A Framework for Arabic Media Analytics
With Special application to Calls Centers Recordings Mining

Dr. Mohsen Rashwan
Dr. Elsayed Hemayed
Dr. Sherif Abdou
Dr. Mohamed Islam

The Big Video Data Sources

Video Data Mining

- Most of today's commercial video retrieval systems still solely rely on meta-data for indexing video content.
- The problem has proven very challenging, and the state of the art is still very limited.
- There is a semantic gap between Natural Language and Video and Audio computational representations.

The Project Objective

- Develop a unified framework for understanding Arabic multimedia content.
- Investigate methods to map
  - Visual
  - Auditory
  - Textual information

  in unstructured videos to a common semantic space

Proposed Framework

- Encoding of video Information
  - Visual Concepts
    - Object Concepts: Person face, specific building,..etc
    - Scene Concepts : City scene, hallway, landscape,..etc
    - Action Concepts : Fighting, running, blowing a candle, etc
  - Audio Concepts : Gunshots, explosions, running water, etc.
  - Speech Recognition Content: Spoken text in the video
  - Textual Content: Text displayed in the video

The Main Challenge

The main challenge is to map all these concepts and contents from a given video to a unified representation that can be the basis to measure:

1. Similarity between videos,
2. Similarity between a video and a query text.

Project Team

- Dr Mohsen Rashwan

  (Cairo University, Communication Dep)

- Dr Sayed Humaid

  (Cairo University, Computer Engineering)

- Dr Mohamed Islam

  (Cairo University, Biomedical Engineering)

- Dr Sherif Abdou

  (Cairo University, Faculty of Computers and Information)

- 8 Master and PhD students

Developed Systems / Tools

- Large Vocabulary Arabic ASR for MSA Egyptian Dialect.
- Automatic Diarization tool for segmenting Broad Cast recordings to speech and non-speech segments and clustering the recording speakers.
- Emotion recognition for call centers recordings.
- Face Verifier: The tool can compare two face and decide if they are identical or not.
- Face Classifier: Trained on 100 different persons and able to recognize them.

Developed Systems / Tools

- Text extractor from news caption: The tool extracts the text from news caption.
- Object detection: The tool is an implementation of Yolo V3 to detect the learned objects in images.
- Ads extraction: The tools extract the learned ads from TV video clips.
- Arabic Name Entitles Extraction.
- Topics classification for call centers recordings.
- Sentiment Analysis for call centers recordings.

Major Achievments

Two major technical progress has been achieved in the project:

- State of art Arabic Speech Recognition based on advanced deep leaning models for both Modern Standard Arabic and Dialectal Egyptian Arabic.

- State of art Face Verifier based Hybrid Siamese Network.

Published Papers

- Ali Alabed, Omar A. Nasr, Elsayed Hemayed, Speeding up Dominant Object Video Segmentation, ICENCO 2017.

- Mohamed Sameer, Elsayed Hemayed, Zero-Shot Learning for Media Mining: Person Spotting and Face Clustering in Talk Shows, Submitted on 2/2019

- Nehal Khalid, Magda Fayek, Elsayed Hemayed, Face Verification and Clustering Using Hybrid Siamese Network, submitted on 2/2019

Thesis

- Amira Alsharkawy, Object recognition using deep convolutional neural networks, MSc, Cairo University, 2017
- Ali Alabed, A Framework for Text Caption Extraction from Arabic News Video, MSc, Cairo University, 2018
- Mohamed Sameer, Zero-Shot Learning for Media Mining: Person Spotting and Face Clustering in Talk Shows, PhD, Cairo University, Expected 2019
- Nehal Khalid, Face Verification and Clustering Using Hybrid Siamese Network, PhD, Cairo University, Expected 201

Arabic Media Analytics Framework

Video Summarization

Shot segmentation
Key fames extraction

Image Processing

Object description
Person faces Identification
Advertisement and logos

Closed Caption Extraction

Segmented caption line images

Arabic OCR

Caption text

Speech Diarization

Speech and non_speeh
Speaker segmentation
Speaker clustering

Speech Recognition

Speech transcripts
(MSA, Dialectal)

Speaker Recognition

Speaker Identity

Text Processing

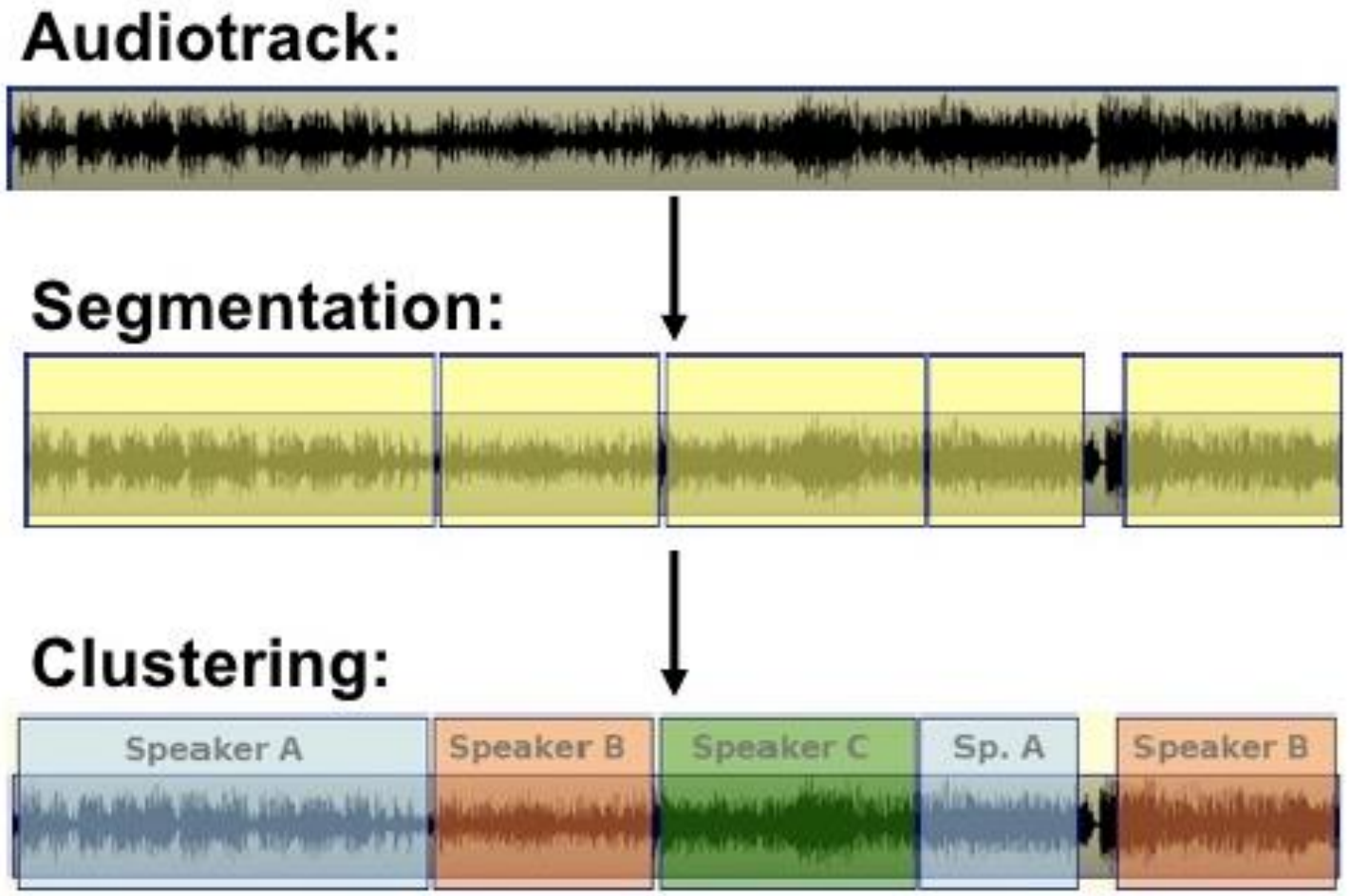
Text meta data

Semantic Layer

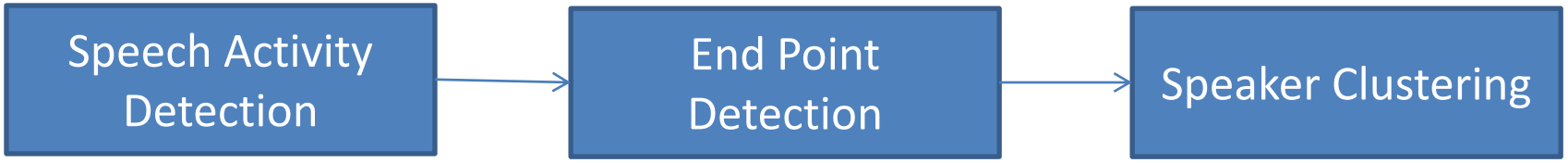
- Key words Search
- Topic classification
- Name Entities
- Sentiment Analysis

