

Oct 16, 2019

Industrialized Capsule Network for Text Analytics

<http://bit.ly/aiconf2019>

publicis
sapient

About the Speaker



Dr. Vijay Agneeswaran

Head of Data Science, Walmart Labs
MS (Research) & PhD , IIT Madras
Post doctoral research fellowship, LSIR Labs
Professional member : ACM, IEEE (Senior)
4 Full US Patents and multiple publications (including IEEE journals)
Regular Speaker @ O'Reilly Strata & AI conference

About the Speaker



Abhishek Kumar

Senior Data Scientist @ Publicis Sapient
Masters from University of California, Berkeley
Google Developer Expert - (Machine Learning)
Speaker @ O'Reilly Strata & AI conference
Pluralsight Author

Recognizing Trends in Social Media

The image shows a Facebook news feed with several posts visible:

- A post from Alyssa Milano (@alyssamilano) with a link to www.latimes.com. It features a photo of Alyssa Milano smiling. The caption reads: "In saying #MeToo, Alyssa Milano pushes awareness campaign about sexual assault and harassment". It has 1,284 shares.
- A post from the Washington Post (@washingtonpost) with a link to www.washingtonpost.com. It features a photo of Alyssa Milano and another woman. The caption reads: "#MeToo Floods Social Media With Stories of Harassment and Assault". It has 37,525 shares.
- A post from CBS News (@CBSNews) with a link to www.cbsnews.com. It features a large red "#METOO" graphic. The caption reads: "#MeToo floods social media with stories of sexual abuse, harassment". It has 147 shares.
- A post from the Washington Post (@washingtonpost) with a link to www.washingtonpost.com. It features a photo of Alyssa Milano. The caption reads: "#MeToo made the scale of sexual abuse go viral. But is it asking too much of survivors? On Twitter, the #MeToo hashtag had been tweeted nearly 22 hours ago · 3,093 shares".

On the right side of the image, there is a vertical column of tweets from various users:

- Adelaide Kane** (@AdelaideKane) · 1h: "And we are not alone. #metoo" (115 replies, 542 retweets, 1.3K likes)
- Zelda Zonk** (@tikibetty) · 2h: "Many men have a #MeToo story - and just as with women, there is no obligation to share it if they don't want to. No one owes anyone an explanation or an apology. #MeToo" (7 replies, 98 retweets, 325 likes)
- Tonia** (@toniahazel) · 43m: "If they were drunk, they didn't know what they were doing. They should have been more careful. Sure. #MeToo" (9 replies, 9 retweets, 45 likes)
- Kent Gökhan Odelli** (@KentOdelli) · 2h: "- Women: Harassments are a part of a woman's life.
- Men: I sympathize with you, I really do, but are sure about what happened. #MeToo" (9 replies, 64 retweets, 198 likes)
- Adi** (@Illumin_Adi) · 46m: "#MeToo is a campaign I am definitely thankful for. It has brought into proximity of a reality I only imagined as distant." (1 reply, 4 retweets, 8 likes)
- Terri Michelle** (@TerriMichelle5) · 1h: "#MeToo, While not all men rape & not all people in power are bad, it is common! We need to do more than make a trending topic" (2 replies, 4 retweets, 4 likes)

Source : <https://www.cbsnews.com/news/me-too-reaches-85-countries-with-1-7-million-tweets/>

How Effective are Political Campaigns? Feedback for Various Policy Decisions

The image is a composite of two main parts. On the left, it shows a screenshot of Prime Minister Narendra Modi's Twitter profile. The profile picture is a portrait of him, and the bio reads "Chowkidar Narendra Modi @narendramodi". Below the bio, there are three engagement metrics: "Tweets 22.6K", "Following 2,121", and "Followers 46.4M". To the right of the profile, there is a large, blurred photograph of a massive political rally or campaign event, featuring a long red banner with white text and a dense crowd of people in the foreground.

Tweets
22.6K

Following
2,121

Followers
46.4M

Chowkidar Narendra
Modi [@narendramodi](#)

Tweets **Tweets & repl**

Chowkidar Narendra Modi
आप सभी चौकीदारों को मेरी

Call Center / Customer Support Performance



Supervised Learning Problems

1. Audience Segment on social platforms
2. Text categorization (Articles, News , Blogs)
3. Tagging (Queries)
4. Spam Detection
5. Reviews Classification

Text Analytics and NLP - Background

Text Analytics and NLP

Syntactic Layer

Microtext Normalization

Sentence Boundary Disambiguation

POS Tagging

Text Chunking, Lemmatization

Semantic Layer

Word Sense Disambiguation

Concept Extraction

NER (Named Entity Recognition)

Subjectivity Detection (e.g Sentiment)

Pragmatic Layer

Personality Recognition

Metaphor Understanding

Aspect Extraction

Polarity Detection

Common Models for NLP: History

First Order Logic (FOL)

All birds fly. Penguins do not fly. Penguins are birds?

Ontology Web Language (OWL)

- Resource Definition Format (RDF) for subject-predicate-object models.
- Suitable more for declarative knowledge, harder to express subjectivity.

Semantic networks

- Graphs of concepts and how they are related.
- Generalization, specialization in definitional networks, propositions in assertional networks.
- Minsky's theory of human knowledge – basis of common-sense knowledge basis for NLP tasks

Production Rules

- Recognize, resolve conflict, act – cycles.
- Scalability issues.

Networks – Bayesian, for example.

- Prior knowledge + evidence of likelihood of events.
- Limited expressiveness + difficult to determine probability of each variable.

Overlapping NLP Curves

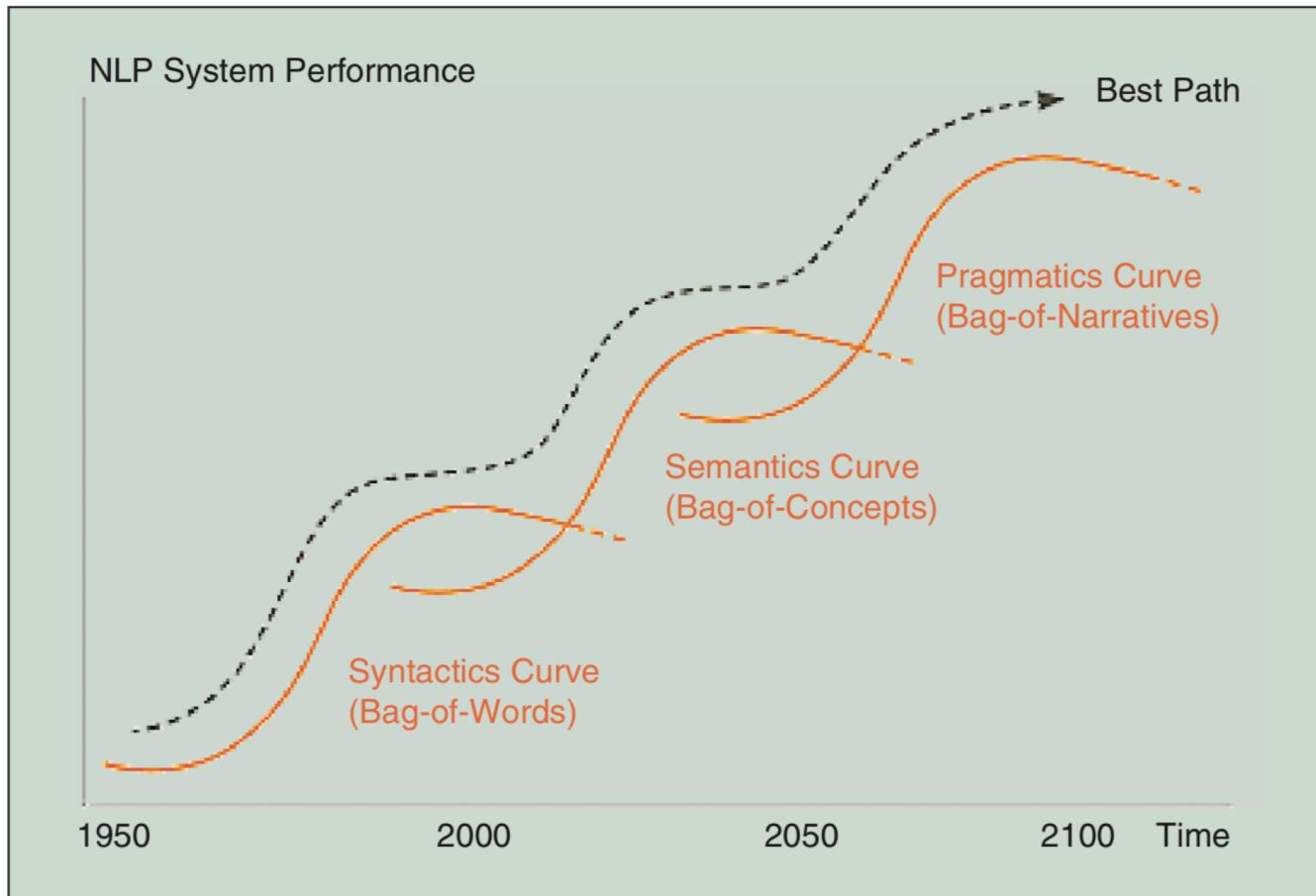


FIGURE 1 Envisioned evolution of NLP research through three different eras or curves.

