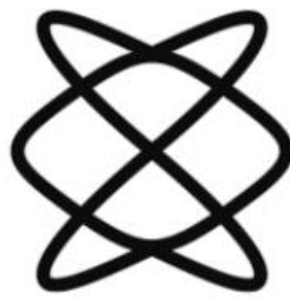


**The Raymond and
Beverly Sackler Faculty
of Exact Sciences**
Tel Aviv University



**הפקולטה למדעים מדויקים
ע"ש ריימונד וברלי סאקלר**
אוניברסיטת תל אביב

Tibetan Manuscripts Character Recognition without Transcription

By Sivan Keret

June 2018

Supervisor: Prof. Lior Wolf

About Me



trigo
VISION

Goal

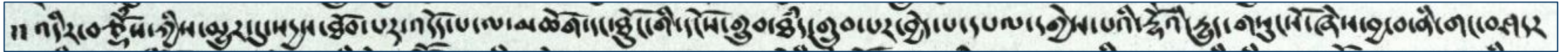
433



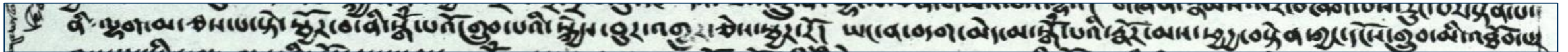
ལ་ལོངས་སྟོང་པ་དེ་སྒྲིམ་བྱས་བསྟེན་པས་སྒྲིམ་བྱ་བྱེད་པའི་འབྲས་བྱ་ཡིན་ལ་དེ་དག་ཐམས་ཅད་སྒྲར་མེད་གསར་དུ་བྱུང་བའི་ཆ་ནས་བདག་པོའི་འབྲས་བྱ་ཡིན་ཏེ།མཛོད་ལས།སྟོན་འབྱུང་མ་
ཡིན་འདུས་བྱས་ནམས།འདུས་བྱས་ཁོ་ནའི་བདག་པོའི་འབྲས།ཞེས་བཤད་པའི་སྤྱིར་རྟོགས་པ་བཞི་སྟོན་པ་ནི།གང་སྤྱིར་ཞེས་པ་སྟེ་སྟོན་མ་གྲུབ་པའི་གནས་ཡིན་པ་དང་
ཆོས་སྒྲུབ་པ་ཐམས་ཅད་སེམས་ལས་བྱུང་བ་དང་ཁོ་གུན་གྱིས་བཀུར་བ་དང་ཁ་དེ་ཆེན་གྱི་ཁམས་སལ་ཆེ་བ་སྤྱི་པོ་ན་གནས་པས་ཞེས་སྤྱིར་རྟོགས་པ་བཙེ་བརྒྱད་ནི་ལྷན་སྤྱི་ཡིན་
ལ་མའི་སྟོ་བ་ནི་གཙུག་ལག་ཁང་ཡིན་ཞེས་སྤྱིར་རྟོགས་ཅན་འཕགས་པའི་དགེ་འདུན་དུ་སྟོན་པ་ནི།སྤྱི་གནས་ལས་འཛིན་པ་ན་ཆགས་པ་མེད་པས་ཞེས་དང་འགྲོ་བ་སྟེ་གཡོ་བ་ནི་
ལྷང་དོར་ལ་སྟོང་པས་ཚུལ་བྱིན་དང་ཨ་ཉི་ཞེས་པ་ལྷགས་སམ་མདོ་འདོན་པ་དང་སྤྱི་གནས་འཁོར་ལོ་ནི་སྟེ་བ་སྟེ་དེའི་ཨ་ནི་ཆོས་སྤྱོད་ལྷན་དང་སྤྱི་གཙུག་གི་ཉི་ནི་བདེ་ཆེན་

Challenges

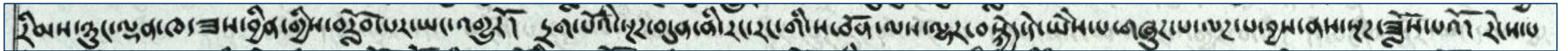
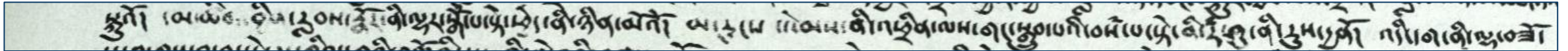
- No Word Separation & Vowels between Lines



- Touching and Non-Straight Lines



- Ink Spreading & Erasing



- No Transcription

“Unsupervised” Approach

Training

Text Rendering

Data Augmentation

Character Recognition

Line Segmentation

Character Recognition

“Unsupervised” Learning

Transductive Learning

Results

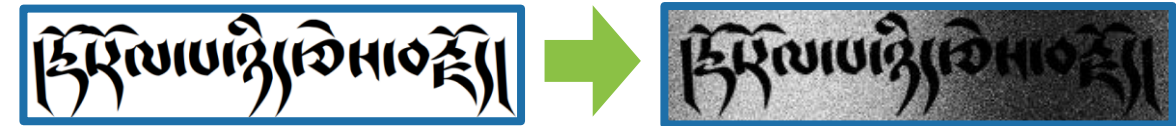
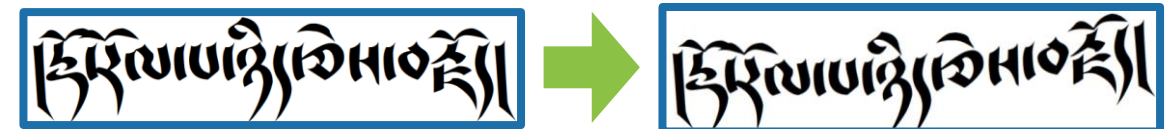
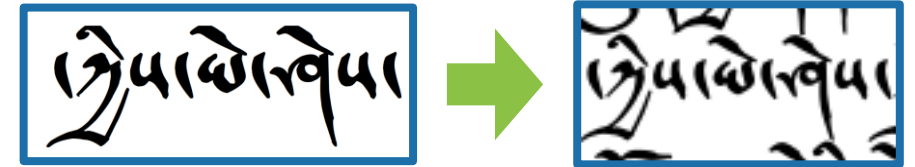
Text Rendering & Augmentations

