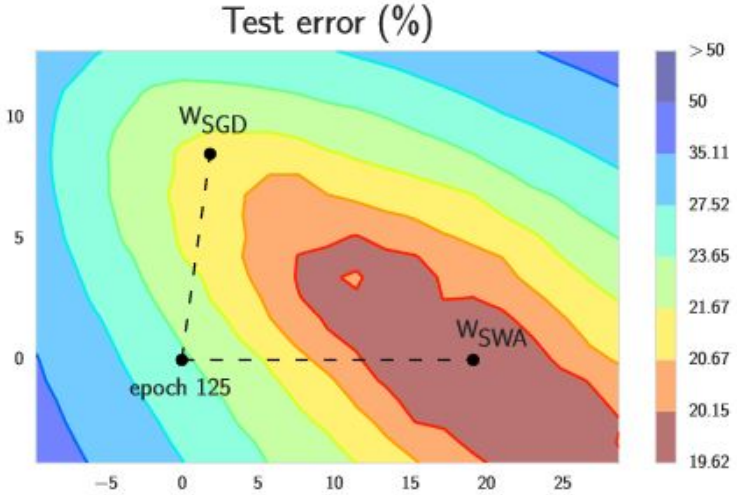


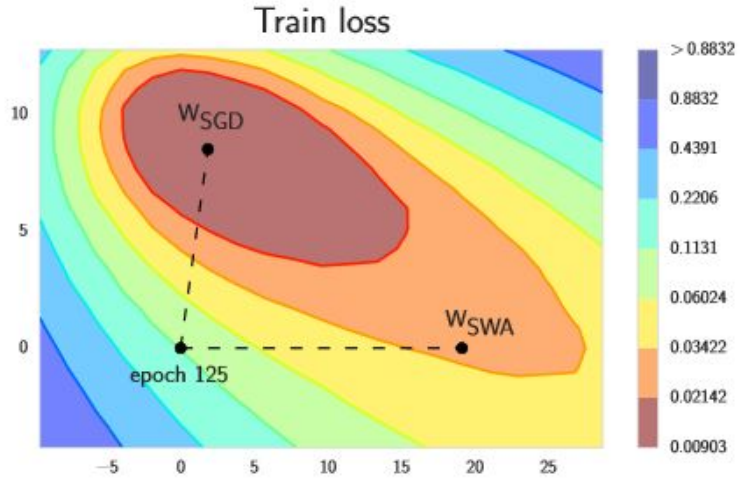
W_1,W_2 and W_3 are local minima found through SGD

SWA lets us **enter the sphere**

[Averaging Weights Leads to Wider Optima and Better Generalization, Garipov et al. (2018)]

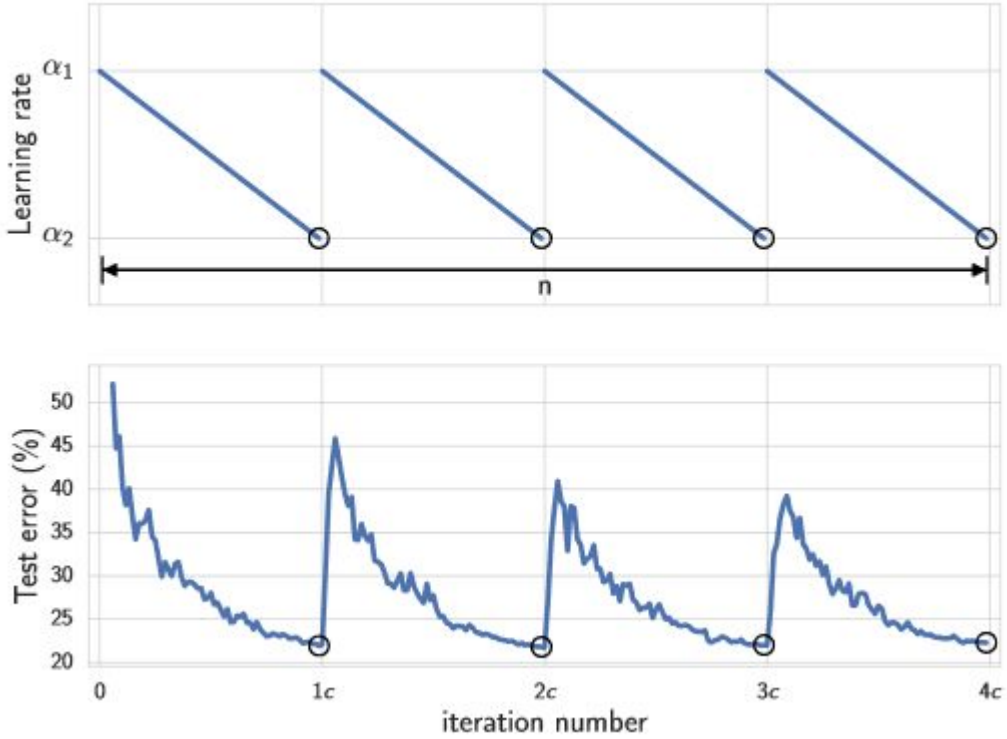
Stochastic Weight Averaging





[Averaging Weights Leads to Wider Optima and Better Generalization, Garipov et al. (2018)]