Erik Cambria and Bebo White. 2014. Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *Comp. Intell.* Mag. 9, 2 (May 2014), 48-57. DOI: <https://doi.org/10.1109/MCI.2014.2307227>

Machine Learning & Deep Learning for NLP

Machine Learning

- assigning labels to words
- extract rich set of hand-crafted features
- task dependent features
- semantic role labeling task may require complex features

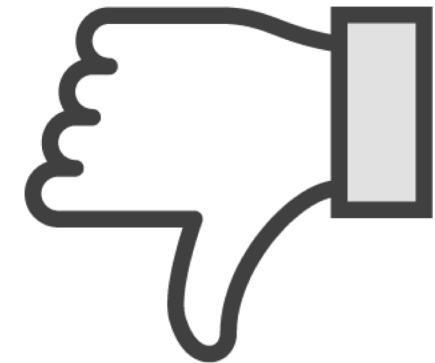
Deep Learning

- Avoid too much pre-processing
- Avoid hand-crafted feature generation
- Avoid task specific features
- Better generalization
- features are learnt by the deep layers of the network taking as input only the sentences

Text Modeling

IMBD Review

A big disappointment for what was touted as an incredible film. Incredibly bad. Very pretentious. It would be nice if just once someone would create a high profile role



Bag-Of-Words

Individual Words are Important

IMBD Review

A **big disappointment** for what was touted as an **incredible** film. **Incredibly bad.** Very **pretentious**. It would be **nice if** just once someone would create a high profile role

Word	Count	TF-IDF
Big
Disappointment
Incredible
Bad
Pretentious
Nice

Missing Context

Word Embedding

Use corpus to use context to create dense word representation

IMBD Review

A **big disappointment** for what was touted as an **incredible** film. **Incredibly bad.** Very **pretentious**. It would be **nice if** just once someone would create a high profile role

Word	Word2Vec	FastText	Glove
Big	
Disappointment	
Incredible	
Bad	
Pretentious	
Nice	

Only Word Level.. But how to model beyond "Word Meaning"

Text as Sequence Modeling Through Deep learning

Modeling spatial Relationship

IMBD Review

A **big disappointment** for what was touted as an
incredible film. **Incredibly bad.** Very **pretentious**. It
would be **nice if** just once someone would create a
high profile role

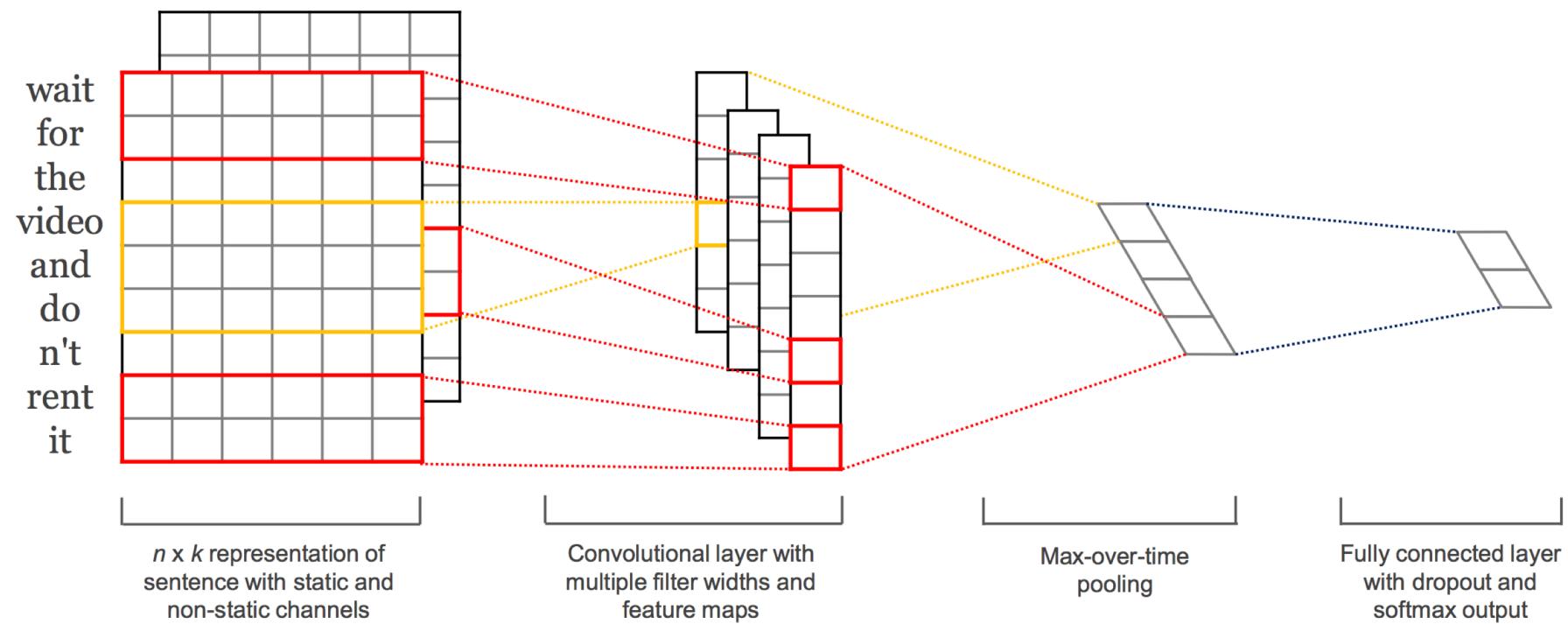
Text as Sequence Modeling Through Deep learning

Use Spatial Patterns

CNN, LSTM Approaches (use window of vector sequences)

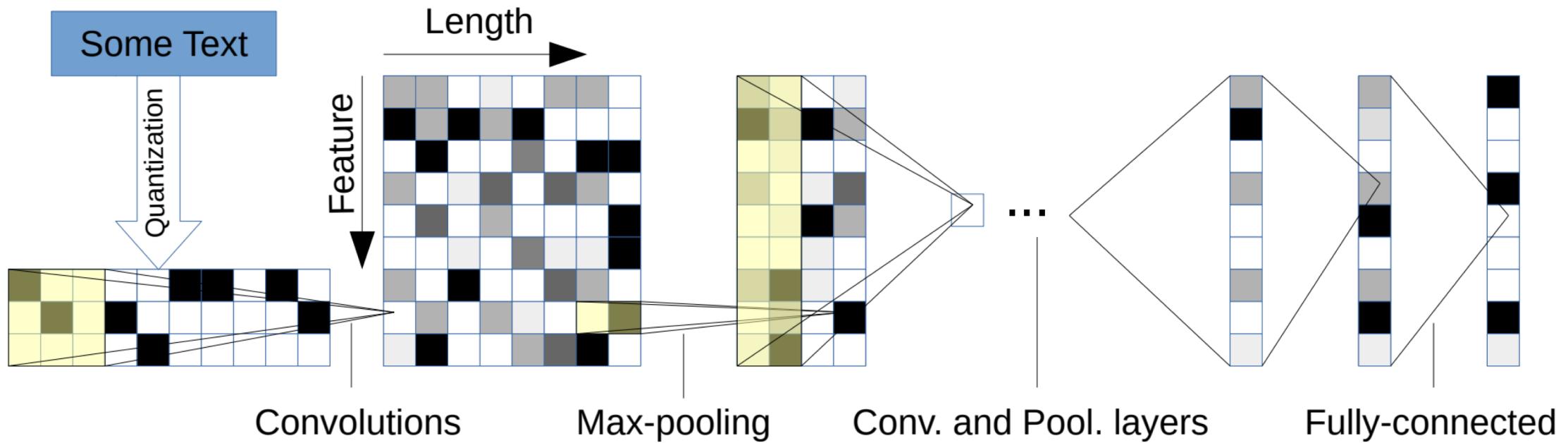
Convolution Neural Networks

Convolution Neural Networks For Sentence Classification, Yoon Kim (2014)



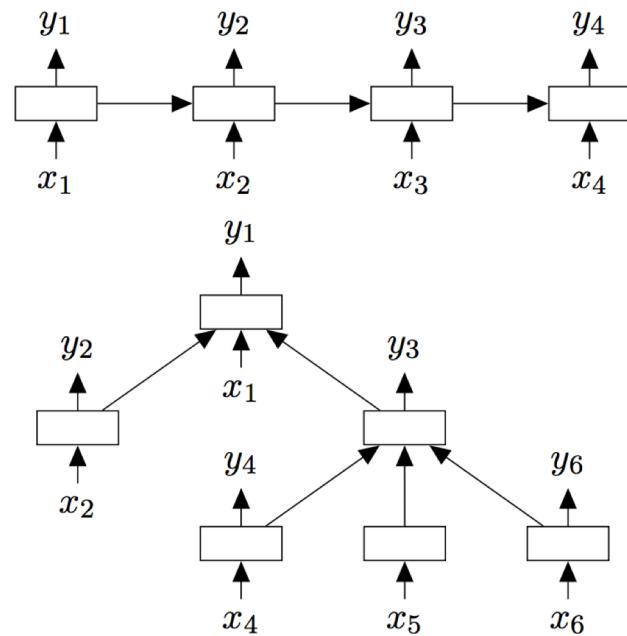
Convolution Neural Networks

Character-Level Convolution Networks For Text Classification, Zhang et.al. (2015)



LSTM (Long Short Term Memory)