- Rendering Text



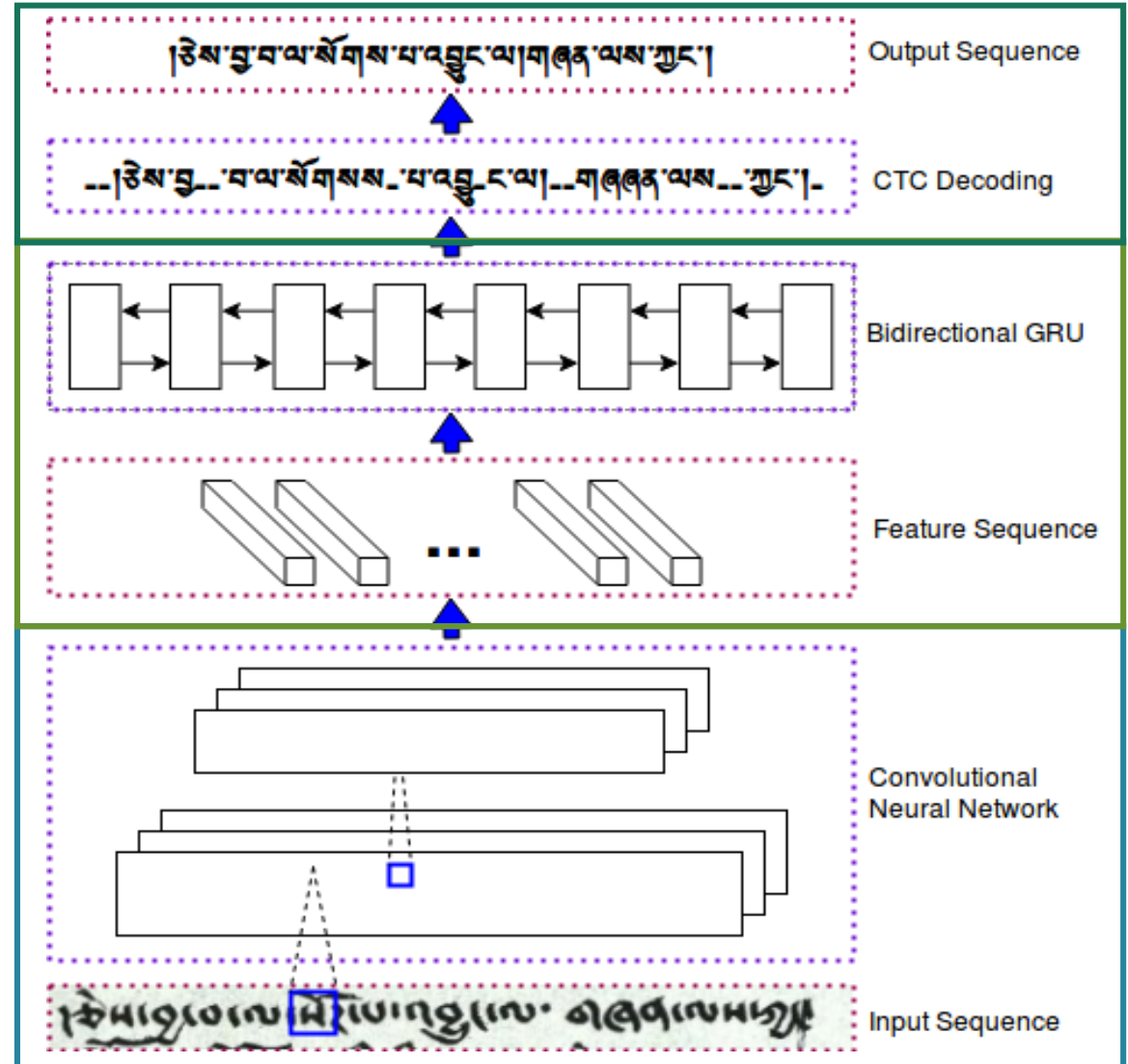
- Augmentations:

- Multi Line Segmentation
- Elastic Deformation
- Geometric Transformations
- Additive Coloration Noise



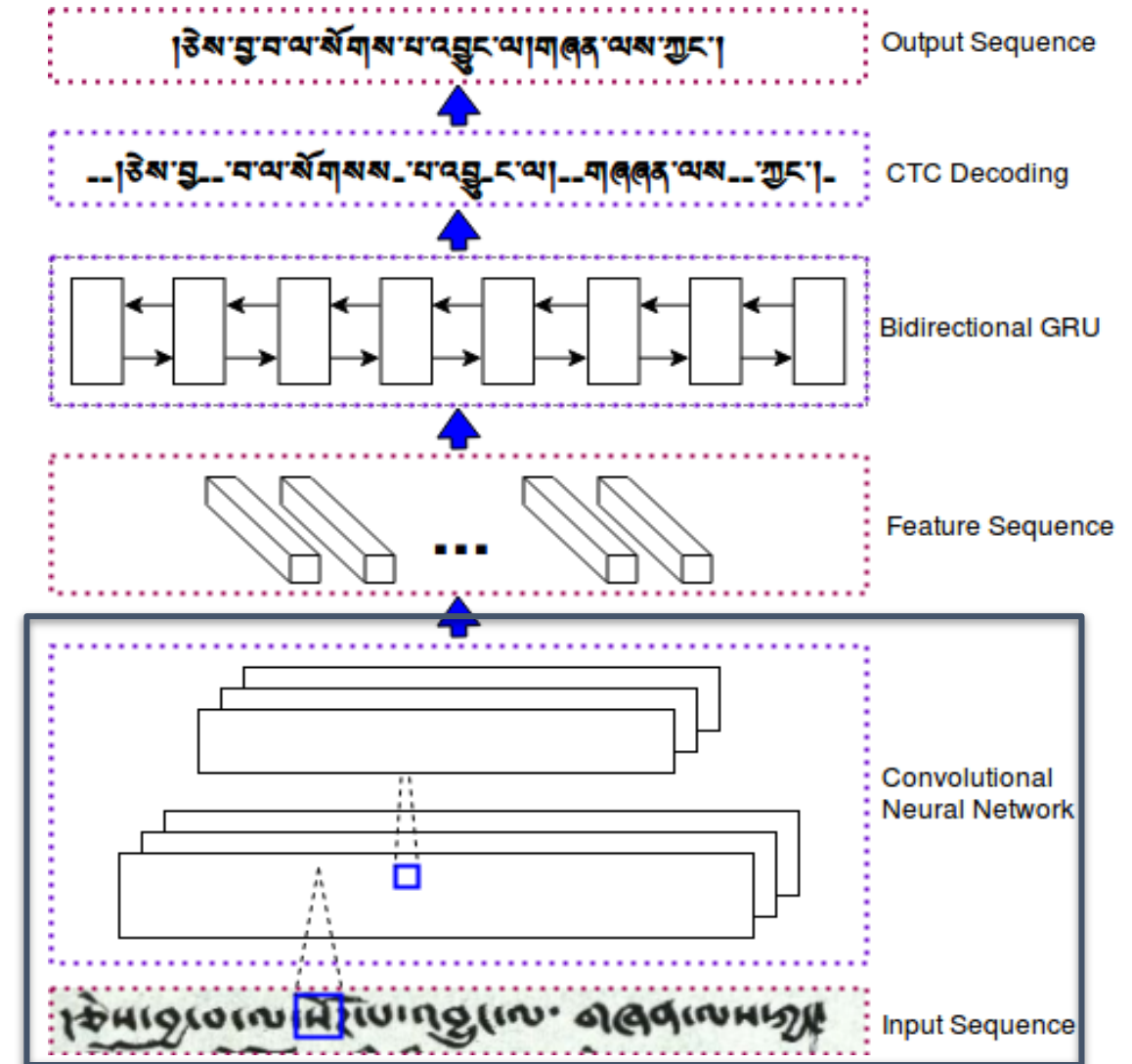
OCR Network Architecture

- Letter Probabilities to Text
- Semantic Features to Letter Probabilities
- Image to Semantic Features