Speech Track Efforts

Video Diarization

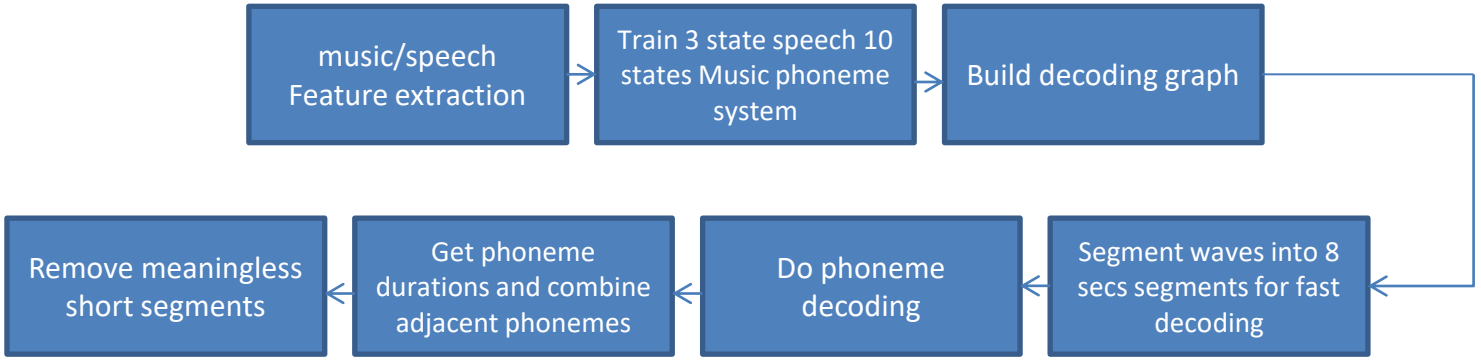


Diarization System Components



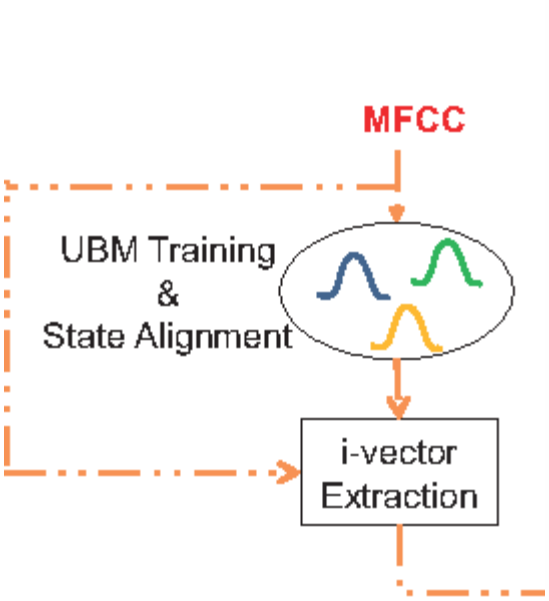
Speech Activity Detection

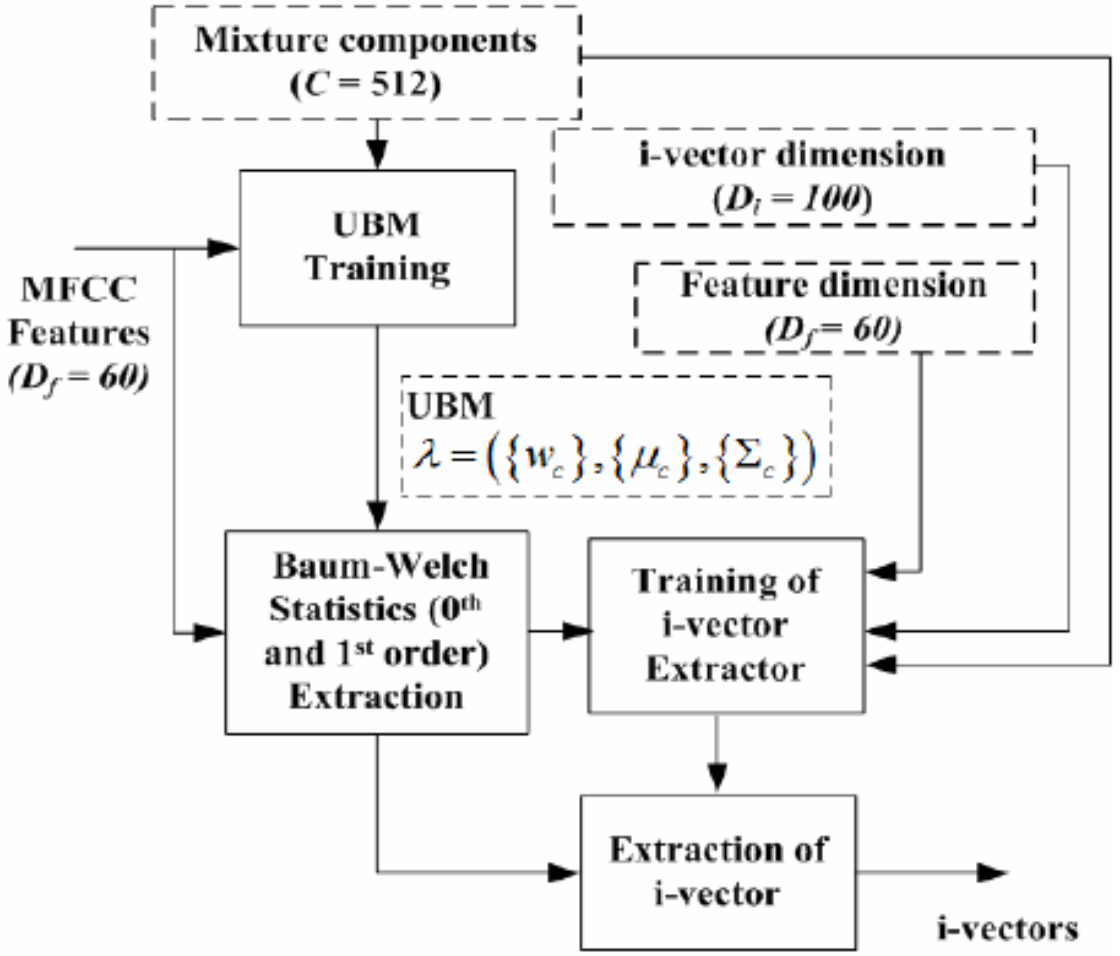
- Phoneme Recognition System for Music/Speech Segmentation



Speaker Clustering

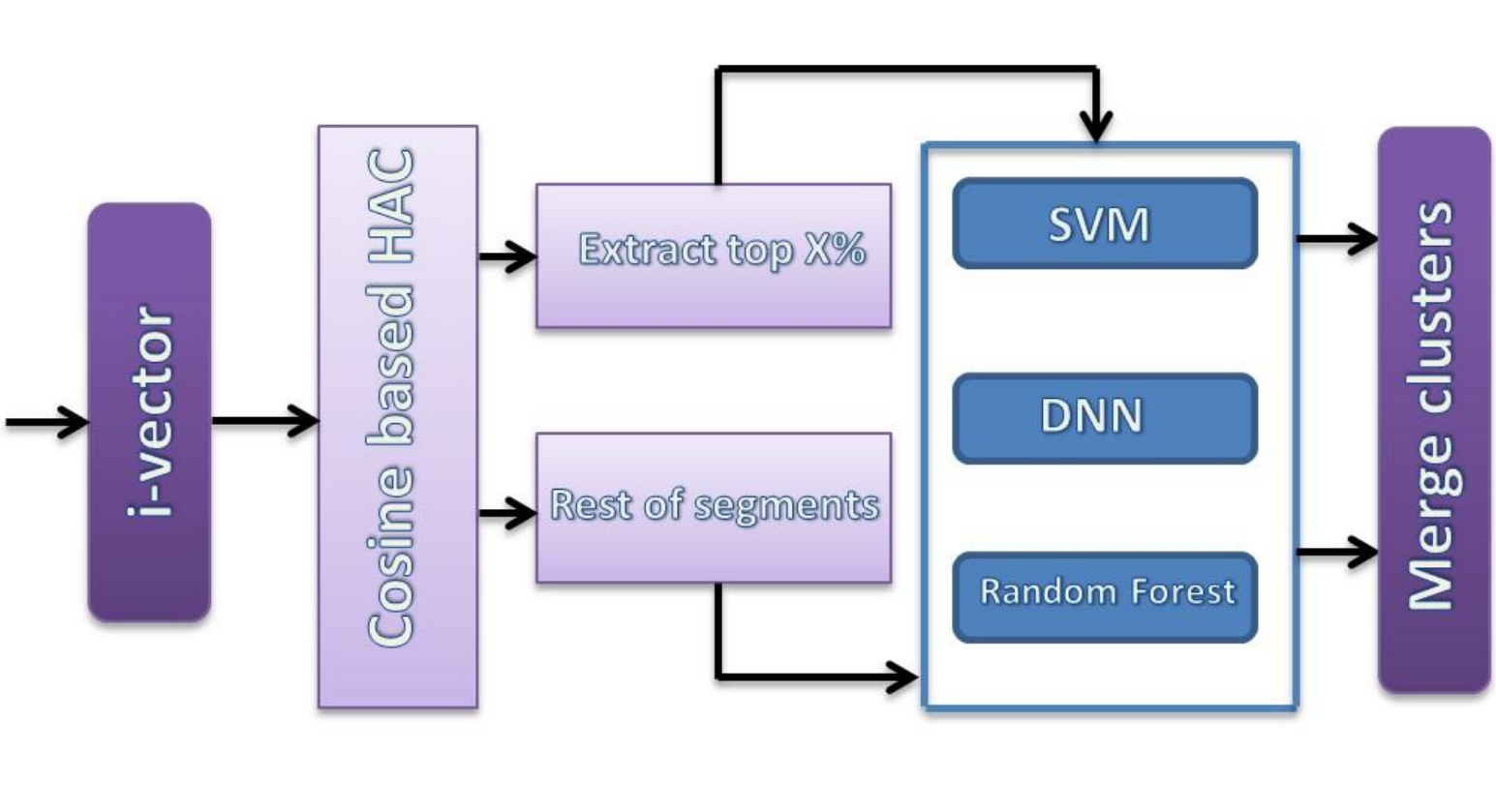
Clustering based on the Speaker Identity vector(I-vectors)





Speaker Diarization System

Proposed Enhanced HAC



Speaker Diarization System

Results

NCLR	SOM +ivectors	GNG + ivectors	HAC ward + ivector	HAC + structured ward	HAC + cosine based
48.8%	28.3%	40%	29.5%	28.5%	27.5%

HAC + cosine based	HAC + cosine + RF	HAC + cosine + DNN	HAC + cosine + SVM
27.5%	27.5%	26.2%	25.8%

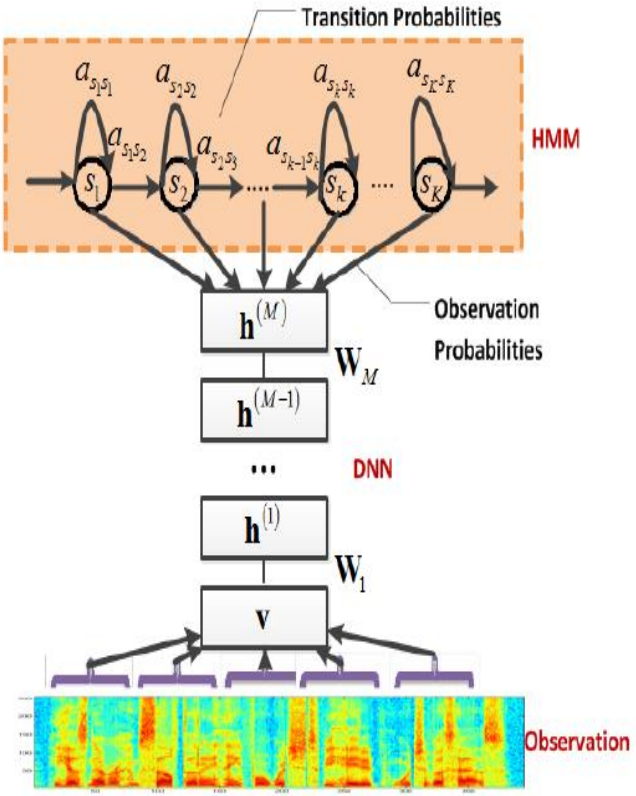
Speaker Diarization system

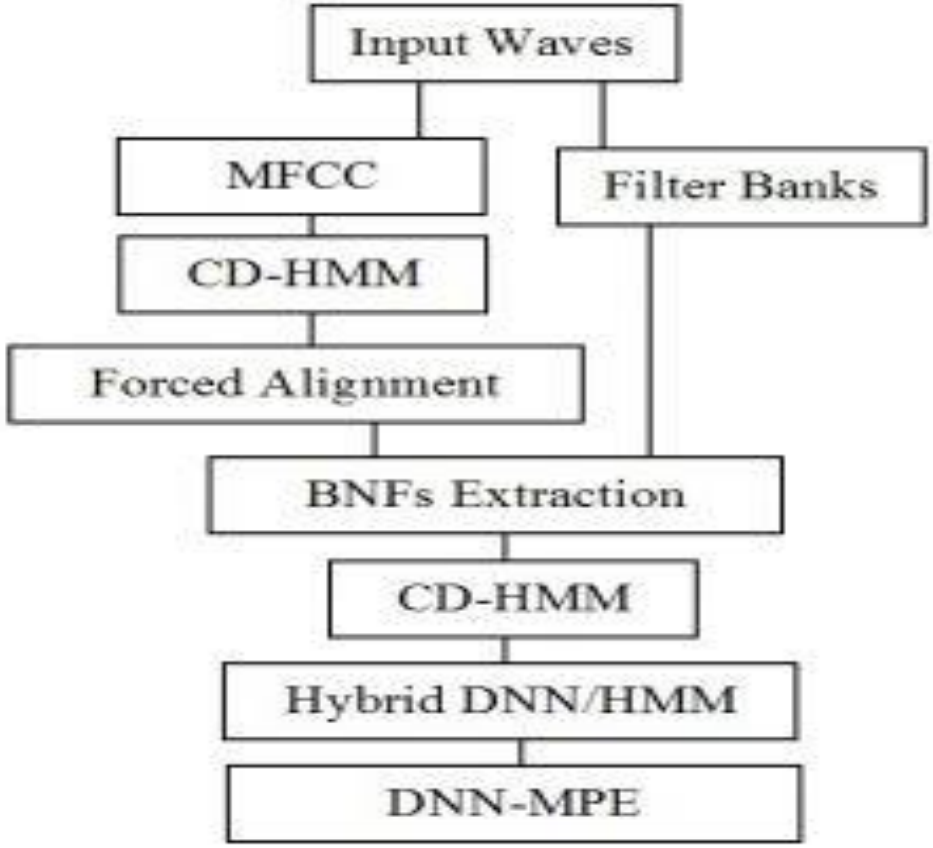
Adjusting Stopping Criterion and Increasing Training Data for UBM

HAC+cosine+ SVM	Adustment performed	Adjustment performed + 500hrs
25.8%	22.7%	21.1%

Automatic Speech Recognition

ASR - first version





ASR - previous version

System	WER %
Tandem (BNFs + fMLLR) (SI)	37.91
Tandem (BNFs + fMLLR) (SD)	36.00
Tandem DNN/HMM-MPE (SD)	27.8

ASR – Current Version

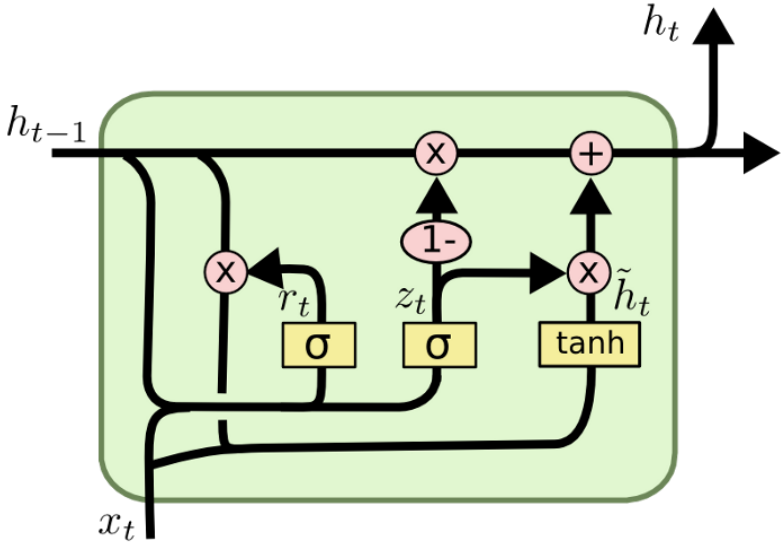
Multiple neural networks with different architectures were trained over the resultant 140-dimensional features.