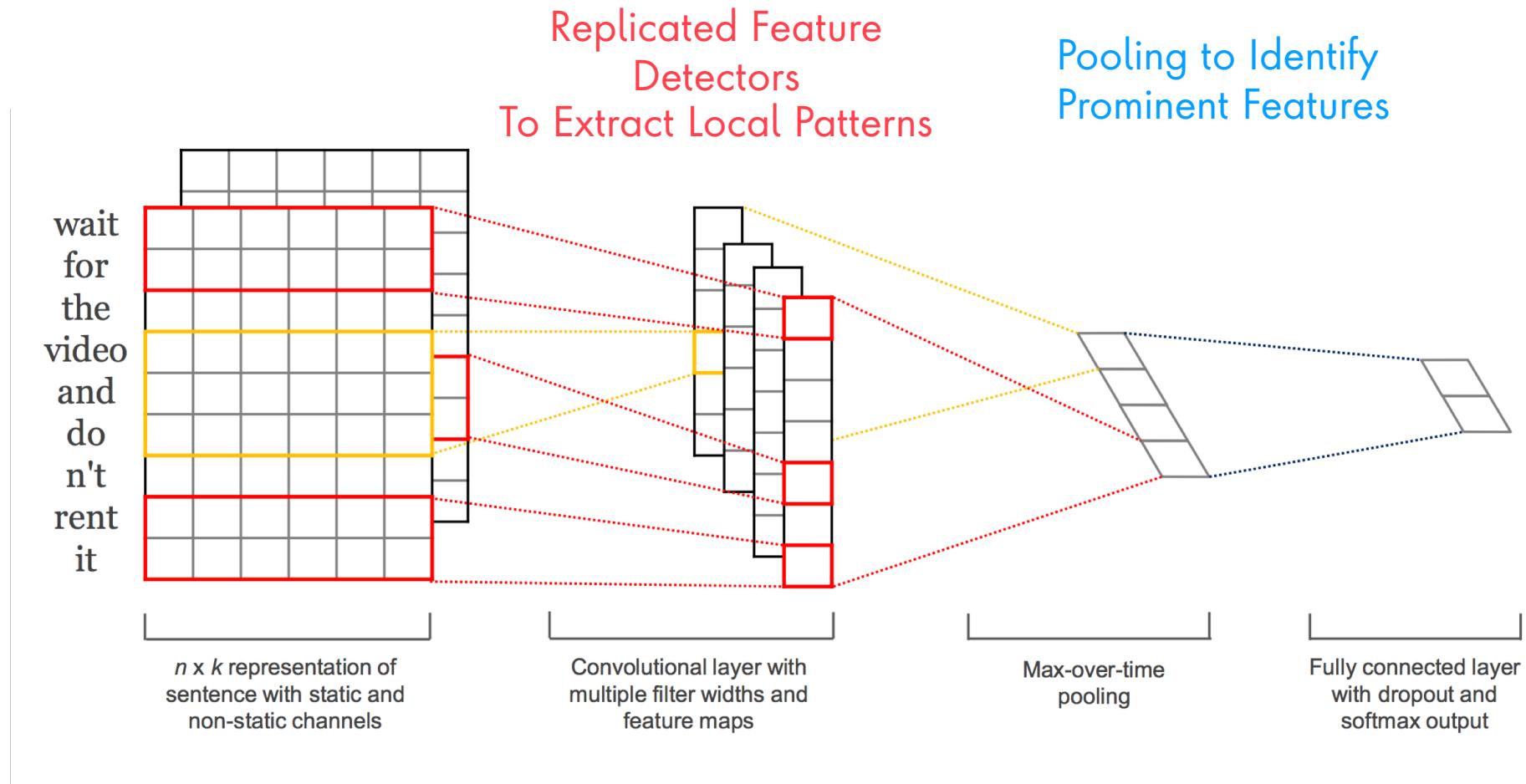
Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks, Tai et.al 2015



Chain-structured LSTM

Tree-structured LSTM

Neural Network Approach



Neural Network Approach

CNNs can deal with translation out of the box, but for robust recognition in the face of all other kinds of transformation (perspective, brightness, local patterns) there are two choices:

 kernels with large dimensions and large overlaps, at the cost of **exponentially** increasing number of parameters to be learned (use Max-Pooling technique)

Or

 Increase the size of the labelled training set in a similarly **exponential** way. (use data-augmentation techniques)

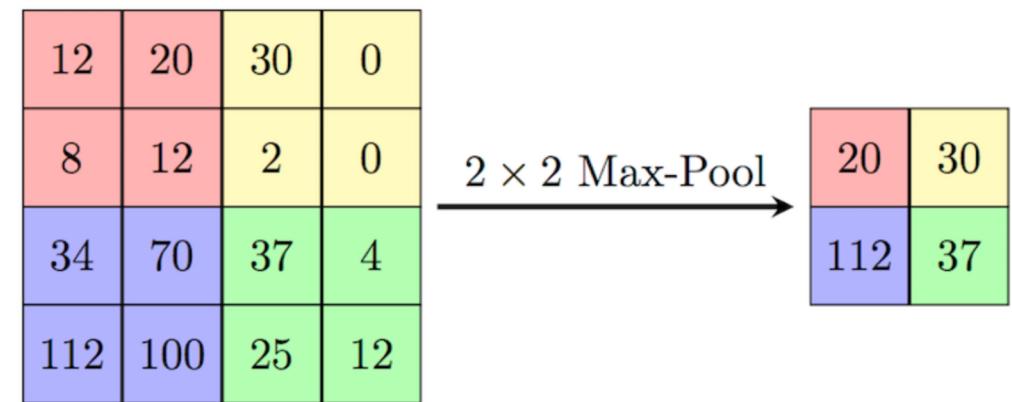
Exponential Inefficiencies

What is the problem with Pooling ?

Most active neurons are passed to next layer

Spatial information is lost

For long text / doc - important concepts will be lost



Dynamic Routing Between Capsules

Sara Sabour

Nicholas Frosst

Geoffrey E. Hinton

Google Brain

Toronto

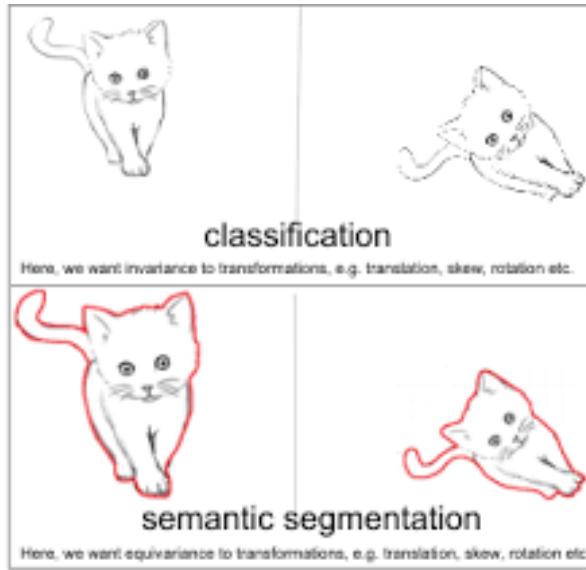
{sasabour, frosst, geoffhinton}@google.com

*Present Spatial Encoding **pooling** operation used in convolutional neural networks is a **big mistake**, and the fact that it works so well is a **disaster!**"*



Key Challenges

Solving for invariance not equivariance



CNN unable to disentangle transformations to the image such as rotation, different lighting conditions or different colours etc.

CNN is invariant to input perturbations, while capsule networks is equivariant – output of capsule nets will change when input changes due to perturbations.

How to strike the right balance?

Spatial Sensitive Approaches
(e.g. CNN)

Spatial Insensitive Approaches
(e.g. Probabilistic Topic Modeling)

Exponential
Inefficiencies

Encode rich sequence structure

Efficient (work on collection and
ignore order or local patterns)

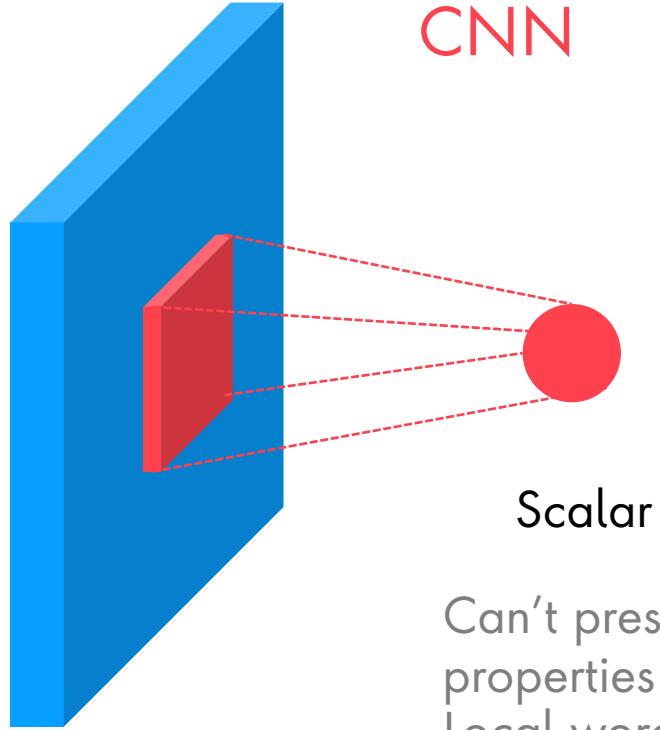
Can't encode rich sequence structure

Capsule Networks (for efficiently encode viewpoint invariance)

Background

Capsule Networks

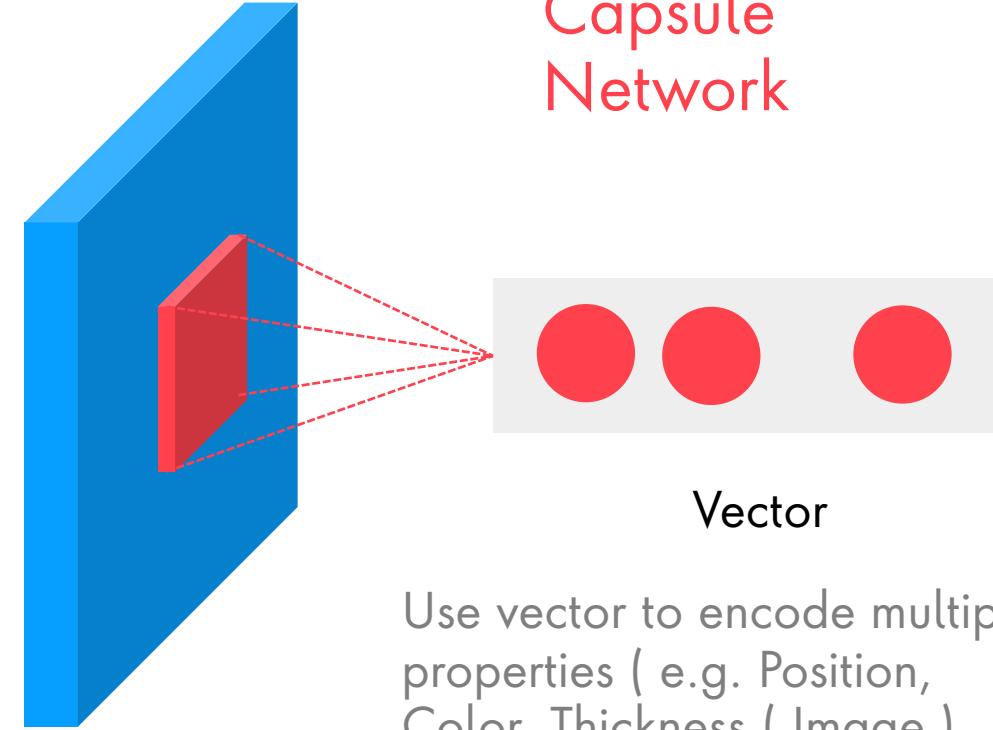
Moving from Scalar to Vector



Can't preserve
properties of Pixels /
Local word Patterns

CNN

Scalar



Use vector to encode multiple
properties (e.g. Position,
Color, Thickness (Image)
Morphology, Semantics (Text)