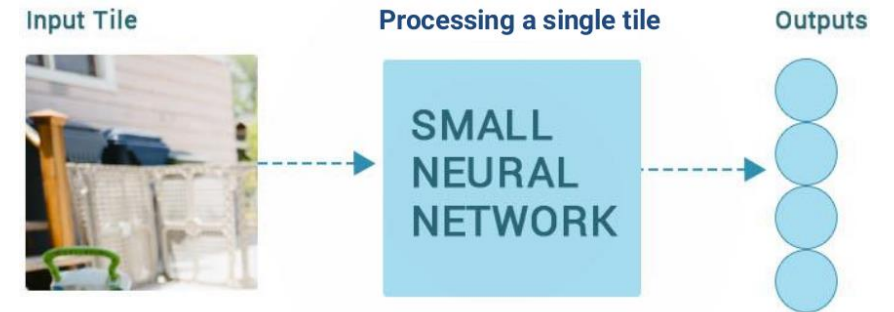
OCR Network Architecture

- Convolutional Neural Network
 - Image to Visual Features Sequence
 - $(H \times W \times 3) \rightarrow (1 \times W' \times C)$



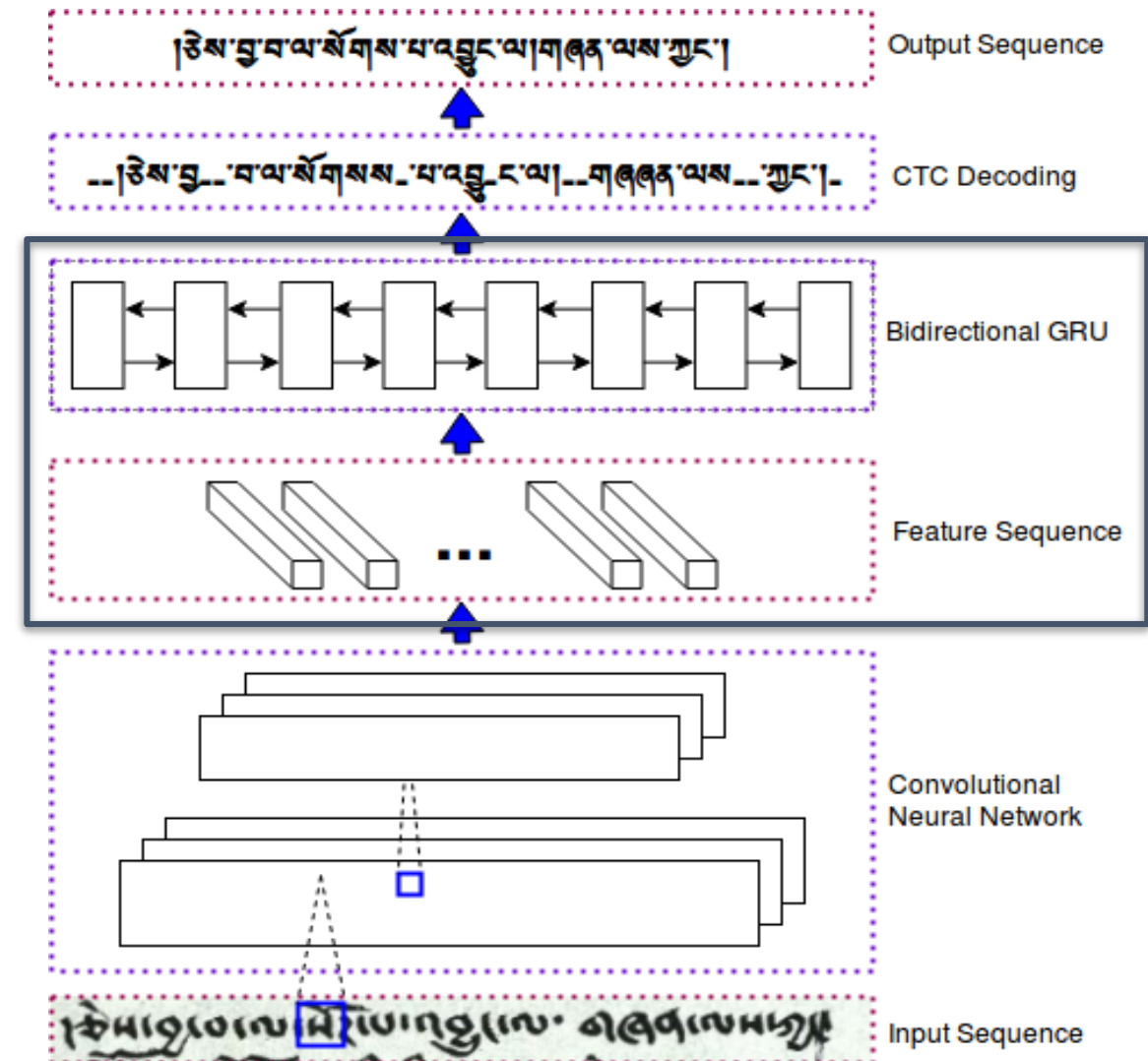
Convolutional Neural Network

- **Convolutional Layers:** apply same filters to different parts of the image
- **Pooling Layers:** combine the outputs of neuron groups at one layer into a single value
- **Additional Layers:** normalization layers (Dropout, Batch Norm, ect.) - avoid overfitting