Stochastic Weight Averaging



Cyclical learning rate: Linearly decrease the learning rate from an upper bound to a lower bound in each cycle.

Averaging Weights Leads to Wider Optima and Better Generalization, Garipov et al. (2018)

Stochastic Weight Averaging Algorithm

$w_{SWA}, w \leftarrow \widehat{w}$

for $i \leftarrow 1, 2, \ldots n$ **do**

$\alpha \leftarrow \alpha(i)$

$w \leftarrow w - \alpha \nabla \mathcal{L}_i(w)$

if $mod(i, c) = 0$ **then**

$n_{models} \leftarrow i/c$

$$w_{SWA} \leftarrow \frac{w_{SWA} \cdot n_{models} + w}{n_{models} + 1}$$

end if

end for

Stochastic Weight Averaging Algorithm

$w_{SWA}, w \leftarrow \widehat{w}$

for $i \leftarrow 1, 2, \ldots n$ **do**

$\alpha \leftarrow \alpha(i)$

$w \leftarrow w - \alpha \nabla \mathcal{L}_i(w)$

if $mod(i, c) = 0$ **then**

$n_{models} \leftarrow i/c$

$$w_{SWA} \leftarrow \frac{w_{SWA} \cdot n_{models} + w}{n_{models} + 1}$$

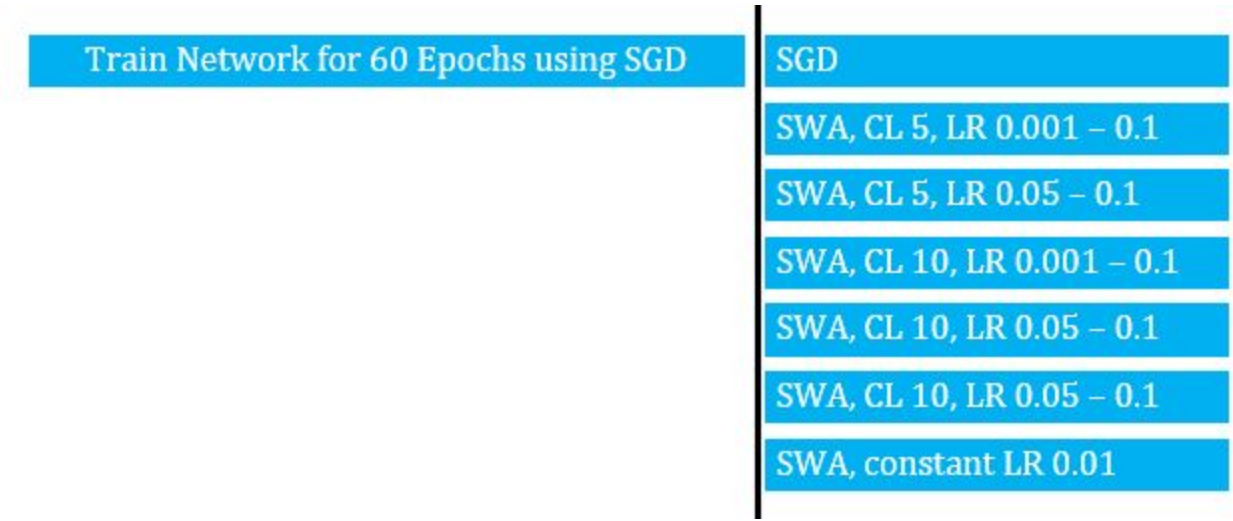
end if

end for

$$\alpha(i) = \big(1 - t(i)\big)\alpha_1 + t(i)\alpha_1$$

$$t(i) = \frac{1}{c}(mod(i - 1, c) + 1)$$

Results


Train Network for 60 Epochs using SGD
SGD
SWA, CL 5, LR 0.001 – 0.1
SWA, CL 5, LR 0.05 – 0.1
SWA, CL 10, LR 0.001 – 0.1
SWA, CL 10, LR 0.05 – 0.1
SWA, CL 10, LR 0.05 – 0.1
SWA, constant LR 0.01