To discriminate which of the conventional and the Lattice-Free version of the Maximum Mutual Information (LF-MMI) criterion models are better, we utilized the following architectures of Neural Networks:

Conventional Models

A- Time Delay Neural Network (TDNN).
B- Long Short Term Memory neural networks (LSTM).
C- Bidirectional LSTM (BLSTM).

LSTM DNN



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

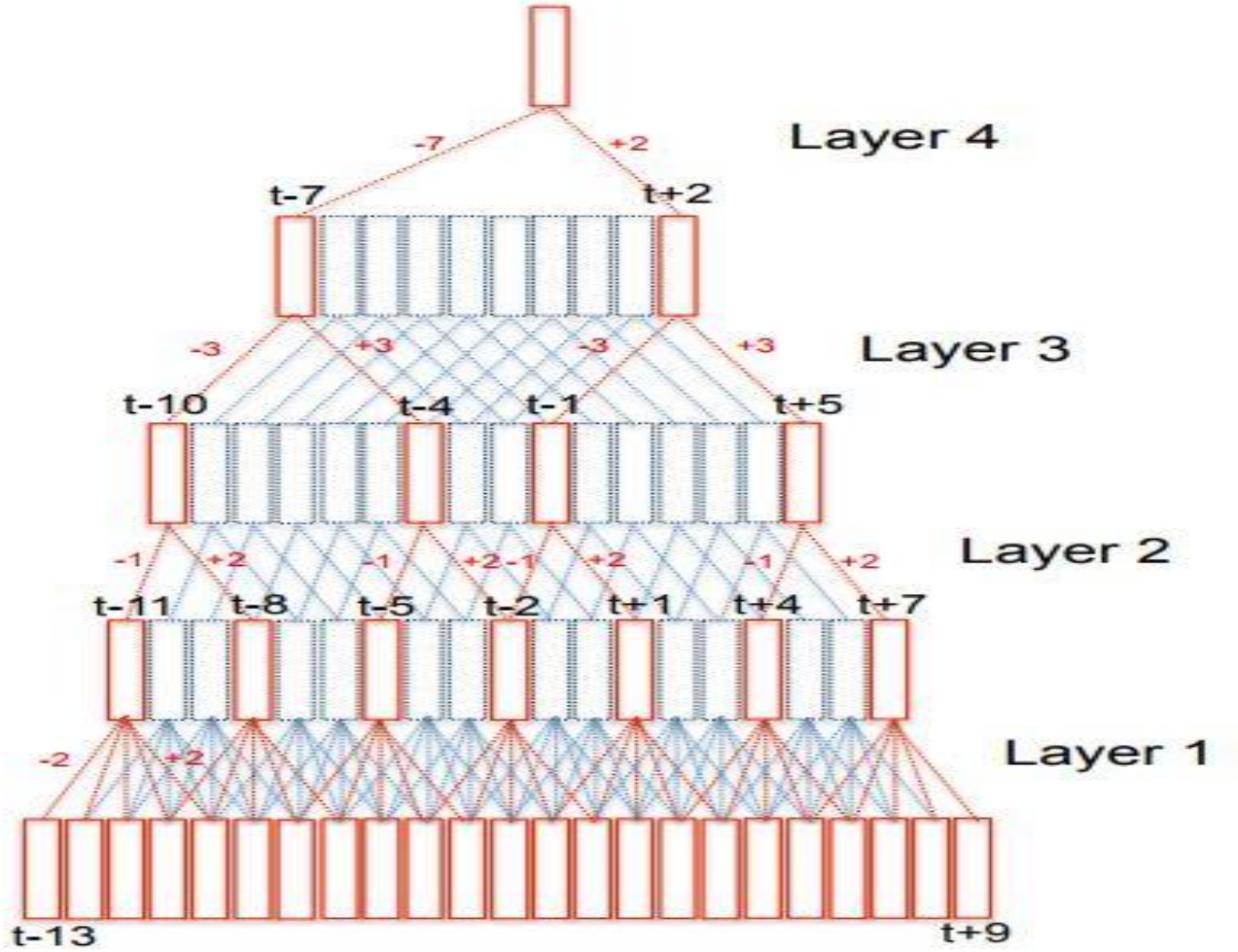
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

ASR – Current Version

TDNN



ASR – Current Version

LF-MMI Models

A- LSTM that got posterior probabilities from a time delay neural network (LF-MMI-TDNN-LSTM).

B- BLSTM that got posterior probabilities from a time delay neural network (LF-MMI-TDNN-BLSTM).

C- LF-MMI-BLSTM

ASR – Current Version

TDNN layers	Splice	Dimension
TDNN-Layer-1	t-2 through t+2	1024
TDNN-Layer-2	t-1 through t+1	1024
TDNN-Layer-3	t-1 through t+1	1024

The output of TDNN neural feeds LSTM or BLSTM .

Type of NN	# cells	Cell dimension
LSTM	3	1024
BLSTM	3	1024

ASR – Current Version

Acoustic modeling experiments

Neural Network (CE)	Features	WER %
TDNN	MFCC-hires	25.24
LSTM	MFCC-hires	24.87
BLSTM	MFCC-hires	25.45
TDNN	Bottle-Neck	22.83
LSTM	Bottle-Neck	22.56
BLSTM	Bottle-Neck	22.98

Neural Network (LF-MMI)	Features	WER %
TDNN-LSTM	Bottle-Neck	20.37
TDNN-BLSTM	Bottle-Neck	21.47
BLSTM	Bottle-Neck	20.34

Speech Recognition System

Dataset

	Dataset	MSA	Coll	Call Centers	Total	Notes
Training Data	***RDI Dataset Mainly***	320 hrs.	400 hrs.	---	720 hrs.	16K sampling rate
	RDI Dataset Call Centers	400 hrs.	500 hrs.	150 hrs.	1050 hrs.	8K sampling rate
	RDI Extended Dataset	140 hrs.	---	---	140 hrs.	Sampled according to model
	Noisy Dataset	125 hrs.	125 hrs.	---	250 hrs.	Randomly selected in both models
	Telephony dataset	100 hrs.	100 hrs.	---	200 hrs.	Randomly selected 8k sampling rate
	Dev Dataset	40 hrs.	50 hrs.	2.5 hrs.	---	Each dataset is separately decoded

Language Model Data

- Text Data
  - Collected 1 Giga words of Modern standard Arabic
    - Sources:
      - Online news papers text
      - TV channels programs scripts
      - The text of the audio data
      - Published books and literacy materials
    - This data set are cleaned from punctuations, English words and some basic normalization for numbers and dates.
  - Collected 0.5 Giga words of Egyptian Arabic
    - Sources:
      - Arabic tweets for one year (50 million words)
      - Famous Egyptian Forums (300 million words
    - Lot of cleaning and normalization: Franco-Arab, large amount of writing mistakes, written emotional effects such the word elongation, reducing the large number of forms for the same word

The Dialectal Arabic

- النهارده النهاردة النهاردا النهرده
  النهردة انهارده انهاردة إنهارده
  إنهاردة إنهاردا

- بأى بئى بئا بئه بقى
  بقا بقه

- دلوقتي دلوأتي دلوئتي دلوءتي

ASR – Current Version

Systems Fusion

First Stage Combination	WER %
TDNN-LSTM +TDNN-BLSTM (C1)	18.94
BLSTM+TDNN-BLSTM (C2)	18.94
TDNN-LSTM+BLSTM (C3)	18.41

Second Stage Combination	WER %
C1+C2 (output 1)	18.2
C1+C3 (output 3)	18.41
C2+C3 (output 2)	18.2

ASR – Current Version

Language modeling experiments:

Using the corpus provided by the organizers and transcriptions of all the used training data, we trained a trigram language model using the SRILM toolkit.

This language model was used for scoring lattices.

In addition, we used another language model for rescoring that was trained using the Faster-RNN-LM toolkit, using 300 hidden units, 500 classes, 300K most frequent words, and order equals to 4.

ASR – Current Version

Language modeling experiments:

Experiment	Data	WER %
Before RNN-LM Rescoring	DEV.	18.2
	TEST	15.9
After RNN-LM Rescoring	DEV.	17.84
	TEST	15.5

MGB2 Challenge Results

	MGB2 WER
2016-best system	14.7
Aalto	**13.2**
NDSC-THUEE	14.5
JHU	16.0
MIT	17.5
BUT	24.7
RDI-CU	16.0

ASR – Current Version

update after competition

- The ASR system was trained using the full MGB-2 data, 1,200 hours of audio.
- This data was augmented by applying speed perturbation, increasing the number of training frames by a factor of 3. (we used only around 3500 hours).

- LF-TDNN-LSTM & BLSTM have been trained.

New Results Dev. Set	16.78%

ASR – Current Version

Conclusion

Version	WER % (Dev)
Previous Version	27.8 %
Current Version	16.78 %

Video Track Efforts

Video Summarization and Image Annotation

Video Summarization

- Video Indexing
- Shot Segmentation and Advertisement

Video Images Annotation

- Objects Description
- Person Faces Identification
- Shot Segmentation and Advertisement

Closed Caption OCR (Done)