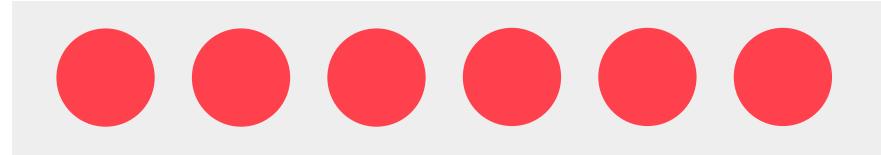


Vector length (norm) is
used to encode probability
for the output class

Capsule
Network

Vector

Capsule



Use vector to encode multiple properties (e.g. Position, Color, Thickness (Image) Morphology, Semantics (Text)

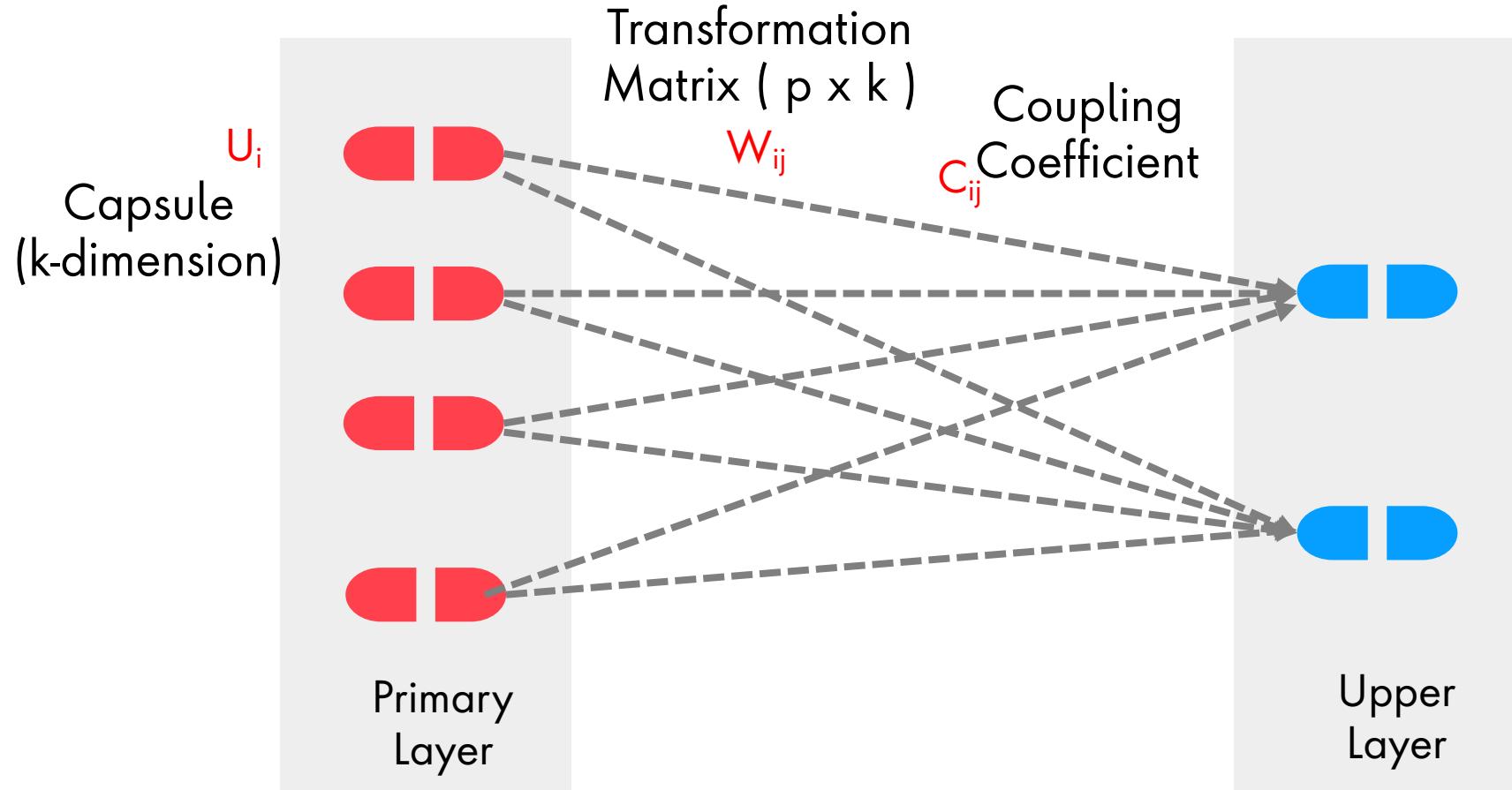
Pose
(Instantiation Parameters)



Vector length (norm) is used to encode probability for the output class

Probability

Capsule Output Calculation



c_{ij} measures how likely capsule i may activate capsule j.

- 1 Apply Transformation on Capsule

$$\hat{u}_{j|i} = W_{ij} u_i$$

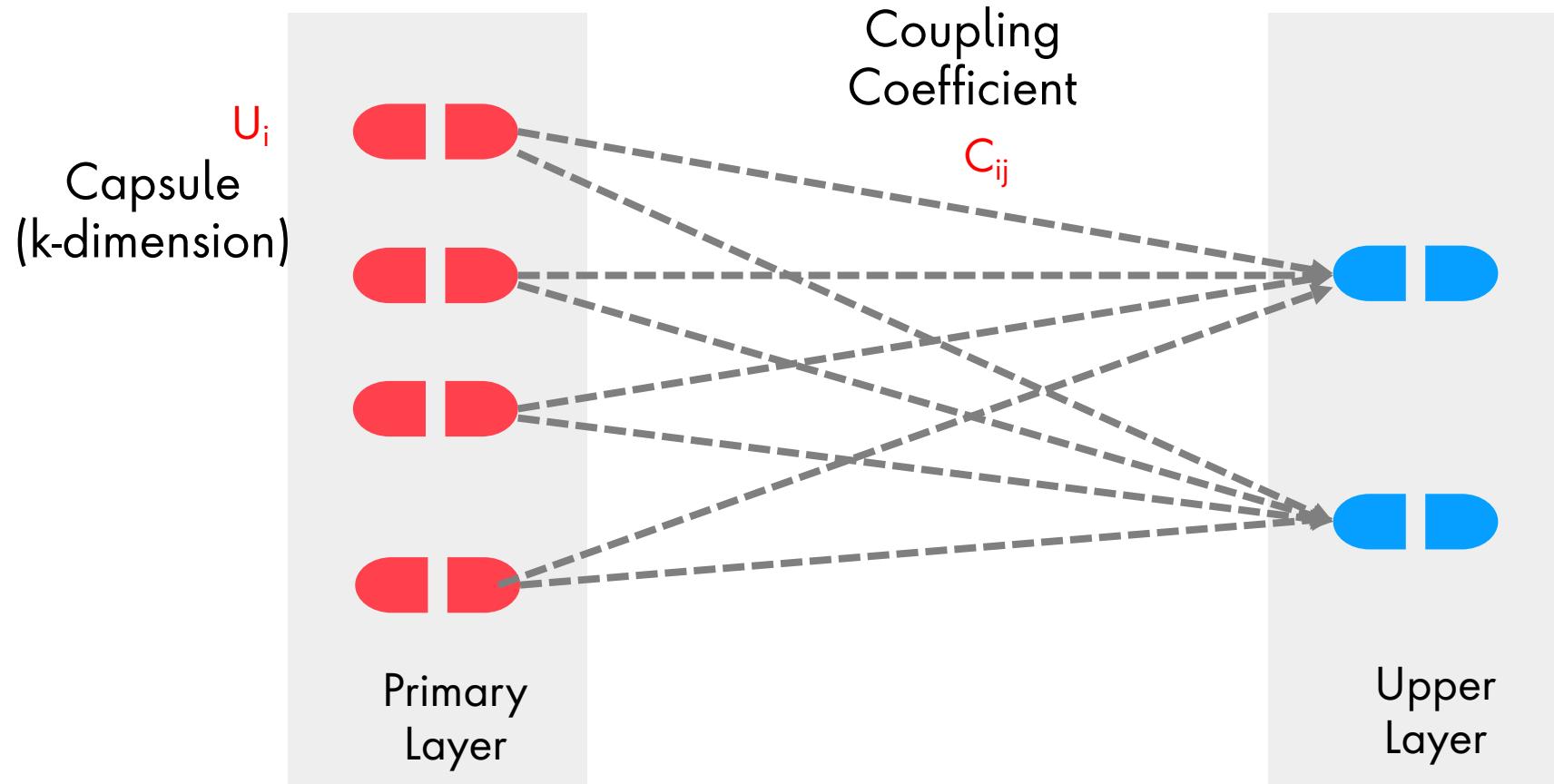
- 2 Weighted sum using coupling coefficients