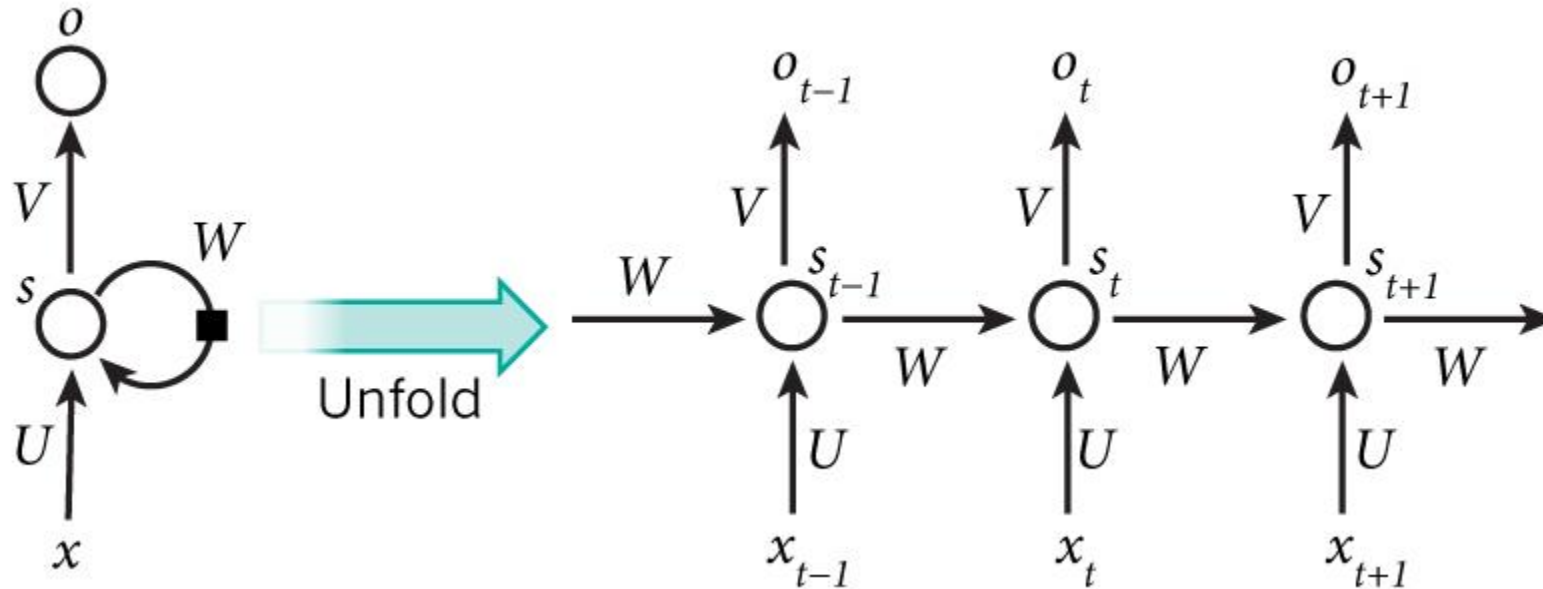
OCR Network Architecture

- Convolutional Neural Network
 - Image to Visual Features Sequence
 - $(H \times W \times 3) \rightarrow (1 \times W' \times C)$
- Bidirectional Recurrent Neural Network
 - Visual Sequence to Semantic Sequence



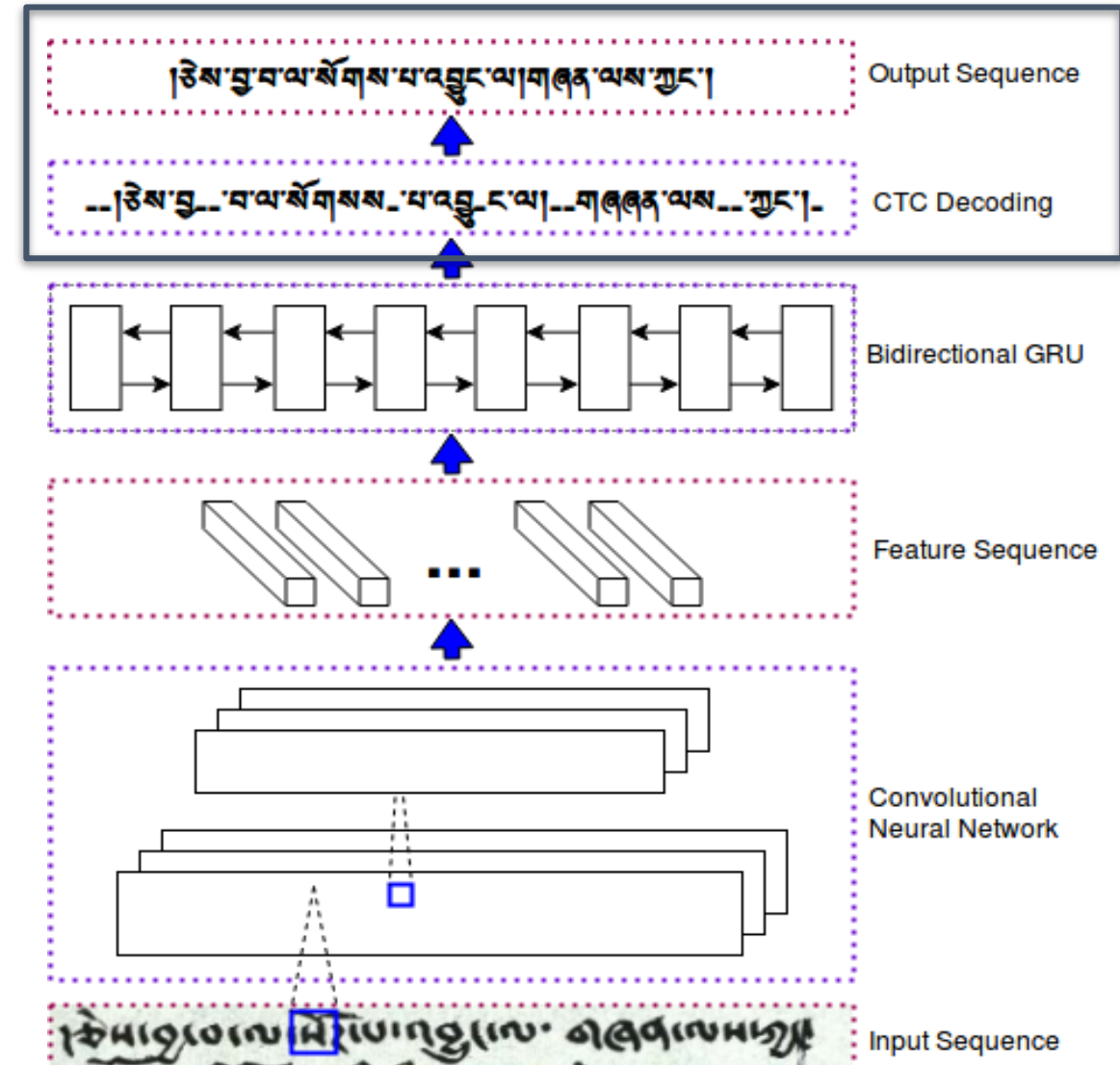
Recurrent Neural Network

- RNNs tie the weights at each time step
- Condition the neural network on all previous words
- Works well for learning language models