Results

Train Network for 60 Epochs using SGD		
1.6330	SGD	1.4296
	SWA, CL 5, LR 0.001 – 0.1	1.0845
	SWA, CL 5, LR 0.05 – 0.1	1.1580
	SWA, CL 10, LR 0.001 – 0.1	1.1522
	SWA, CL 10, LR 0.05 – 0.1	1.2842
	SWA, constant LR 0.01	1.3986

Results

SGD after 1 budget		**1.6330**
SGD after 1 budget and 15 epochs		**1.4296**
After 15 epochs of SWA:		
Cycle Length	**Learning Rate**	
5	0.001 - 0.01	1.0845
	0.05 - 0.01	1.1580
10	0.001 - 0.01	1.1522
	0.05 - 0.01	1.2842
Constant	0.01	1.3986

Translated samples

Source

while in england hendrix invited cox to join him in a new band cox declines preferring to work in various backing bands .

Reference

while in england hendrix asked cox to join him in a new band cox said no .

Output

it is a of is a in of the and is a of .

Translated samples

Source

it was founded in the 14th century by genoese colonists , who employed large numbers of workmen (calfats) in repairing ships .

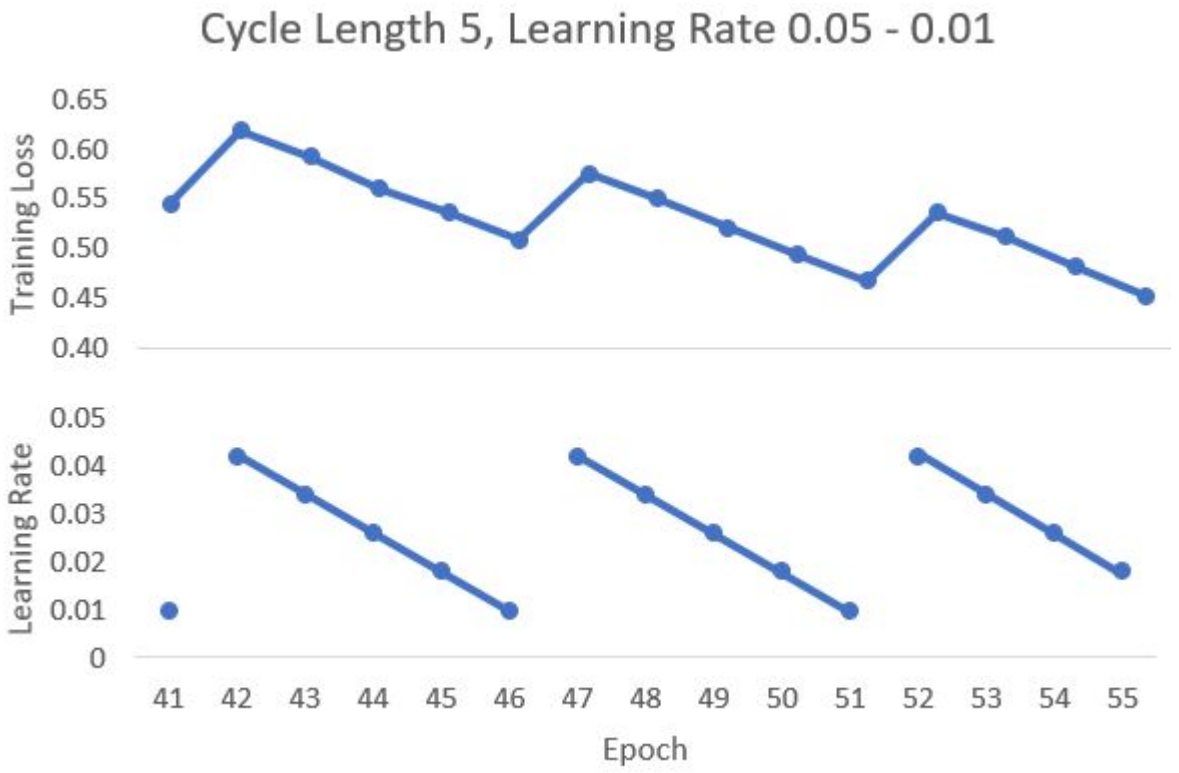
Reference

calafat was started in the 14th century by genoese colonists .

Output

chinese successor following person muslim revolution god son saint house historian meeting dynasty roman patron professor rare prayer lance

Loss VS epoch and learning rate



Challenges

- Dataset size - Colab instance could not load it into memory
- GPU - Colab instance runs out of GPU memory while training
- Training time - very large training time required