- Closed Caption Extraction
- Arabic OCR

Caption Text Extraction

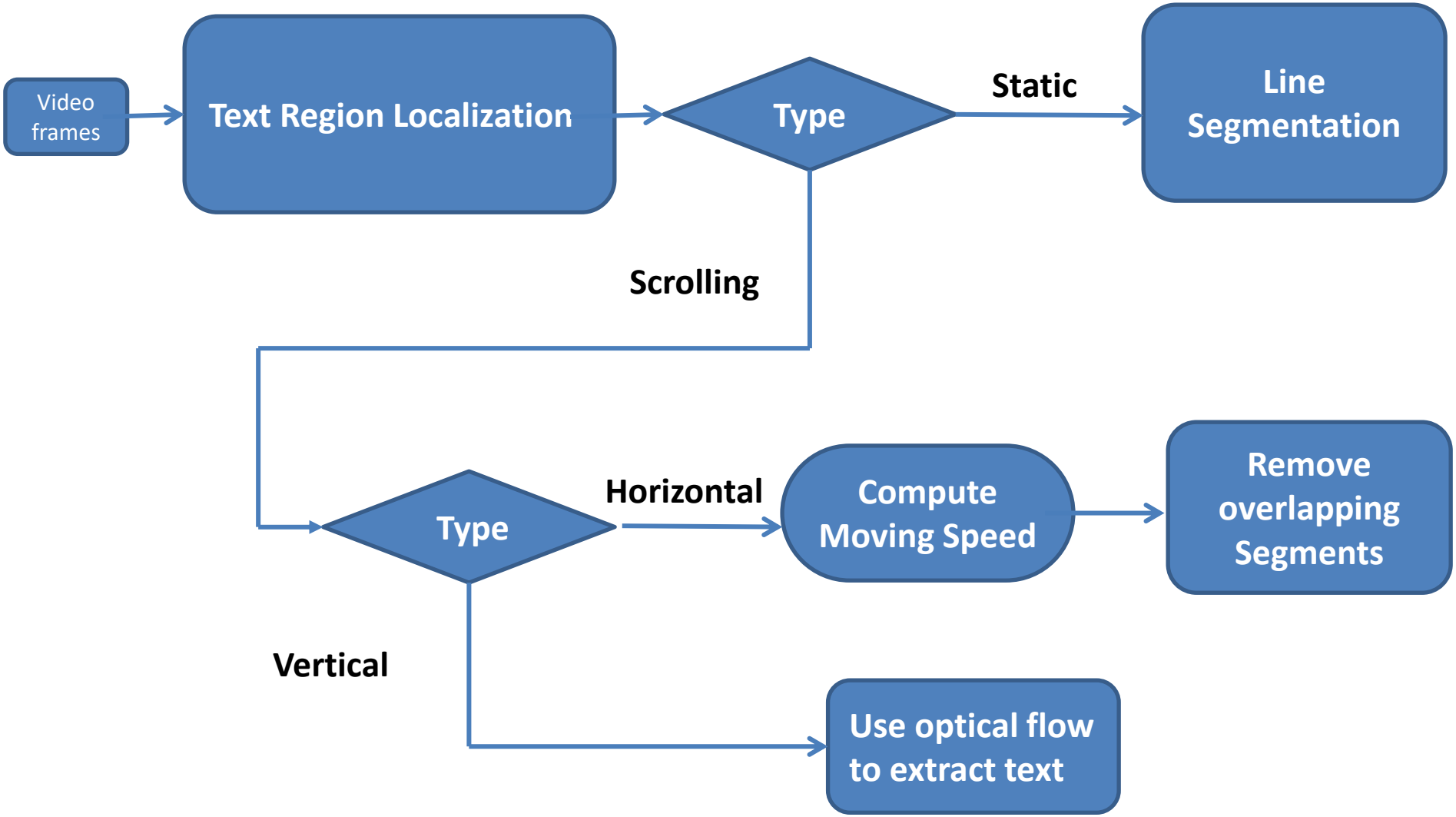
- Location in the lower third of the video frame.
- It can be static or scrolling
- The scrolling can be horizontally or vertical.

Caption Text Extraction

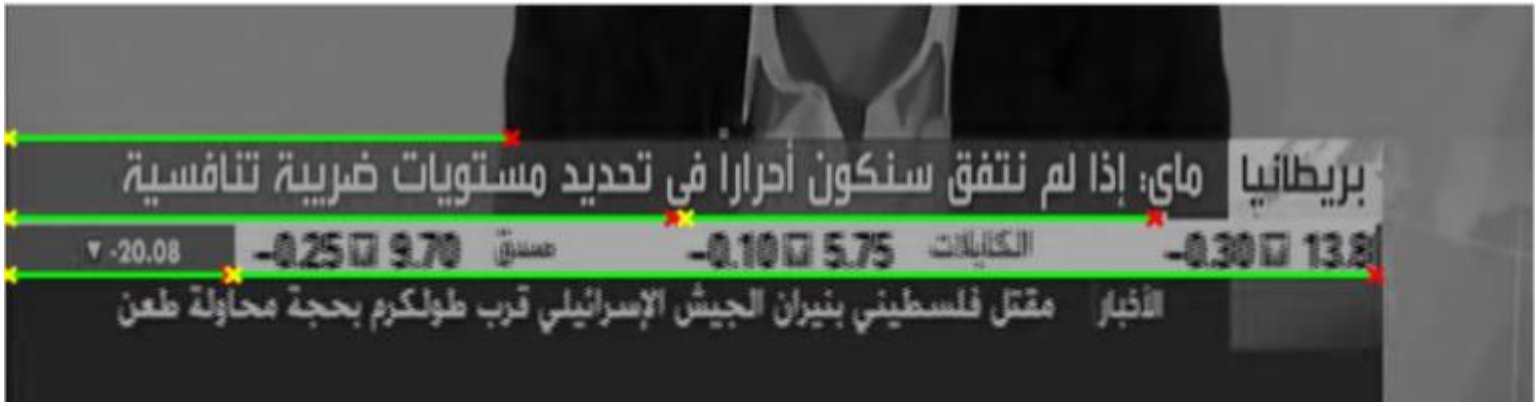
Method

- Text Region Line Segmentation
  - Based on Hough Transform
- Text Region Type Detection
  - Based on Optical flow
- Sentence Extraction from Horizontal Scrolling
  - Calculate frame speed
- Sentence Extraction from Static and Vertical Scrolling

Caption Text Extraction



Caption Text Extraction



Caption Text Extraction

Before matching

After matching

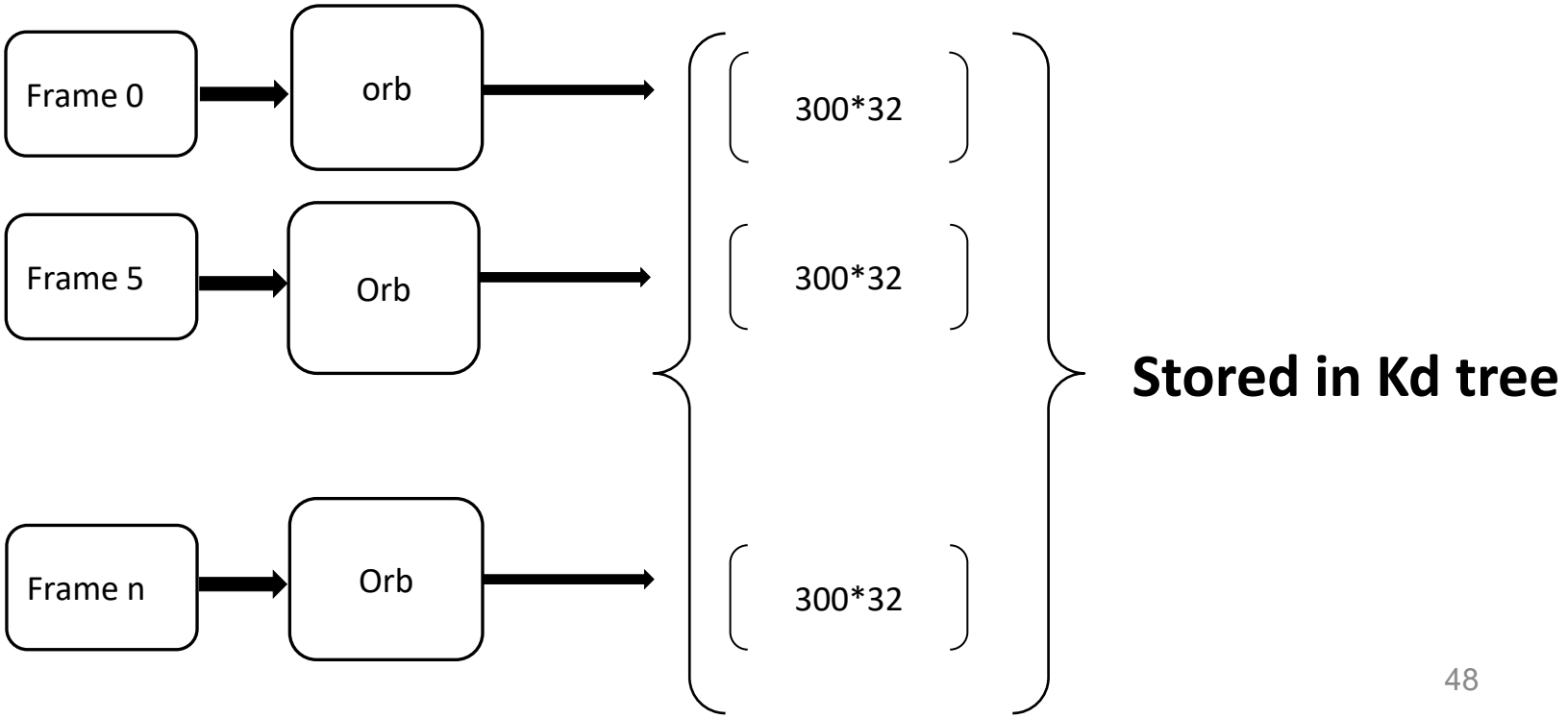
Caption Text Extraction

TV Channel	Static accuracy	Vertical accuracy	Horizontal accuracy	Total accuracy
Al-Arabiya	4/4=100%	---	27/27=100%	100%
BBC	5/5=100%	---	24/24=100%	100%
Aljazeera Live	4/4=100%	60/60=100%	18/18=100%	100%
Aljazeera	3/4=75%	----	18/20=90%	87.5%
Nile	1/1=100%	----	33/33=100%	100%
CNBC	6/6=100%	----	43/43=100%	100%
CBC extra	6/6=100%	----	24/24=100%	100%
SKY	2/2=100%	----	20/20=100%	100%
Alghad	1/2=50%	----	20/24=83.3%	80.7%
Overall	91.67%	100%	97.03%	96.47%

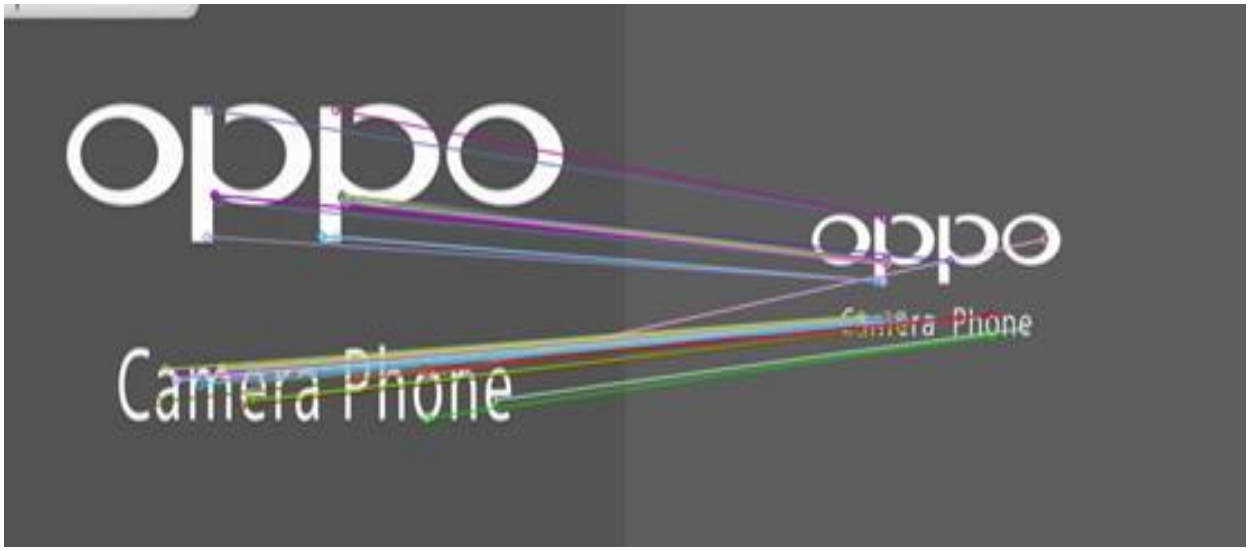
Video Summarization

Video Indexing

- Extract Oriented fast and Rotated Brief (ORB) features
  - 300 features per frame, each of length 32.