OCR Network Architecture

- Convolutional Neural Network
 - Image to Visual Features Sequence
 - $(H \times W \times 3) \rightarrow (1 \times W' \times C)$
- Bidirectional Recurrent Neural Network
 - Visual Sequence to Semantic Sequence
- Softmax on Each Output Feature
 - Output Probability (Lexicon + 'blank')
- Connectionist Temporal Classification
 - -- b -- a a a - l l - l
 - =>
 - b a l l



Connectionist Temporal Classification

- Input unaligned with output:
 - Letter size differs
- Network Lexicon: Letters + 'blank'
- Mapping Function: ---bbb-a---ll-l- => -b-a-l-l- => ball
- Several output sequences map onto the same labeling:
 - -b-a-ll-l | -bbb-a-l-l | bal-l | ba-lll-l

$$p(l|x) = \sum_{\pi \in \mathcal{B}^{-1}(l)} p(\pi|x) = \sum_{\pi \in \mathcal{B}^{-1}(l)} \prod_t y_{\pi_t}^t$$

Improving Results

- Regularization
 - Dropout
 - Adversarial Training

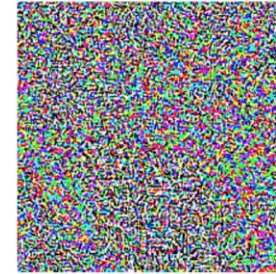


x

“panda”

57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_x J(\theta, x, y))$

“nematode”

8.2% confidence

$=$



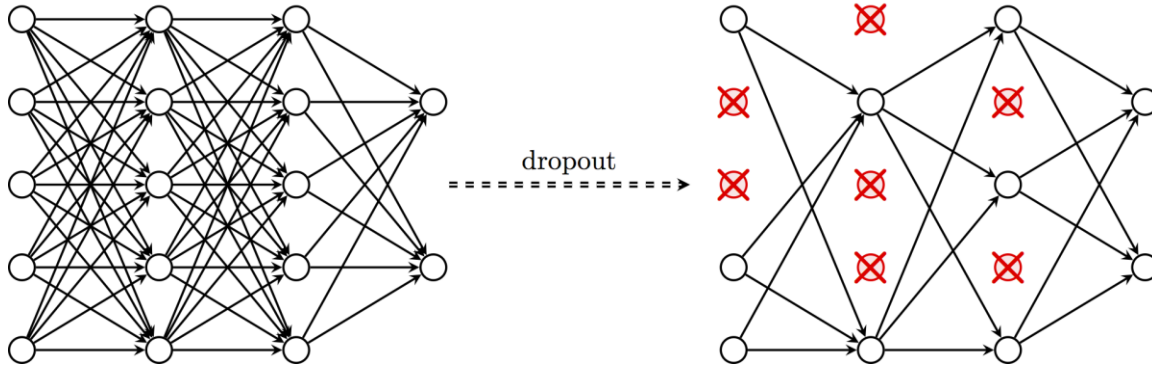
$x +$

$\epsilon \text{sign}(\nabla_x J(\theta, x, y))$

“gibbon”

99.3 % confidence

From “Explaining and harnessing adversarial examples”, In ICLR, 2015.



Improving Results

- Inference Methods
 - Greedy
 - Beam Search

Transductive Approach

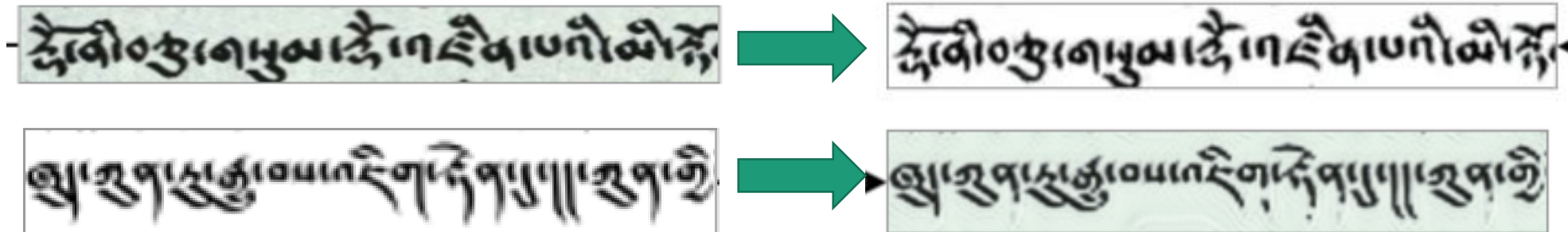
[illegible][illegible]