$$s_j = \sum_i c_{ij} \hat{u}_{j|i}$$

- 3 Apply Squashing to have output between 0 and 1

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|}$$

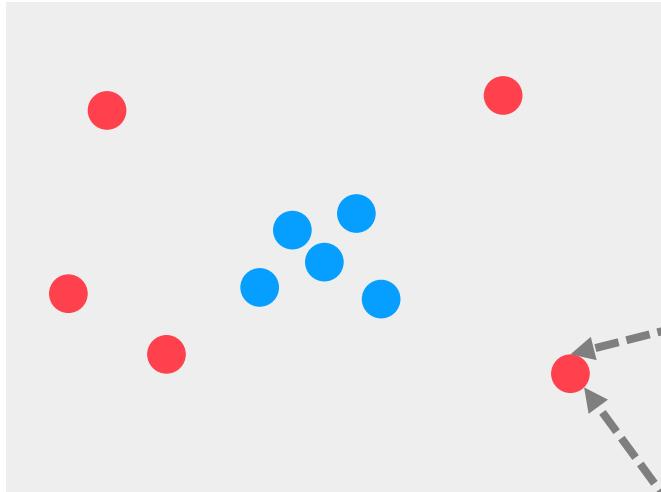
Dynamic Routing



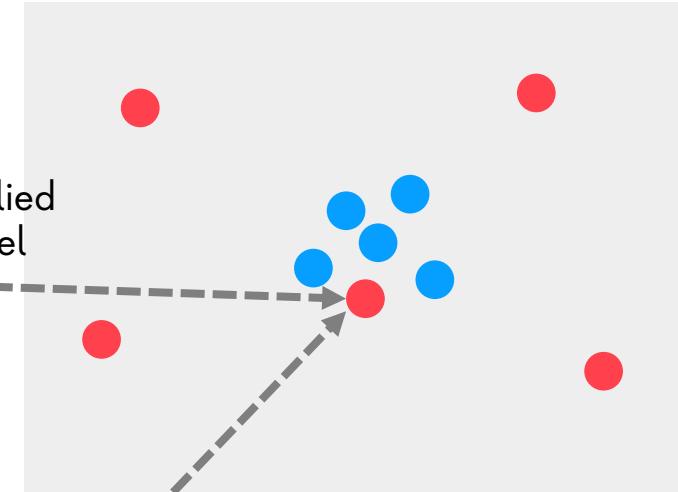
Coupling Coefficients (Scalar weights) determined through "**Dynamic Routing**" Process

Dynamic Routing (Routing by Agreement)

Higher Level Capsule - A



Higher Level Capsule - B



Send Less
(Low Coupling Coefficient)

Lower Level
Capsule



$$\hat{u}_{j|i} = W_{ij}u_i$$

Matrix Weight Multiplied
Output of Lower Level
Capsule

Send More
(High Coupling Coefficient)

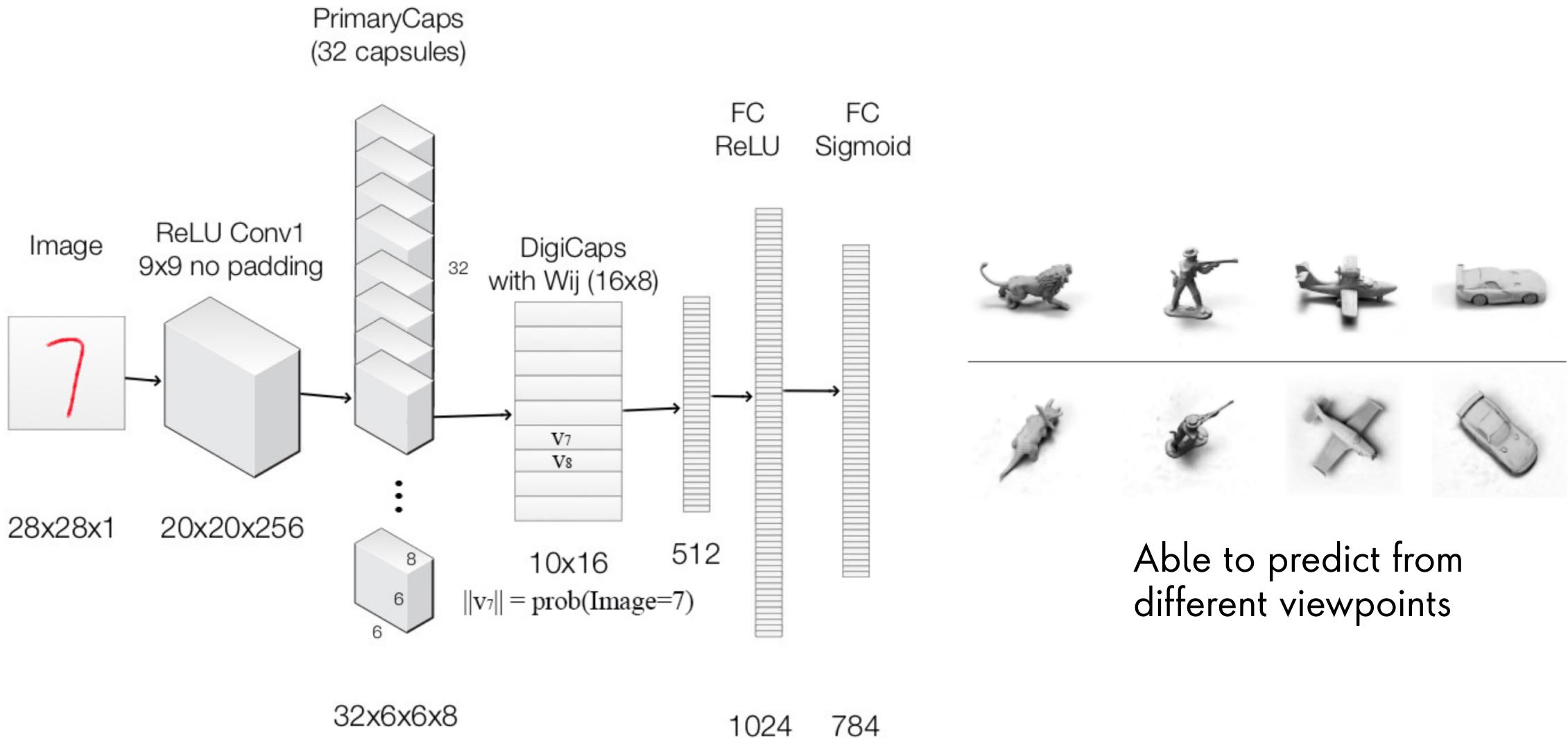
Softmax
Routing

$$c_{ij} = \frac{\exp b_{ij}}{\sum_k \exp b_{ik}}$$

Dynamic Routing (Routing by Agreement)

- 1 Calculate Prediction vector (prediction from capsule i to capsule j) $\hat{u}_{j|i} = W_{ij} u_i$
- 2 Calculate Activity vector v_j (capsule j output) $s_j = \sum_i c_{ij} \hat{u}_{j|i}$ $v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|}$
- 3 Check if activity is closer to prediction using dot product $b_{ij} \leftarrow \hat{u}_{j|i} \cdot v_j$
- 4 Calculate coupling coefficient as softmax of b_{ij} $c_{ij} = \frac{\exp b_{ij}}{\sum_k \exp b_{ik}}$
- 5 Iteratively update b_{ij} $b_{ij} \leftarrow b_{ij} + \hat{u}_{j|i} \cdot v_j$

Capsule Network Architecture for Image Classification



Background

Capsule Networks For Text

Capsule Network For Text

Investigating Capsule Networks with Dynamic Routing for Text Classification

Wei Zhao^{1,2}, Jianbo Ye³, Min Yang^{1*}, Zeyang Lei⁴, Soufei Zhang⁵, Zhou Zhao⁶