Video Indexing



Video Indexing

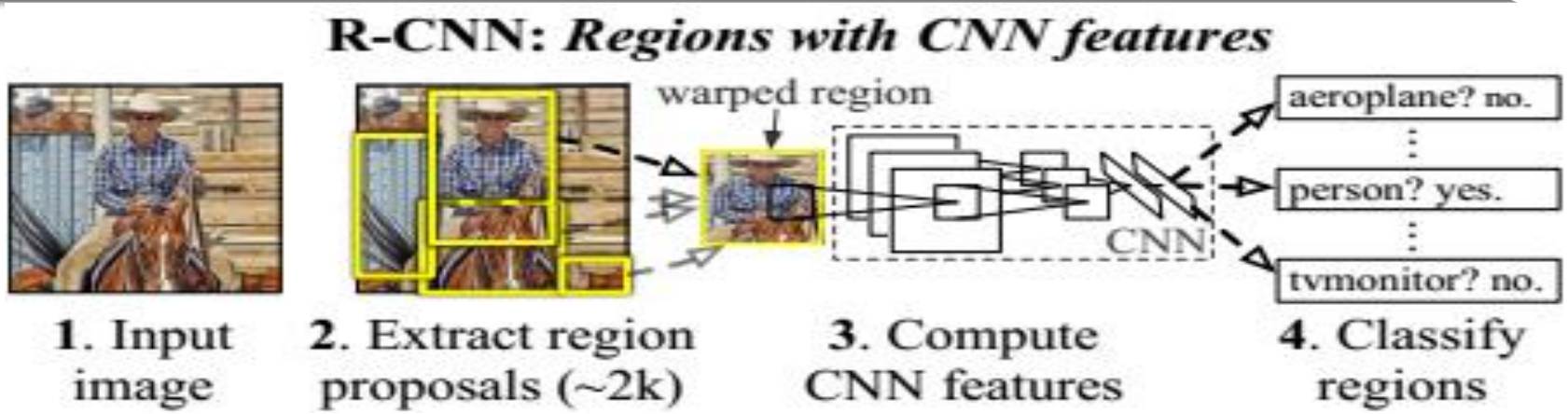
Hour	Indexing time	Searching time 60 adds	Accuracy
1	32 min	15 sec	100%
2	28 min	14 sec	100%
3	31 min	13.4 sec	100%
4	23 min	14 sec	100%
5	27 min	13 sec	100%
6	34 min	12.4 sec	100%
7	32 min	12.6 sec	100%
8	27 min	13 sec	100%

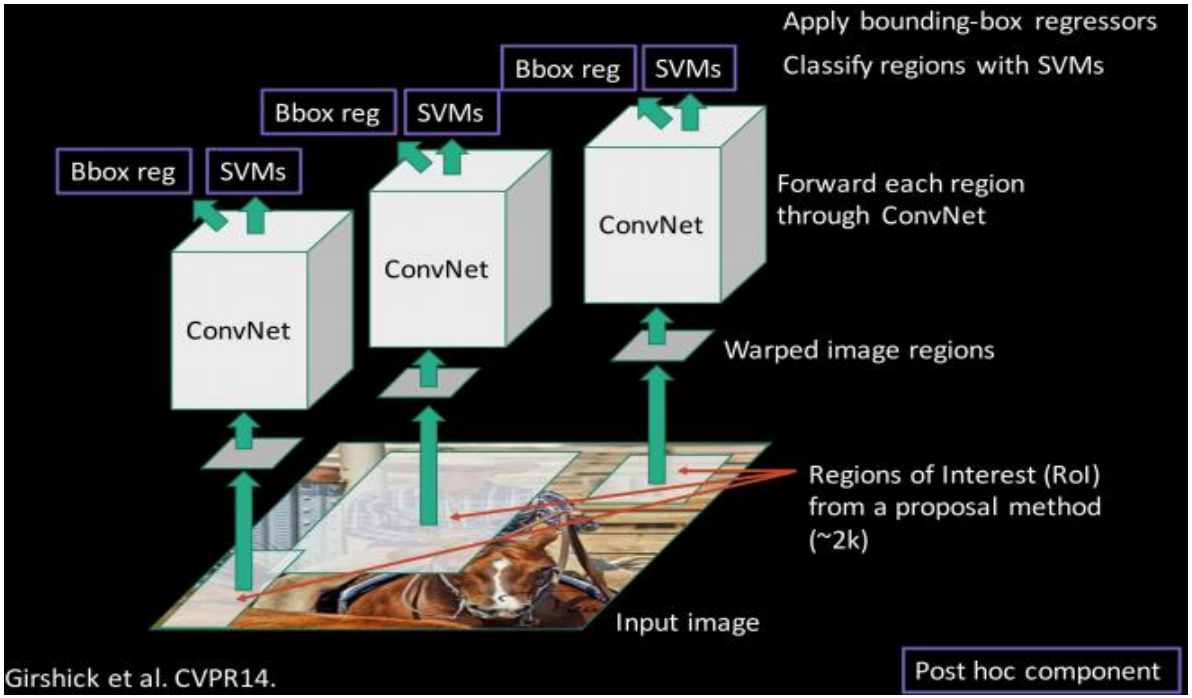
Testing Environment: Intel core i7 processor with 8 GB RAM

Video Images Annotation

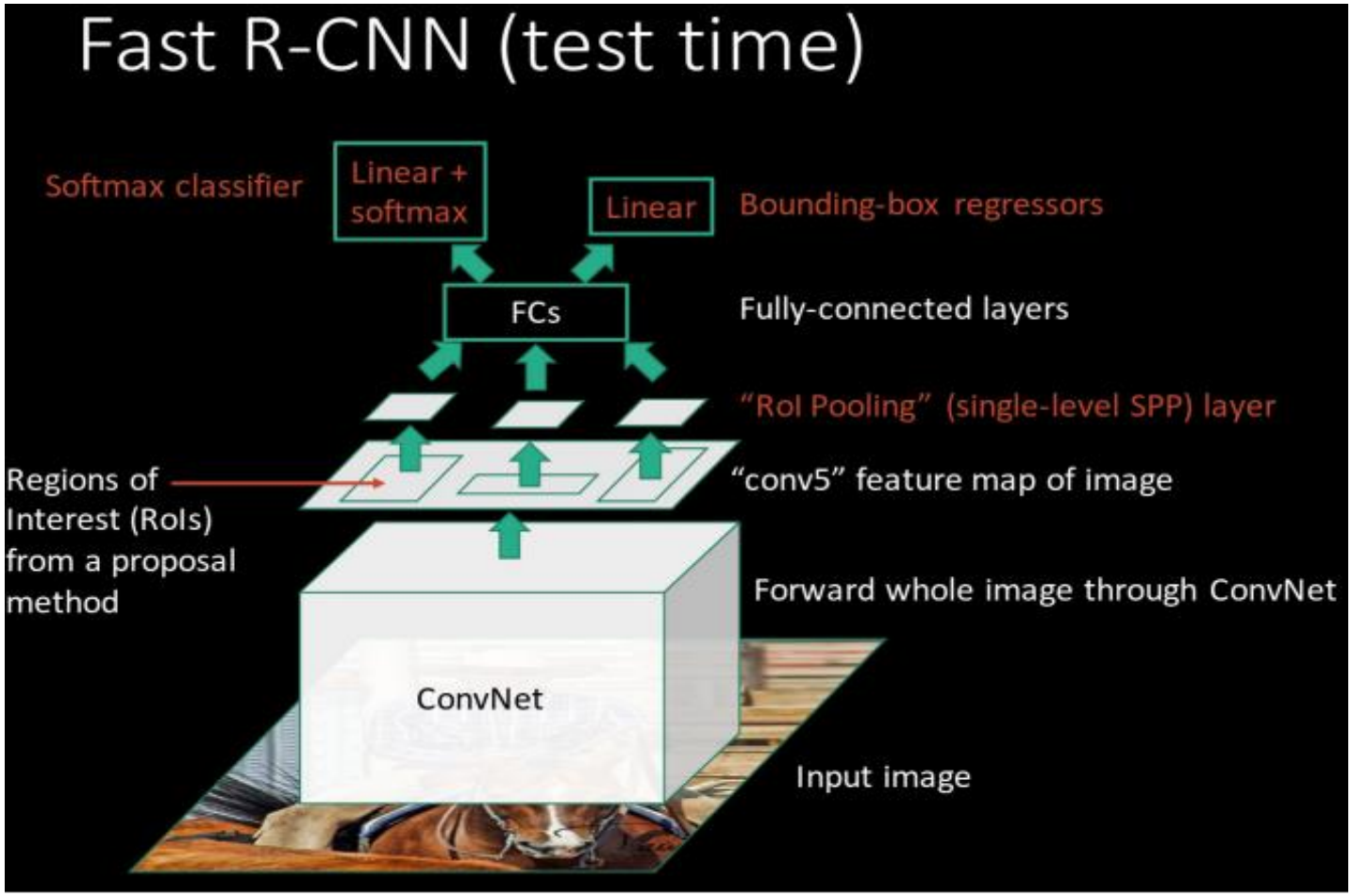
- Used pretrained CNN models on the imagnet dataset
- Fine-tuned on extracted frames from our 2000 hrs video data
- The best models are
  - RCNN
  - Fast RCNN
  - Faster RCNN
  - YOLO (You Look Only Ones)

RCNN

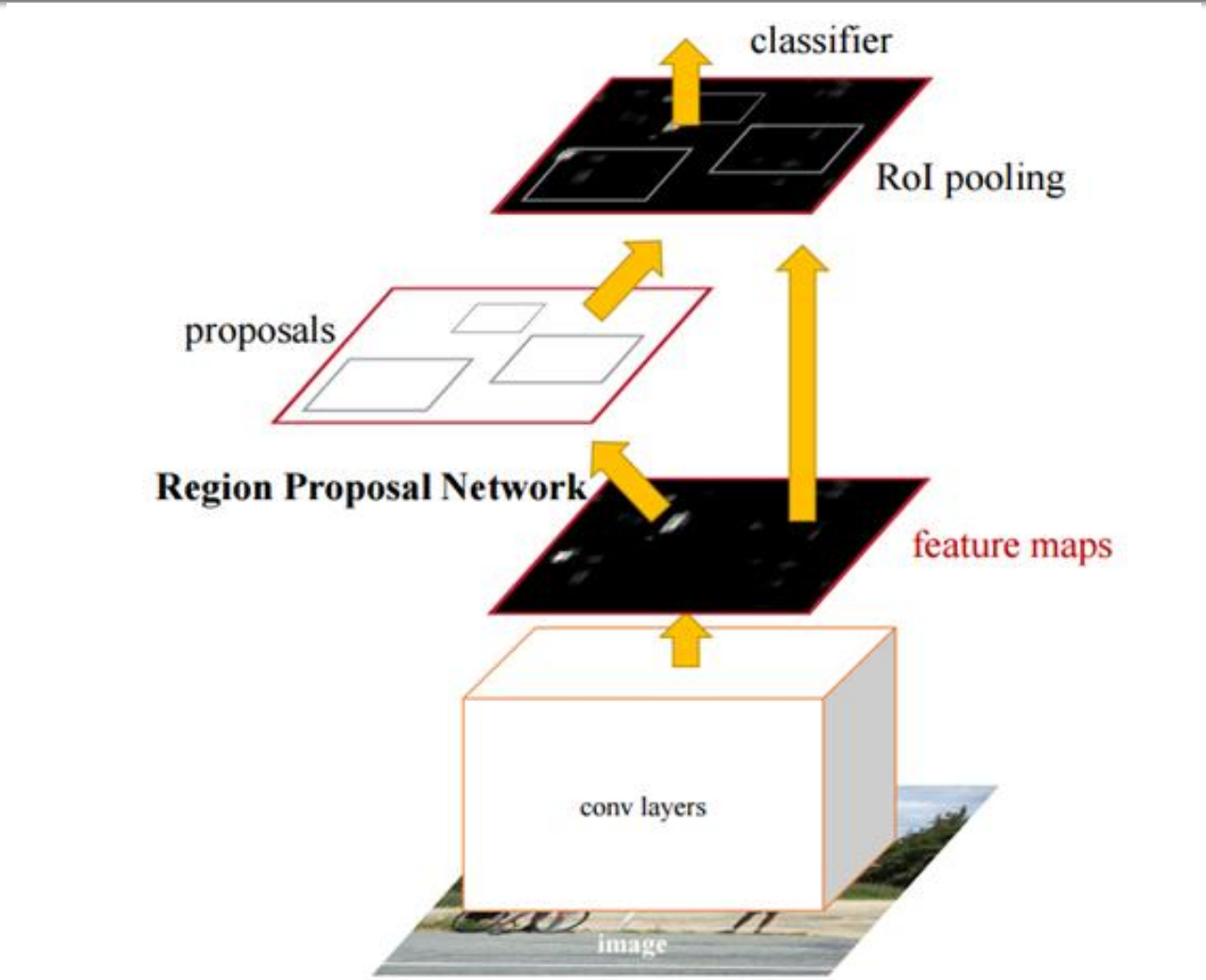

R-CNN: Regions with CNN features
warped region
CNN
aeroplane? no.
person? yes.
tvmonitor? no.
1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions


Apply bounding-box regressors
Classify regions with SVMs
Bbox reg
SVMs
ConvNet
Forward each region through ConvNet
Warped image regions
Regions of Interest (RoI) from a proposal method (~2k)
Input image
Girshick et al. CVPR14.
Post hoc component

Fast RCNN


Fast R-CNN (test time)
Softmax classifier
Linear + softmax
Linear
Bounding-box regressors
FCs
Fully-connected layers
"RoI Pooling" (single-level SPP) layer
Regions of Interest (RoIs) from a proposal method
"conv5" feature map of image
Forward whole image through ConvNet
ConvNet
Input image

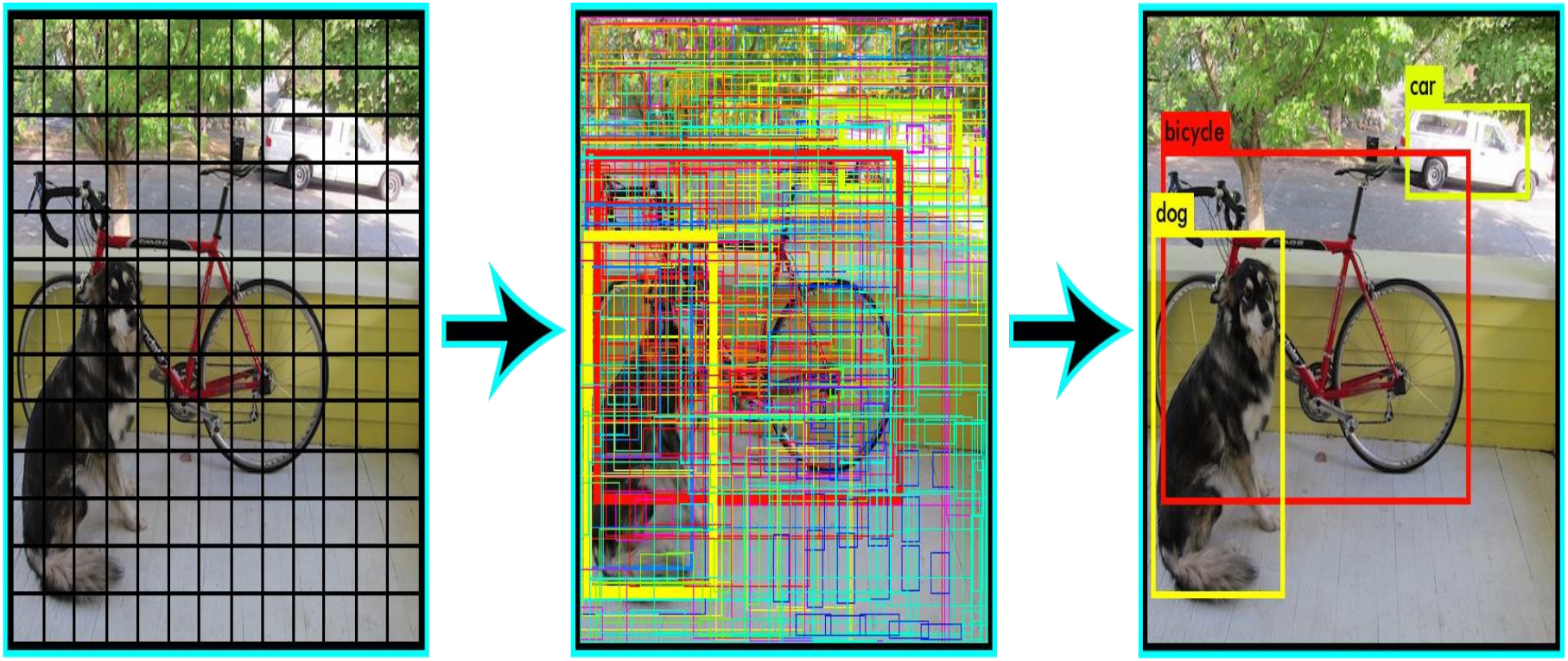
Faster RCNN



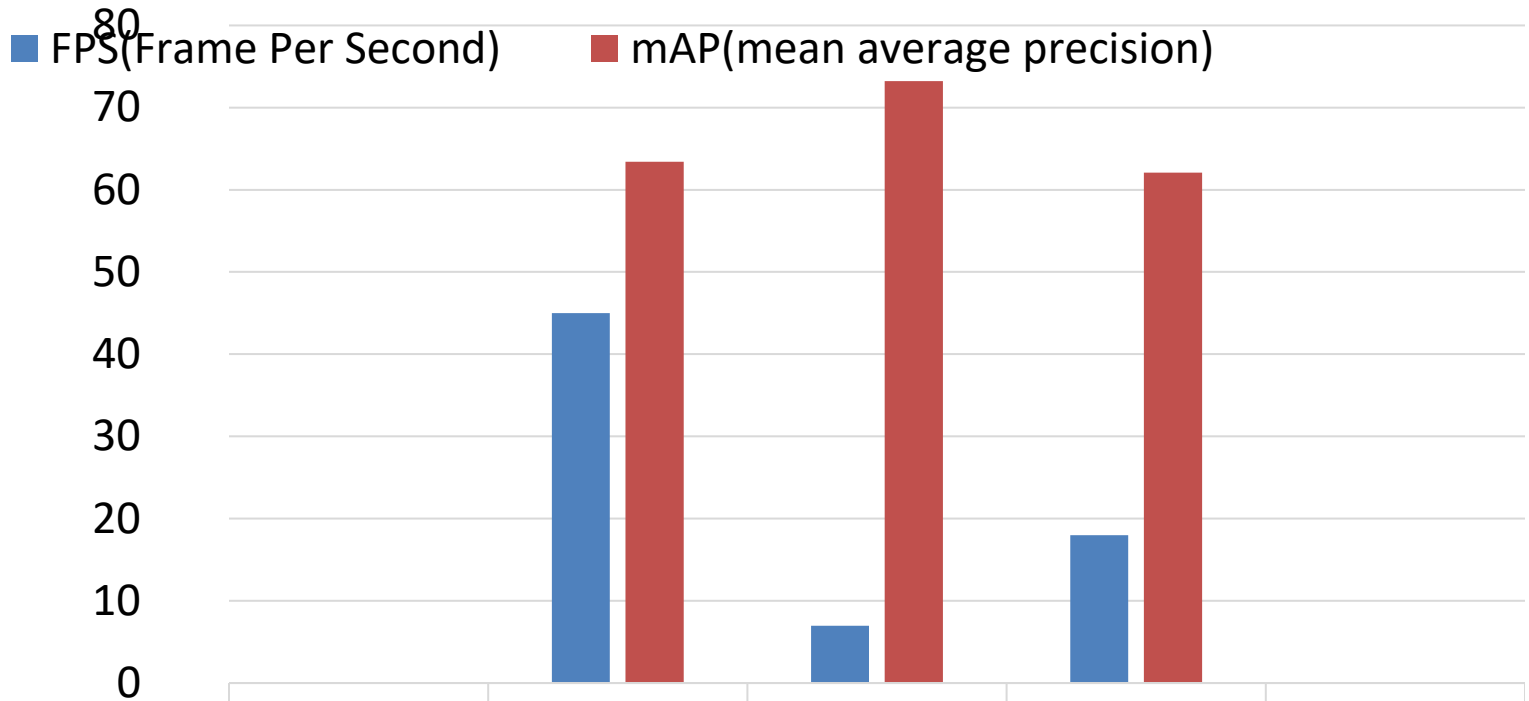
Faster R-CNN workflow

Yolo CNN

It is 100 times faster than Fast RCNN

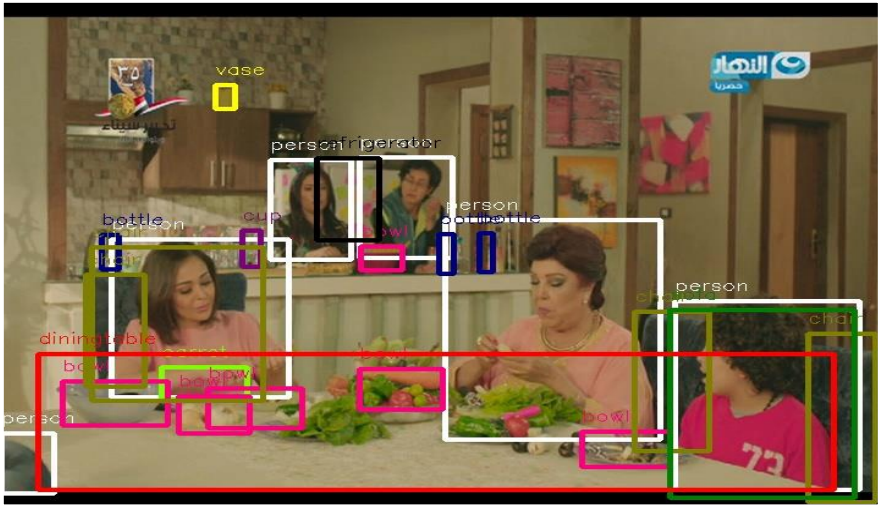


Yolo vs. Faster R-CNNs

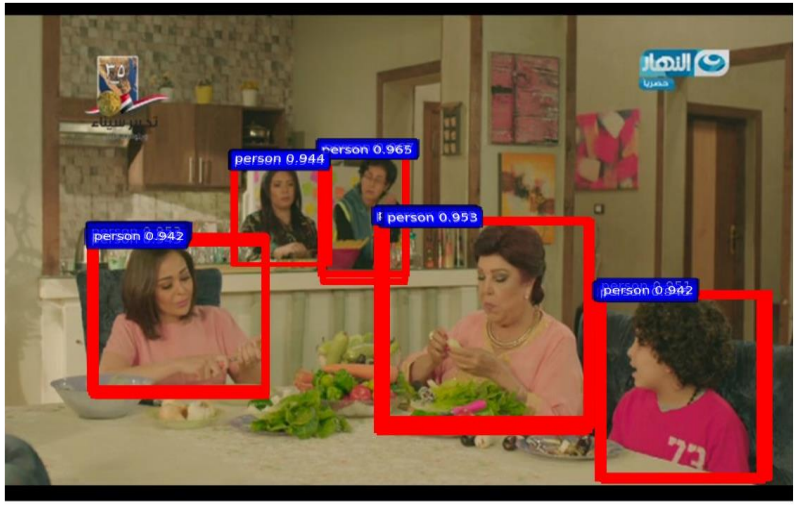


		YOLO	Faster R-CNN with VGG-16	Faster R-CNN with ZFNet	
FPS(Frame Per Second)		45	7	18	
mAP(mean average precision)		63.4	73.2	62.1	