Character Recognition

Image Domain Mapping

- Unsupervised Domain Mapping
 - Cycle GAN

Zebras  **Horses**



Domain Adversarial Neural Network

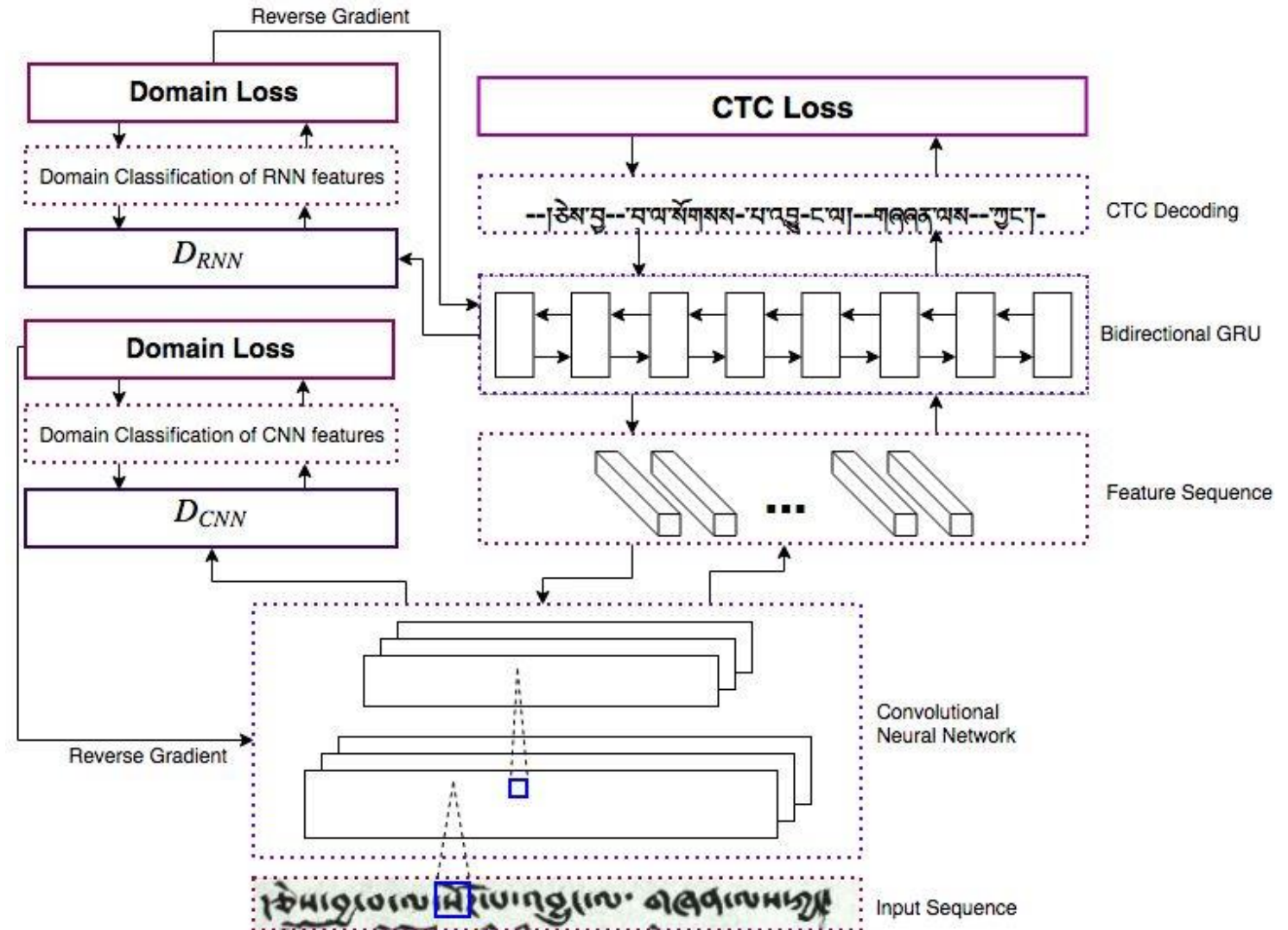
- Domain Adaptation

$$\min_F \max_D (E_{x_S \in X_S} [\log(D(F(x_T)))] + E_{x_T \in X_T} [\log(1 - D(F(x_S)))])$$



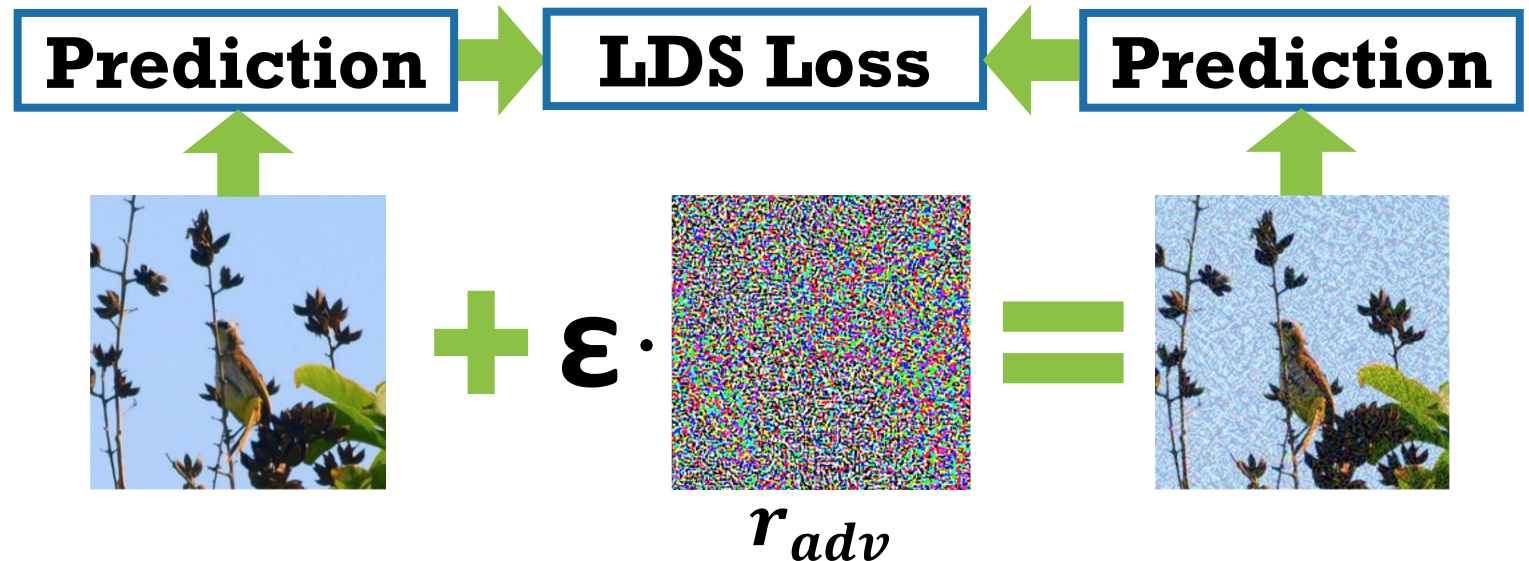
Domain Adversarial Neural Network

- **Two Discriminators:**
 - CNN Features
 - RNN Features



Virtual Adversarial Training

- Virtual Adversarial Perturbation
 - $r_{adv} = \underset{r; \|r\|_2 \leq \epsilon}{\operatorname{argmax}} D[p(y|x, \theta), p(y|x + r, \theta)]$
- Local Distributional Smoothness
- On testing samples



Testing

- 167 Documents
- 829 Lines
- Accuracy Measure: $\frac{\textit{Levenshtein Distance}(\textit{Label}, \textit{Prediction})}{\textit{Length}(\textit{Label})}$
- Problems with Transcription:
 - “1” vs. “one”
 - “I’ll” vs. “I will”

॥ गी० १० ॥ अथ श्रुत्वा पुराणं प्रोक्तं तदा श्रीगुरुः प्रोक्तवान् ॥
 [गी० १०] अथ श्रुत्वा पुराणं प्रोक्तं तदा श्रीगुरुः प्रोक्तवान् ॥
 गी० १० ॥ अथ श्रुत्वा पुराणं प्रोक्तं तदा श्रीगुरुः प्रोक्तवान् ॥