¹ Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

² Tencent

³ Pennsylvania State University

⁴ Graduate School at Shenzhen, Tsinghua University

⁵ Nanjing University of Posts and Telecommunications

⁶ Zhejiang University

<https://arxiv.org/pdf/1804.00538.pdf>

Text Classification using Capsules

Jaeyoung Kim, Sion Jang and Sungchul Choi

TEAMLAB, Gachon University

teamlab.gachon@gmail.com

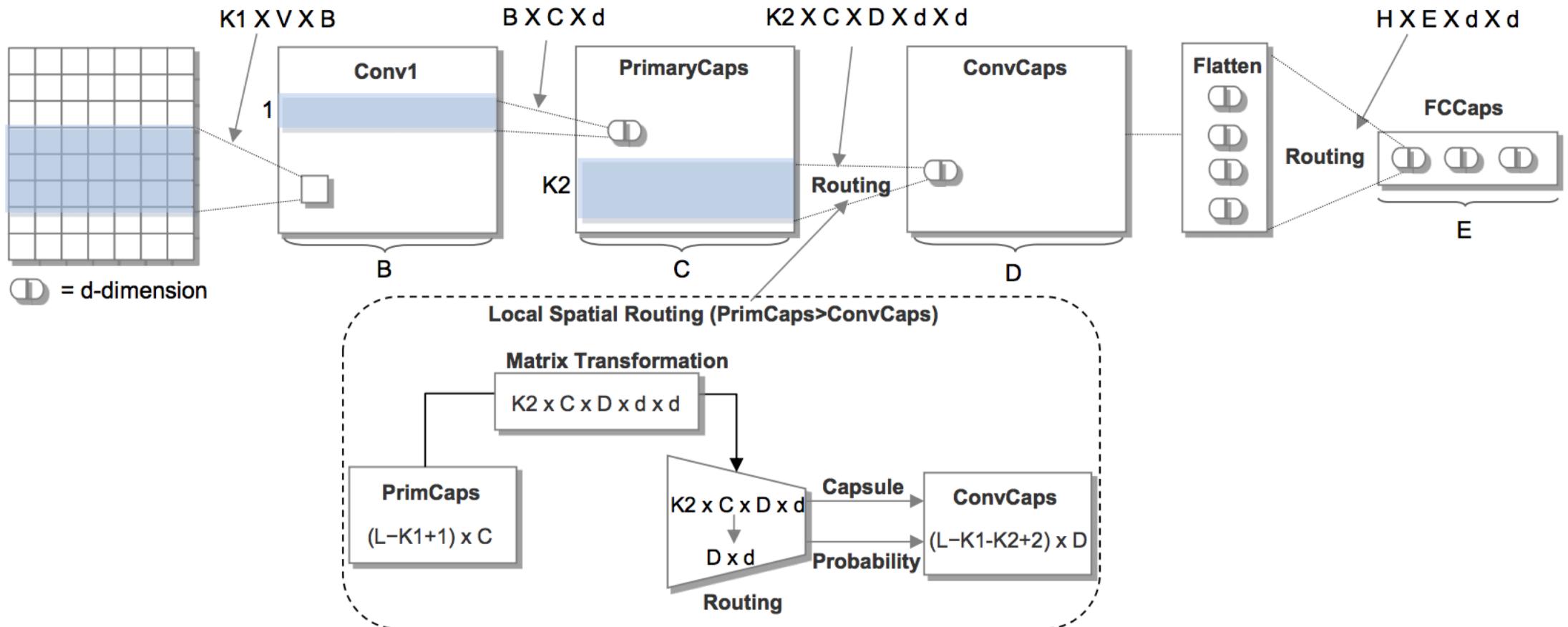
Eunjeong Park

NAVER

lucy.park@navercorp.com

<https://arxiv.org/pdf/1808.03976.pdf>

Capsule Network For Text



<https://arxiv.org/pdf/1804.00538.pdf>

Capsule Network For Text : Dealing with Noise

Orphan Category

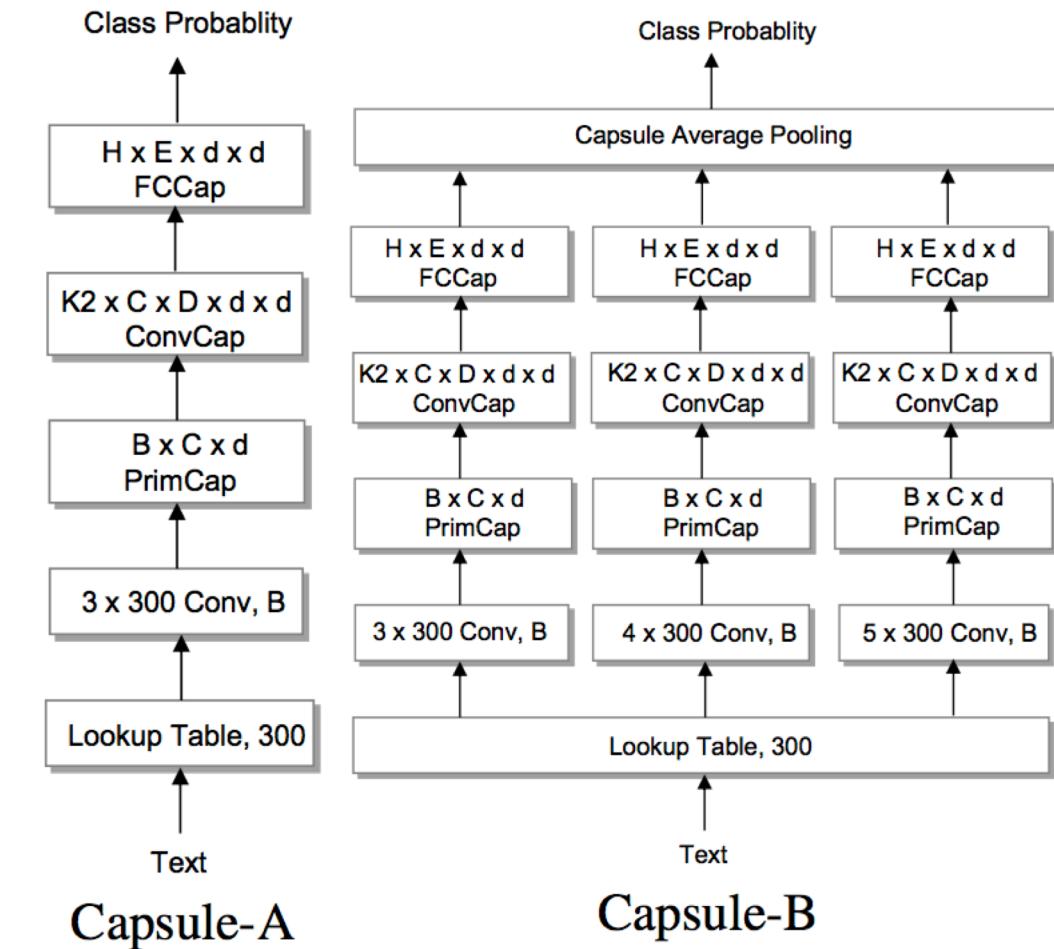
- Add extra class with no-category in final layer
- Background noise such as Stop Words or unimportant words redirected towards Orphan category

Leaky Softmax

- Leaky Softmax instead of regular Softmax
- To route the noise to orphan Category without any additional parameters and computation consuming.

Capsule Network For Text

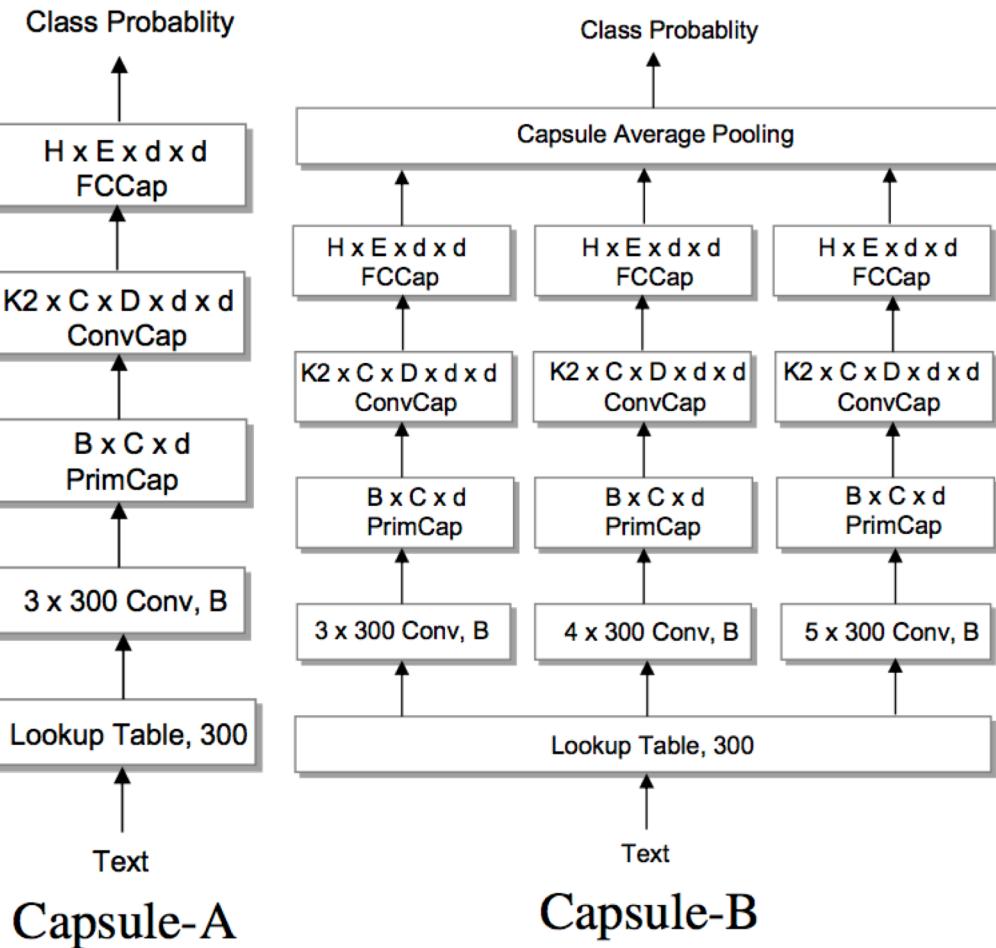
Dataset	Train	Dev	Test	Classes	Classification Task
MR	8.6k	0.9k	1.1k	2	review classification
SST-2	8.6k	0.9k	1.8k	2	sentiment analysis
Subj	8.1k	0.9k	1.0k	2	opinion classification
TREC	5.4k	0.5k	0.5k	6	question categorization
CR	3.1k	0.3k	0.4k	2	review classification
AG's news	108k	12.0k	7.6k	4	news categorization



<https://arxiv.org/pdf/1804.00538.pdf>

Capsule Network For Text

	MR	SST2	Subj	TREC	CR	AG's
LSTM	75.9	80.6	89.3	86.8	78.4	86.1
BiLSTM	79.3	83.2	90.5	89.6	82.1	88.2
Tree-LSTM	80.7	85.7	91.3	91.8	83.2	90.1
LR-LSTM	81.5	87.5	89.9	-	82.5	-
CNN-rand	76.1	82.7	89.6	91.2	79.8	92.2
CNN-static	81.0	86.8	93.0	92.8	84.7	91.4
CNN-non-static	81.5	87.2	93.4	93.6	84.3	92.3
CL-CNN	-	-	88.4	85.7	-	92.3
VD-CNN	-	-	88.2	85.4	-	91.3
Capsule-A	81.3	86.4	93.3	91.8	83.8	92.1
Capsule-B	82.3	86.8	93.8	92.8	85.1	92.6