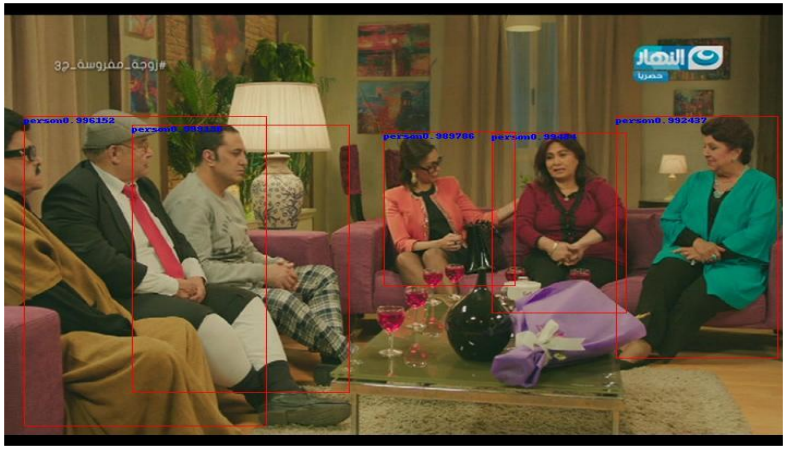
Sample Object Detection For Our Data



YOLO V2



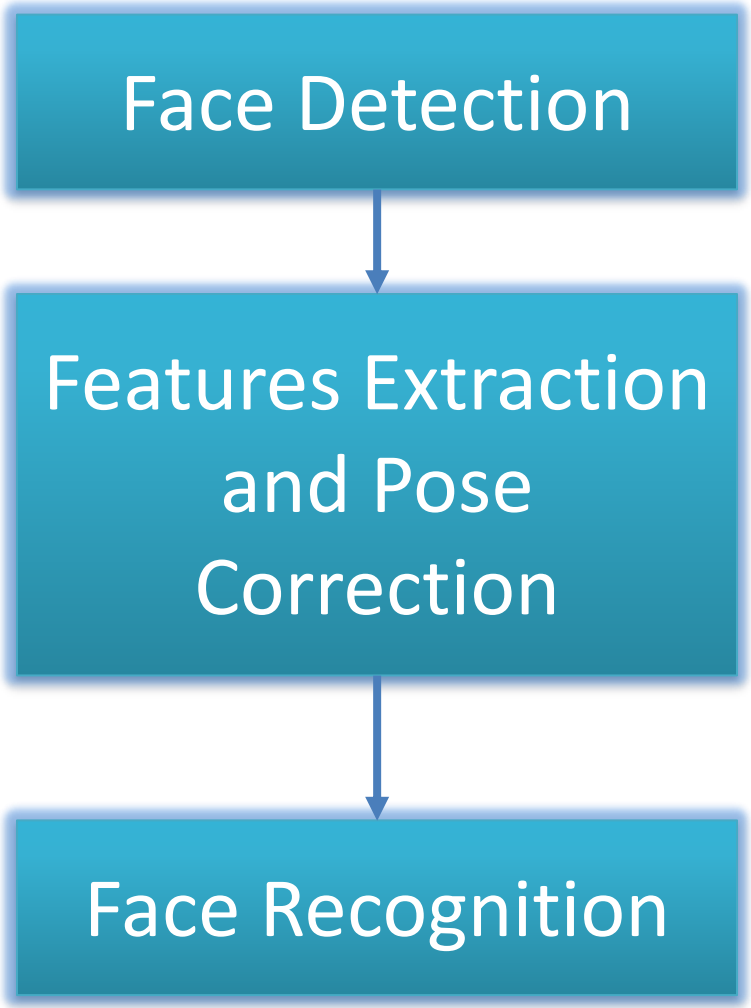
Faster R-CNN



SSD

Face Recognition

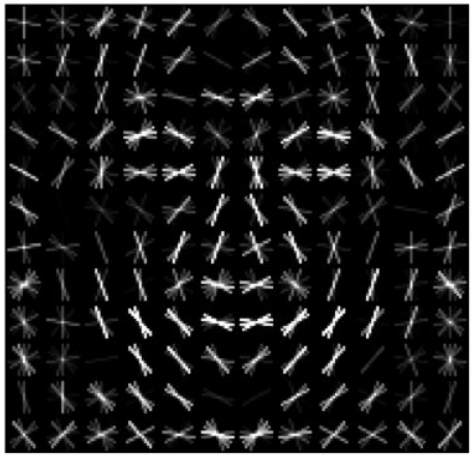
Face Recognition



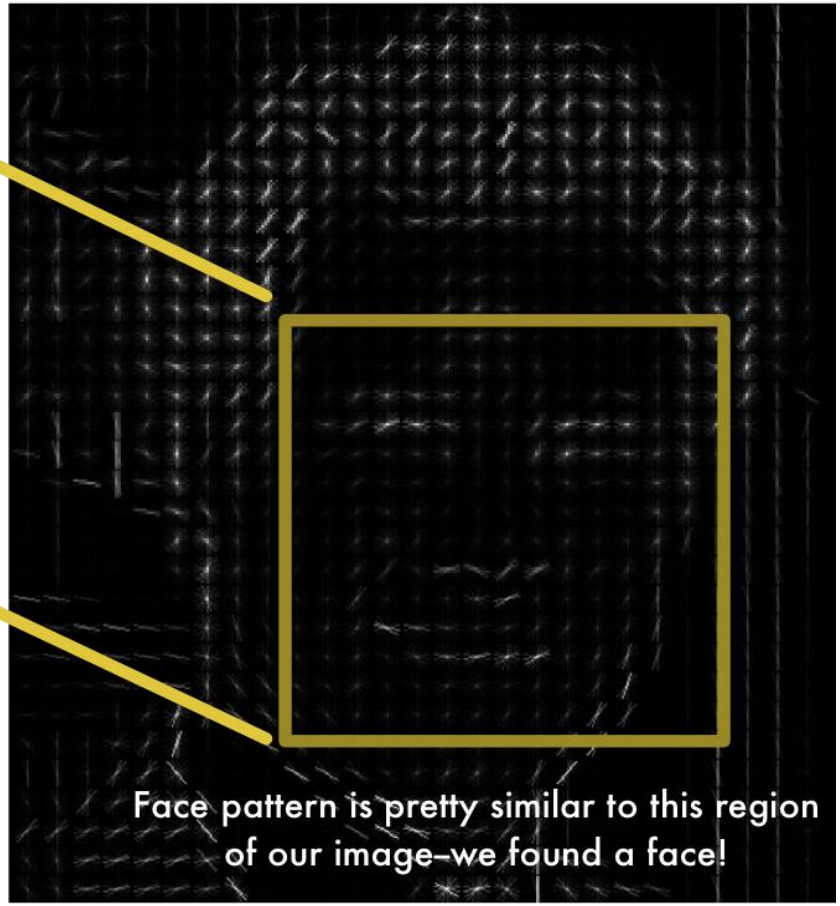
Face Detection

Histogram of Oriented Gradients (HOG)

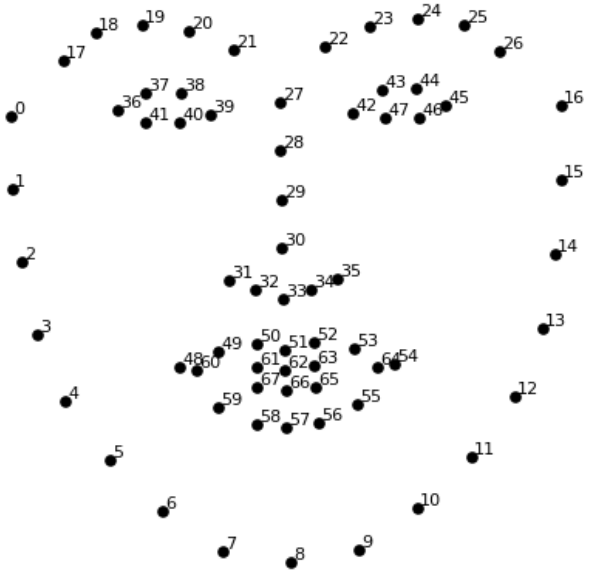
HOG face pattern generated from lots of face images

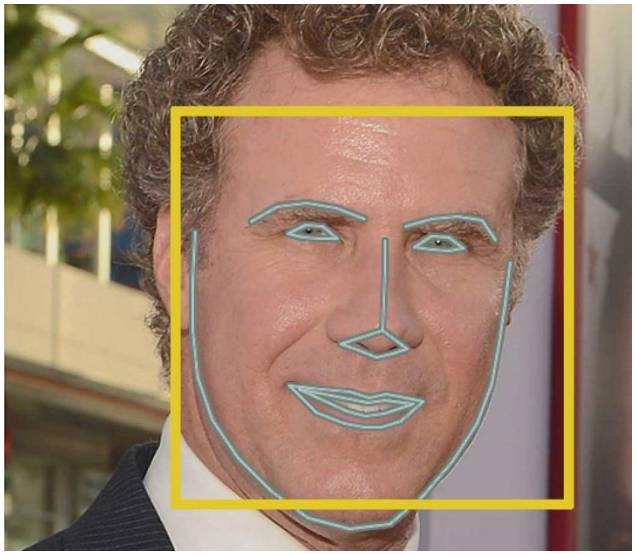


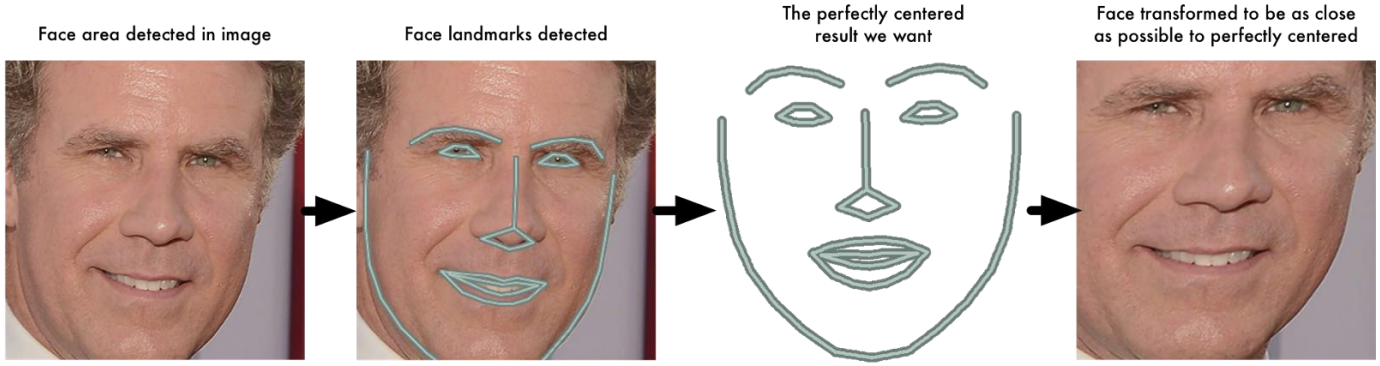
HOG version of our image



Posing and projecting faces process



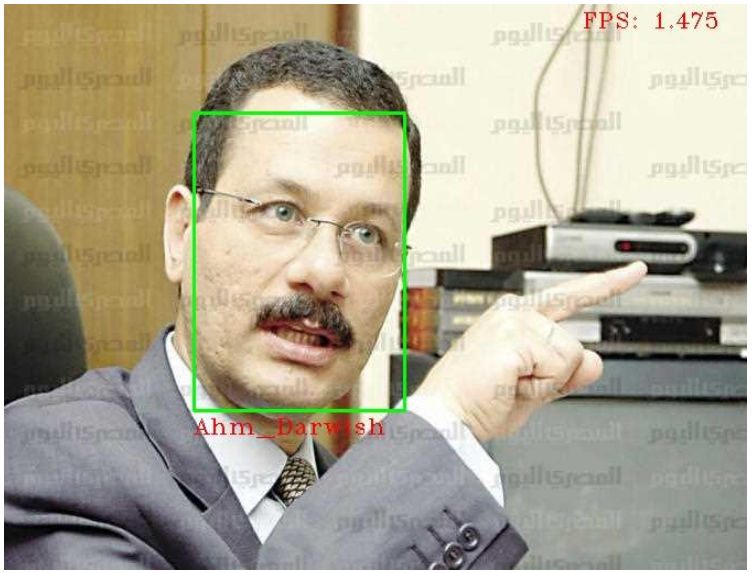


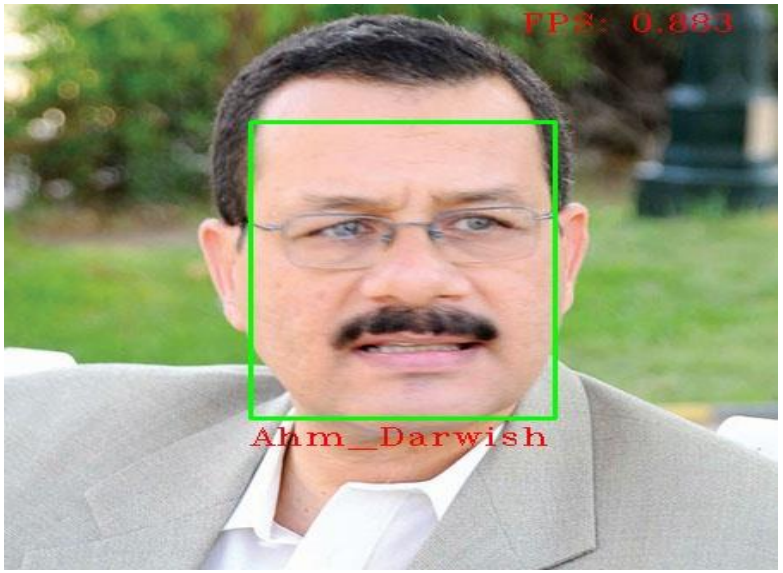


Person Faces Identification

- Recognition model
  - FaceNet fine tuned for our 200 persons
- Data
  - Collect 20,000 photos for 100 persons.
  - 200 photos for each persons (100 from video and 100 from web)
- Results
  - Overall Accuracy = 99.6% dropped to 98.2% with unknown faces
  - F1 measure = 0.996