Results in Numbers – Unsupervised

Augmentation Omitted	CER
Multi Line Omitted	$92.58 \pm 0.96 \%$
Color Tran. Omitted	$59.38 \pm 4.03 \%$
Multi Font Properties Omitted	$32.74 \pm 0.98 \%$
Elastic Omitted	$27.95 \pm 0.24 \%$
Geometric Tran. Omitted	$27.49 \pm 0.96 \%$
Rotation Omitted	$26.97 \pm 1.48 \%$
Color Gaussian Omitted	$26.29 \pm 1.06 \%$
Geom Sine Omitted	$25.775 \pm 1.0 \%$
Color Sine Omitted	$25.09 \pm 0.45 \%$
All Augmentations	$24.23 \pm 0.94 \%$

Regularization Method	CER
Comparison Loss	$26.59 \pm 0.89 \%$
Only L2 and BN	$25.796 \pm 0.69 \%$
Dropout (RNN Hidden X 2)	$25.654 \pm 0.56 \%$
Dropout	$24.23 \pm 0.94 \%$
Dropout + EMA	$23.669 \pm 0.32 \%$
Dropout + AT	$22.23 \pm 0.7 \%$

Inference Method	CER
Greedy Inference	$24.23 \pm 0.94 \%$
Beam Search	$23.8 \pm 0.77 \%$

Results in Numbers – Transductive

Transductive Method	CER
Cycle on Test	$29.45 \pm 0.6 \%$
Cycle on Train - Best	$22.92 \pm 0.296 \%$
Best no Tran.	$22.23 \pm 0.7 \%$
DANN RNN ratio 1	$20.23 \pm 1.32 \%$
DANN CNN ratio 1	$19.23 \pm 1.34 \%$
DANN CNN+RNN ratio 1	$18.85 \pm 1.09 \%$
DANN CNN+RNN ratio 10	$16.8 \pm 0.6 \%$
VAT ratio 1	$18.24 \pm 0.35 \%$
VAT ratio 10	$15.436 \pm 0.24 \%$

The Good

[illegible]

සිංහලයාගේ ස්වභාවය

සිංහලයාගේ ස්වභාවය

[illegible][illegible][illegible]

ପଞ୍ଚମୋକ୍ଷାବଳୀ

[illegible][illegible][illegible][illegible][illegible][illegible]

ପ୍ରତିଷ୍ଠାକାରୀଙ୍କ ନାମ : ଶ୍ରୀ ରାଧାନାଥ ମହାନ୍ତି । ପ୍ରତିଷ୍ଠାକାରୀଙ୍କ ବର୍ଣ୍ଣନା : ଶ୍ରୀ ରାଧାନାଥ ମହାନ୍ତି ଏକ ବିଶିଷ୍ଟ ଶିକ୍ଷକ ଓ ସାମାଜିକ କର୍ମୀ ।

[illegible]

The Bad

પ્રબોધની વીરુદ્ધે બહુ જાણનારા એવા સંન્યાસીઓએ પણ આ પ્રકારે વિવિધ ભૂતલોકોમાં રહેલા અજ્ઞાનને દૂર કરવા માટે પ્રયત્ન કર્યો હતો.

[illegible][illegible]

418

[illegible][illegible][illegible][illegible][illegible]

What's Next?

- More Languages
- Dealing with Erased Letters
- Dealing with Curvilinear Lines

ཐུགས་རྟེན་དག་ཆེ།

Results in Numbers – Unsupervised

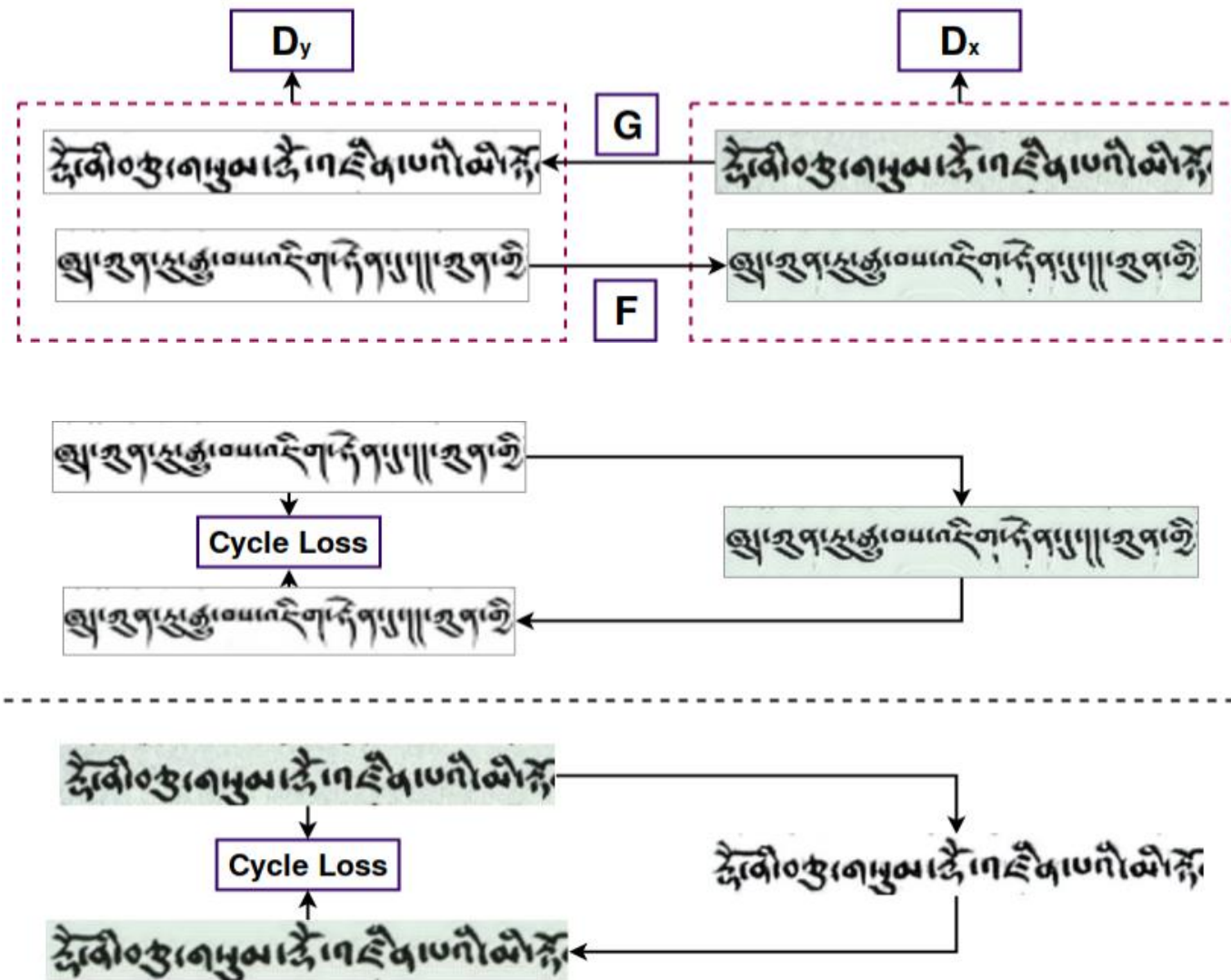
Font Omitted	CER
Shangshung Sgoba-KhraChen	$40.12 \pm 0.92 \%$
Qomolangma-Betsu	$35.83 \pm 0.71 \%$
Qomolangma-Drutsa	$32.11 \pm 0.57 \%$
Shangshung Sgoba-KhraChung	$30.02 \pm 0.78 \%$