<https://arxiv.org/pdf/1804.00538.pdf>

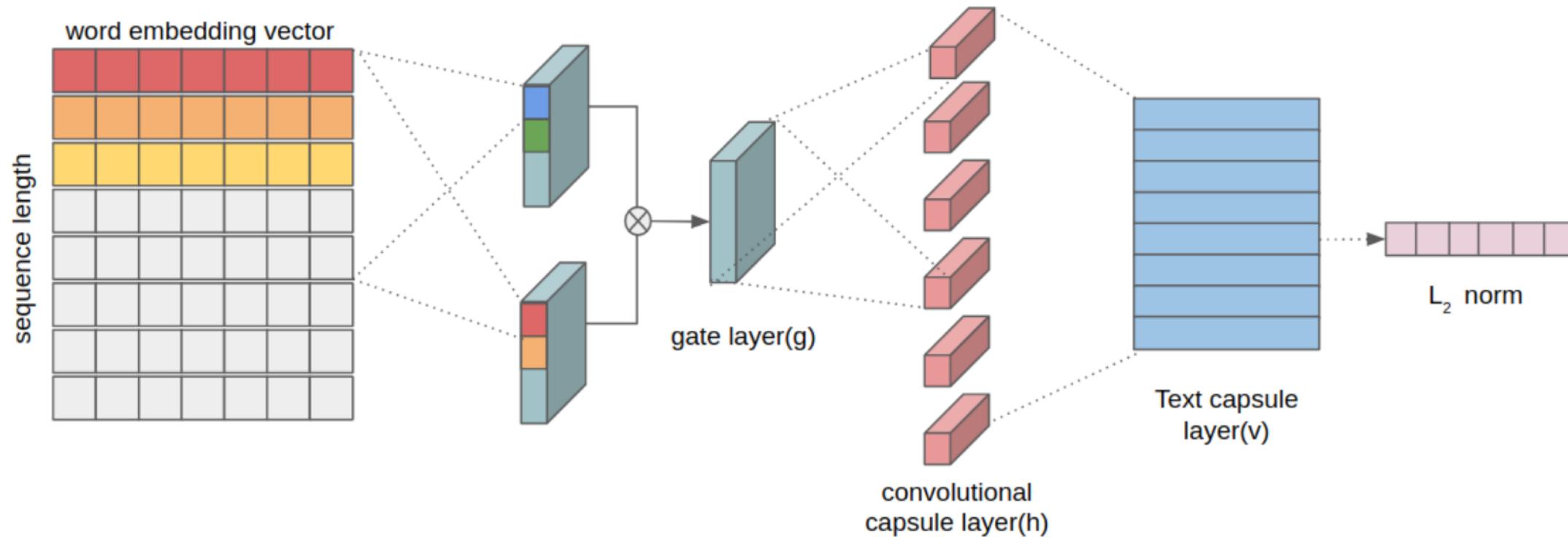
Capsule Network For Text : Multi-Label Classification

Reuters-Multi-label				
	ER	Precision	Recall	F1
LSTM	23.3	86.7	54.7	63.5
BiLSTM	26.4	82.3	55.9	64.6
CNN-rand	22.5	88.6	56.4	67.1
CNN-static	27.1	91.1	59.1	69.7
CNN-non-static	27.4	92.0	59.7	70.4
Capsule-A	57.2	88.2	80.1	82.0
Capsule-B	60.3	95.4	82.0	85.8

Multi-Label Classification Challenges

- Label space is expanded from n to 2^n
- More labelled dataset required for CNN and LSTM to work
- Capsule performed well (better generalization by capturing patterns using capsules) without extensive labeled data

Capsule Network For Text



- Uses Gate Linear Unit (Dauphin et al., 2016) for selecting features to be activated
- Unlike Pooling , ELU gate unit doesn't lose spatial information

<https://arxiv.org/pdf/1808.03976.pdf>

CapsNet for Text : Multi-Task Learning

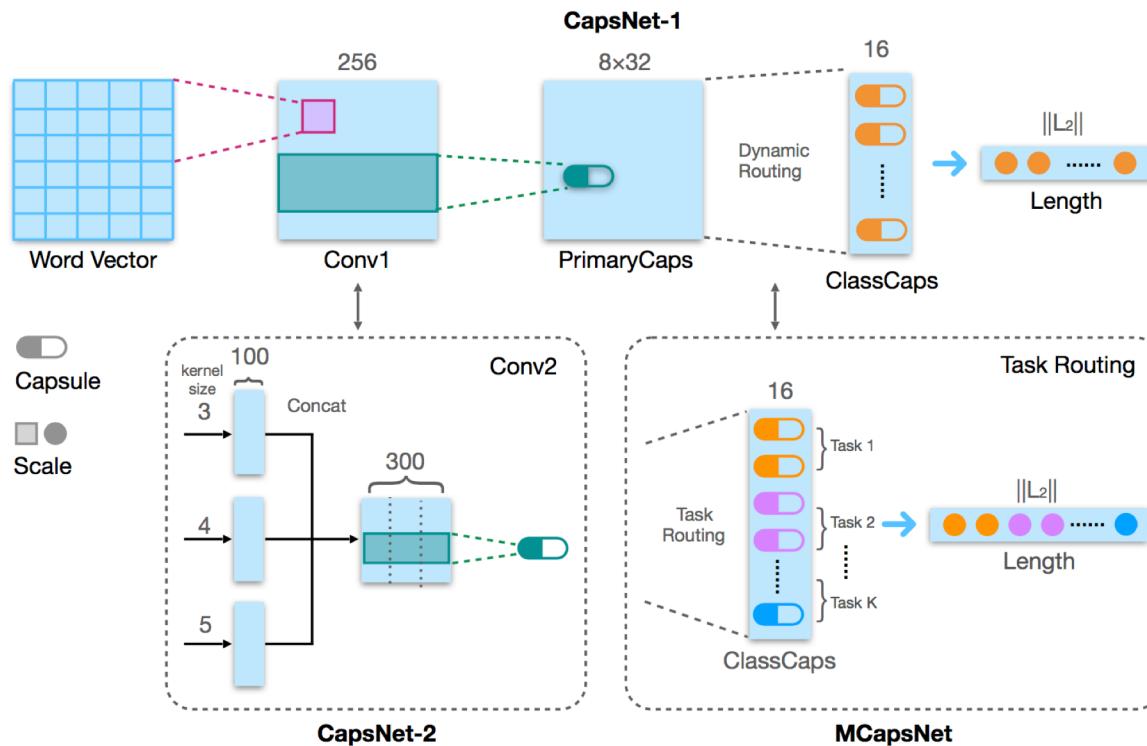
MCapsNet: Capsule Network for Text with Multi-Task Learning

Liqiang Xiao^{1,2}, Honglun Zhang^{1,2}, Wenqing Chen^{1,2}, Yongkun Wang³, Yaohui Jin^{1,2}

¹ State Key Lab of Advanced Optical Communication System and Network,
Shanghai Jiao Tong University

² MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

³ Network and Information Center, Shanghai Jiao Tong University
{jinyh}@sjtu.edu.cn

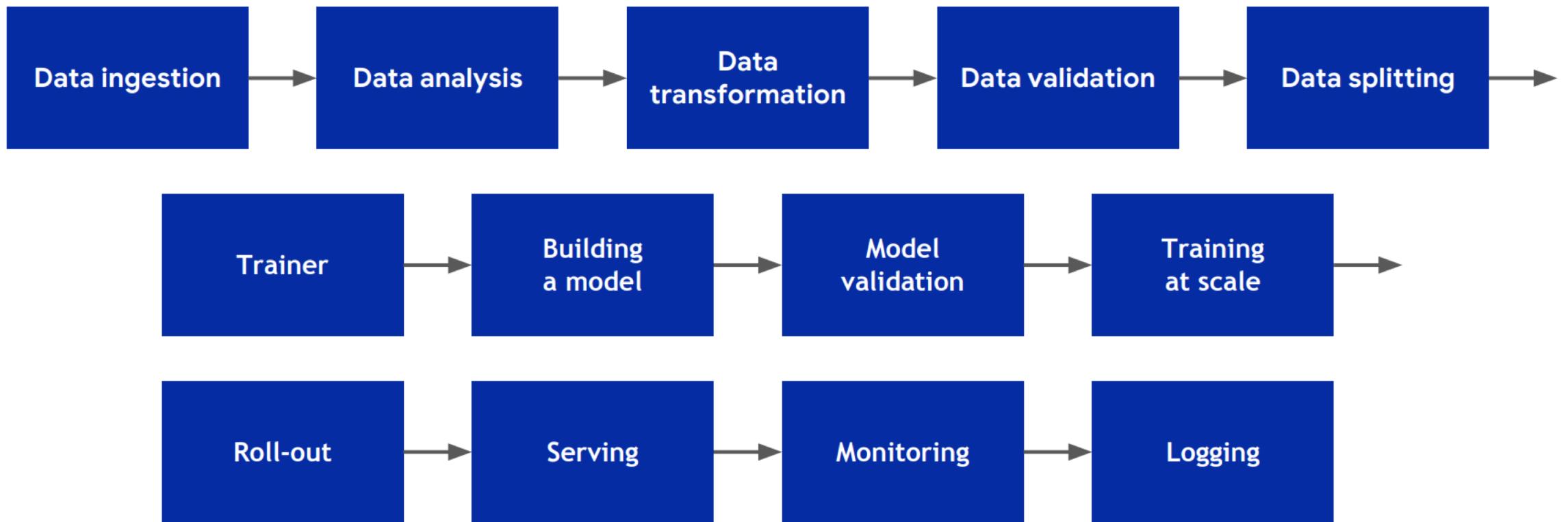


Dataset	MR	SST-1	SST-2	Subj	TREC	AG's	Avg. Δ
BiLSTM	79.3	46.2	83.2	90.5	89.6	88.2	+0
MT-GRNN	-	49.2	87.7	89.3	93.8	-	+2.6
MT-RNN	-	49.6	87.9	94.1	91.8	-	+3.5
MT-DNN	82.1	48.1	87.3	93.9	92.2	91.8	+2.9
MT-CNN	81.6	49.0	86.9	93.6	91.8	91.9	+3.0
CapsNet-1	81.5	48.1	86.4	93.3	91.8	91.1	+2.5
CapsNet-2	82.4	48.7	87.8	93.6	92.9	92.3	+3.3
MCapsNet	83.5	49.7	88.6	94.5	94.2	93.8	+4.6

Section 2

Industrialization of Capsule Networks

Platform



What is Kubeflow ?