Preprocessing	CER
No Transformation	$24.23 \pm 0.935\%$
Gray-scale	$24.09 \pm 0.79\%$
Binarization w. Color Aug.	$40.40 \pm 2.13\%$
Binarization w.o. Color Aug.	$56.93 \pm 0.648\%$

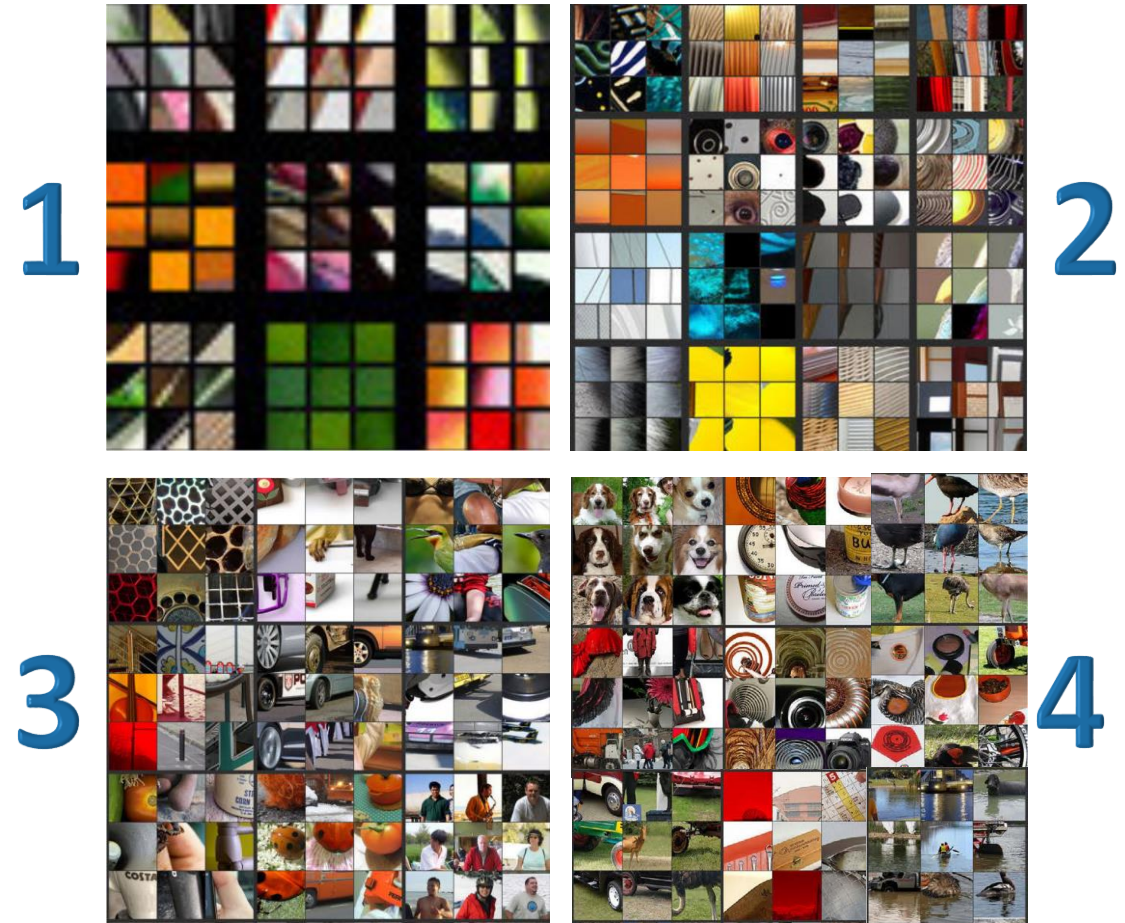
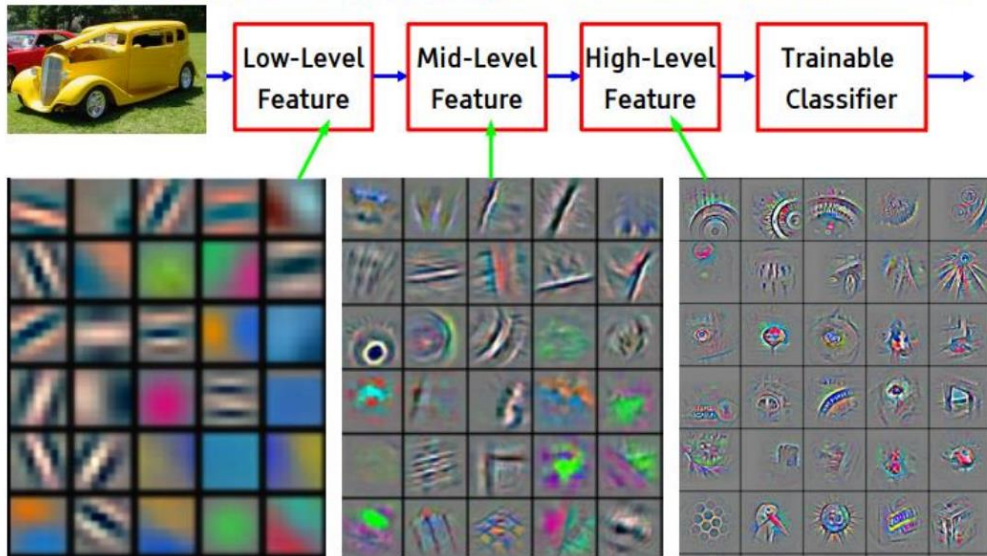
Cycle GAN

- $$\min_G \max_{D_Y} \left(E_{y \sim p_{data}(Y)} [\log D_Y(y)] + E_{x \sim p_{data}(X)} [\log (1 - D_Y(G(x)))] \right)$$

- $$\min_{F,G} \left(E_{y \sim p_{data}(Y)} [\|G(F(y)) - y\|_1] + E_{x \sim p_{data}(X)} [\|F(G(x)) - x\|_1] \right)$$



Convolutional Neural Network



“Visualizing and Understanding Convolutional Networks”

Matthew D. Zeiler and Rob Fergus 2013

“Unsupervised” Learning

Transductive Learning

Results