Kubeflow

“The Machine Learning Toolkit for Kubernetes”

A curated set of compatible tools and artifacts that lays a foundation for running production ML apps



Notebook



TF Model Training



TF serving
Seldon,
TensorRT



Kubeflow Pipelines



Pipelines



AMBASSADOR



argo



Pachyderm



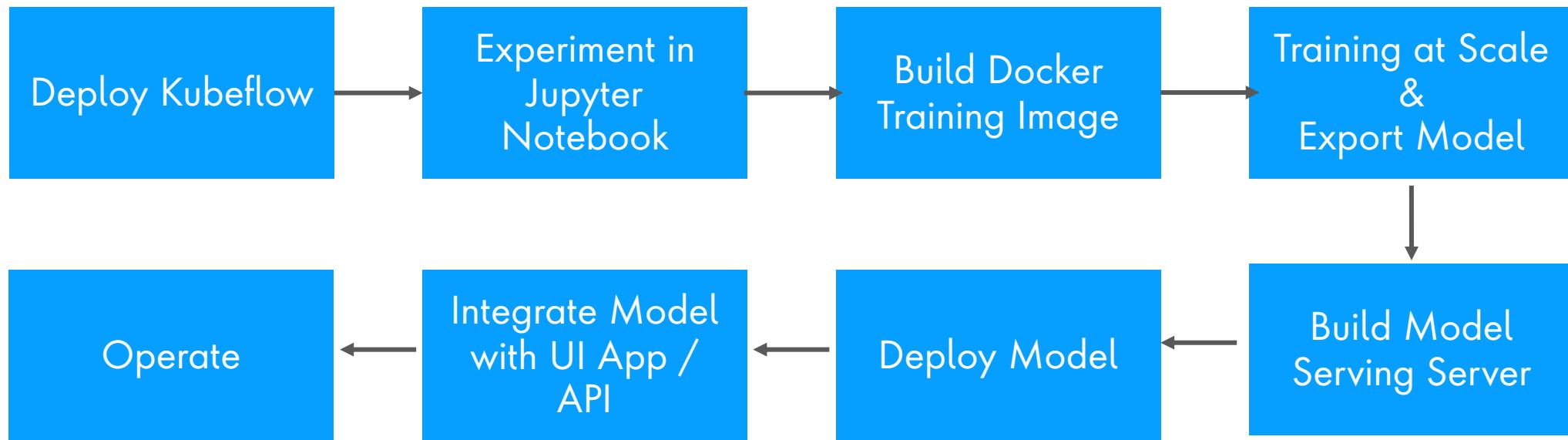
PyTorch



Chainer

Multi-framework
Integration

Typical Workflow



Key Summary and Demo

- Kubeflow : running capsnet on Kubeflow
- Notebooks with GPU configuration (Multi-GPU training Strategy)
- TF Job : Distributed jobs
- Katib for Hyper-Parameter Tuning

Basic Implementation using Tensorflow

```
def squash(x, axis=-1):
    s_squared_norm = K.sum(K.square(x), axis, keepdims=True) + K.epsilon()
    scale = K.sqrt(s_squared_norm)/ (0.5 + s_squared_norm)
    return scale * x

#define our own softmax function instead of K.softmax
def softmax(x, axis=-1):
    ex = K.exp(x - K.max(x, axis=axis, keepdims=True))
    return ex/K.sum(ex, axis=axis, keepdims=True)
```

```
def call(self, u_vecs):
    if self.share_weights:
        u_hat_vecs = K.conv1d(u_vecs, self.W)
    else:
        u_hat_vecs = K.local_conv1d(u_vecs, self.W, [1], [1])

    batch_size = K.shape(u_vecs)[0]
    input_num_capsule = K.shape(u_vecs)[1]
    u_hat_vecs = K.reshape(u_hat_vecs, (batch_size, input_num_capsule,
                                         self.num_capsule, self.dim_capsule))
    u_hat_vecs = K.permute_dimensions(u_hat_vecs, (0, 2, 1, 3))
    #final u_hat_vecs.shape = [None, num_capsule, input_num_capsule, dim_capsule]

    b = K.zeros_like(u_hat_vecs[:, :, :, 0]) #shape = [None, num_capsule, input_num_capsule]
    for i in range(self.routings):
        c = softmax(b, 1)
        o = K.batch_dot(c, u_hat_vecs, [2, 2])
        if i < self.routings - 1:
            o = K.l2_normalize(o, -1)
            b = K.batch_dot(o, u_hat_vecs, [2, 3])
    return self.activation(o)

def compute_output_shape(self, input_shape):
    return (None, self.num_capsule, self.dim_capsule)
```

Distribution Strategy

- Converting Keras Model to Estimator
- Use Mirrored Strategy

```
config = tf.estimator.RunConfig(  
    experimental_distribute=tf.contrib.distribute.DistributeConfig(  
        train_distribute=tf.contrib.distribute.CollectiveAllReduceStrategy(  
            num_gpus_per_worker=0),  
        eval_distribute=tf.contrib.distribute.MirroredStrategy(  
            num_gpus_per_worker=0)))
```

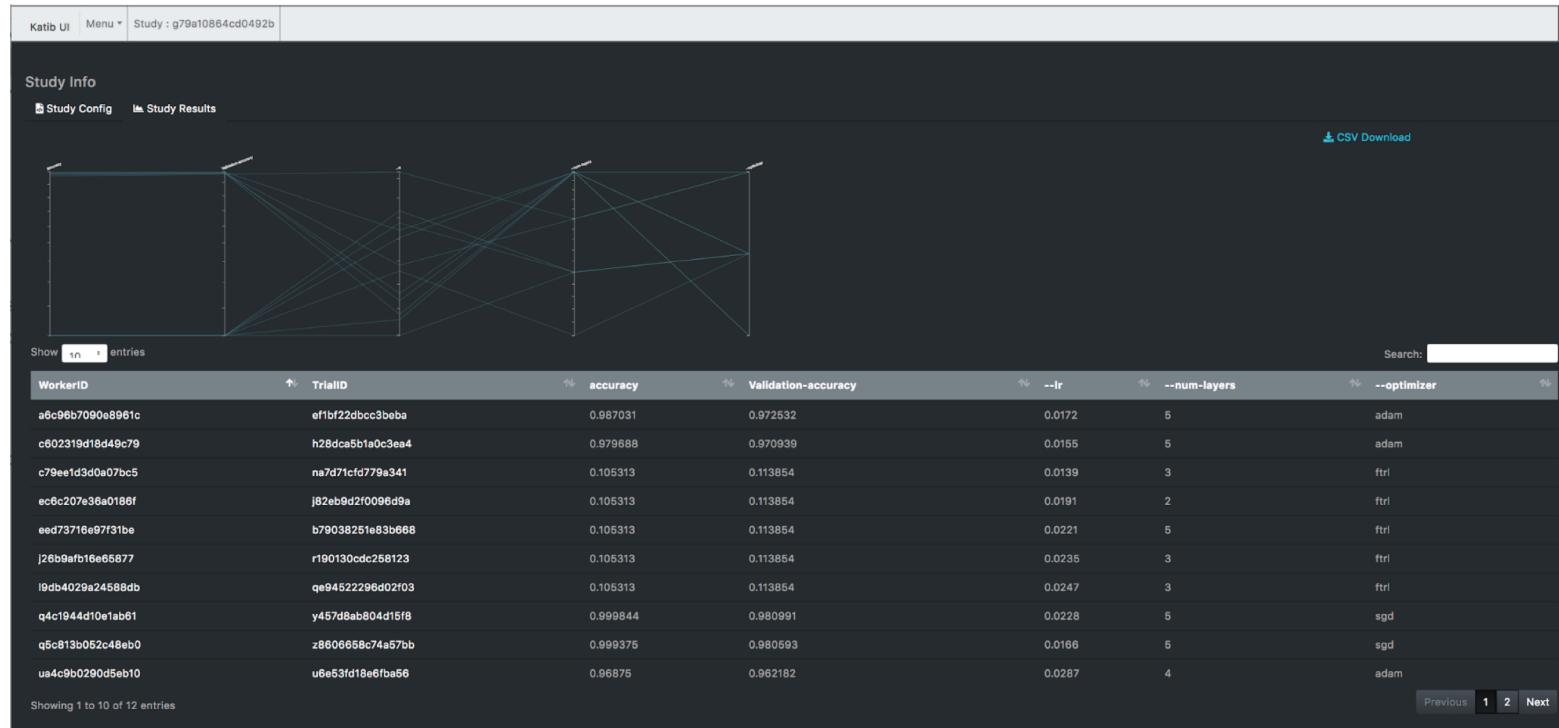
Hyper-Parameter Tuning

Katib (by NTT)

- Pluggable micro-service
- Multiple architecture for Hyper-Parameter tuning (Grid, Random, Bayesian)
- Different optimization algorithms
Different frameworks

StudyJob (K8s CRD)

- Hides complexity from user
- No code needed to do hyper-parameter tuning



Further Exploration (Efficiency Side)

- Trying Efficient Routing Algorithm
- Use Matrix Capsules instead of Vector Capsules
- Tensorflow Code Optimization

Fast Dynamic Routing Based on Weighted Kernel Density Estimation

Suofei Zhang¹, Wei Zhao², Xiaofu Wu¹, Quan Zhou¹

¹Nanjing University of Post and Telecommunication

²SIAT, Chinese Academy of Sciences

<https://arxiv.org/pdf/1805.10807.pdf>

MATRIX CAPSULES WITH EM ROUTING

Geoffrey Hinton, Sara Sabour, Nicholas Frosst

Google Brain

Toronto, Canada

{geoffhinton, sasabour, frosst}@google.com

<https://openreview.net/pdf?id=HJWLfGWRb>

Next Steps

- Provide feedback on the tutorial
- Session Content
 - <http://bit.ly/aiconf2019>
 - Know more on Kubeflow (Strata SFO, 2019)
 - <http://bit.ly/deep-recsys>
- Share
 - Progress, Issues, Use-cases
 - Connect on LinkedIn
 - Twitter
 - [@a_vijaysrinivas](#)
 - [@meabhishekkumar](#)

thank you