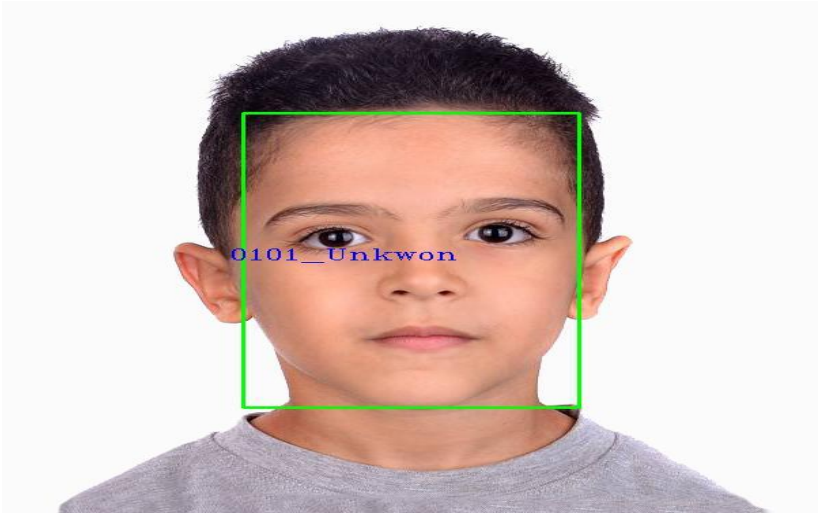
Person Faces Identification
Sample Results

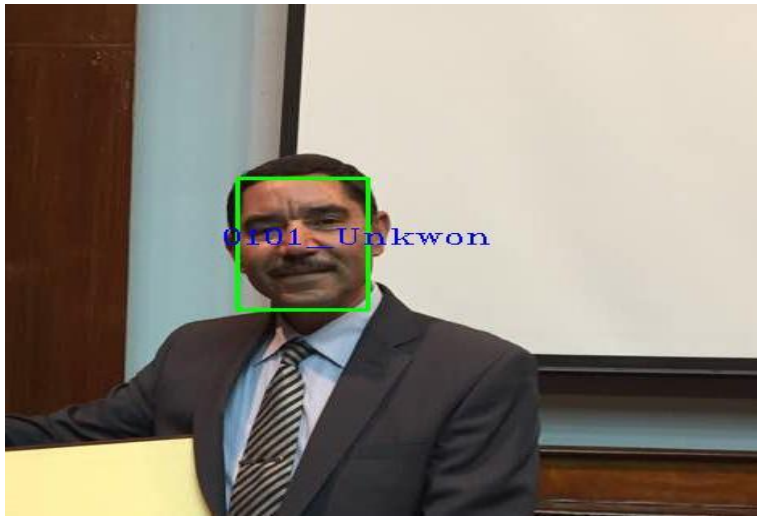




Person Faces Identification
Sample Results

Sample for Unknown Faces

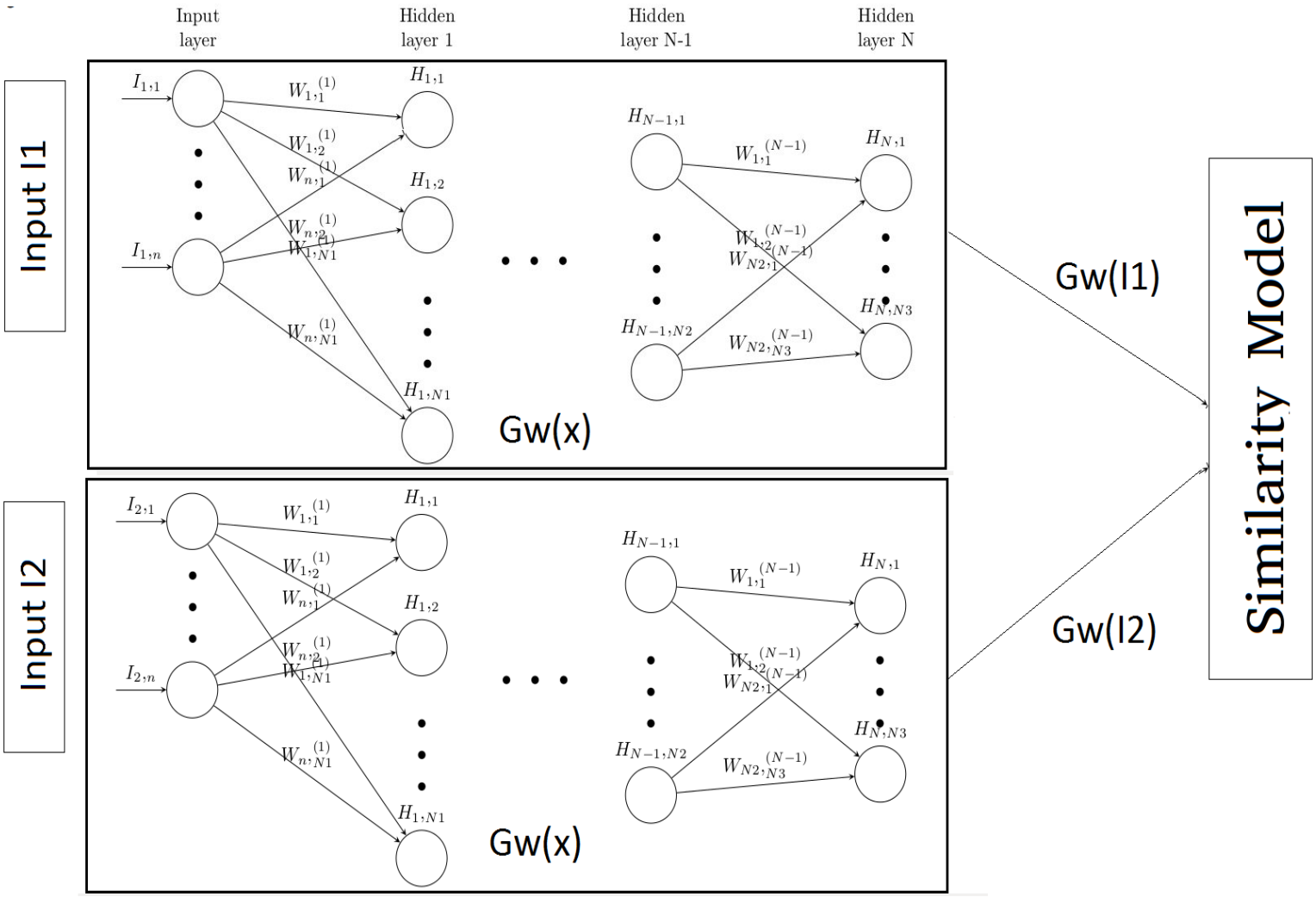




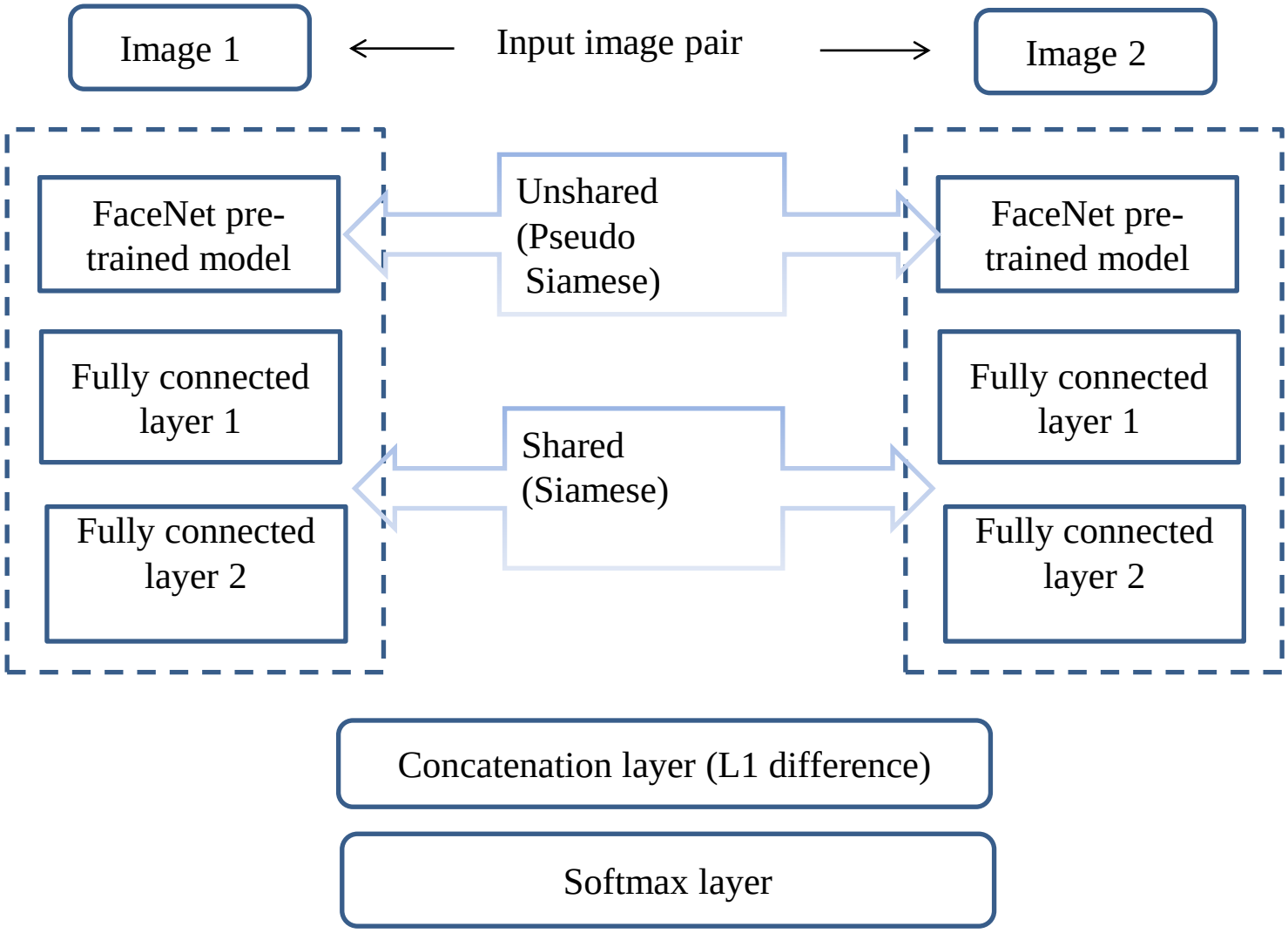
Face Verifier and Clustering Approach

- Motivation:
  - We cannot build an accurate classifier for thousands or millions of faces.
  - Several hundreds of images of the face to recognize should be available.
  - Not all the faces appear in the media are of interested at the moment they appear, but it could be latter when the interest starts to build up
- Approach:
  - Build Face Verifier to decide in two faces are similar or not.
  - Use this face verifier to cluster faces together as a method of indexing
  - Use it again to search for the cluster of a new face.

Siamese Network



Hybrid Siamese Network Architecture



Cluster 1


Sherif_amer_F (1)


Sherif_amer_F (1)_00156


Sherif_amer_F (1)_00157


Sherif_amer_F (1)_00158


Sherif_amer_F (1)_00159


Sherif_amer_F (1)_00160


Sherif_amer_F (1)_00161


Sherif_amer_F (1)_00162

Cluster 2


img_00324


img_00324_00325


img_00324_00333


img_00324_00368


img_00324_00369


img_00324_00371


img_00324_00380

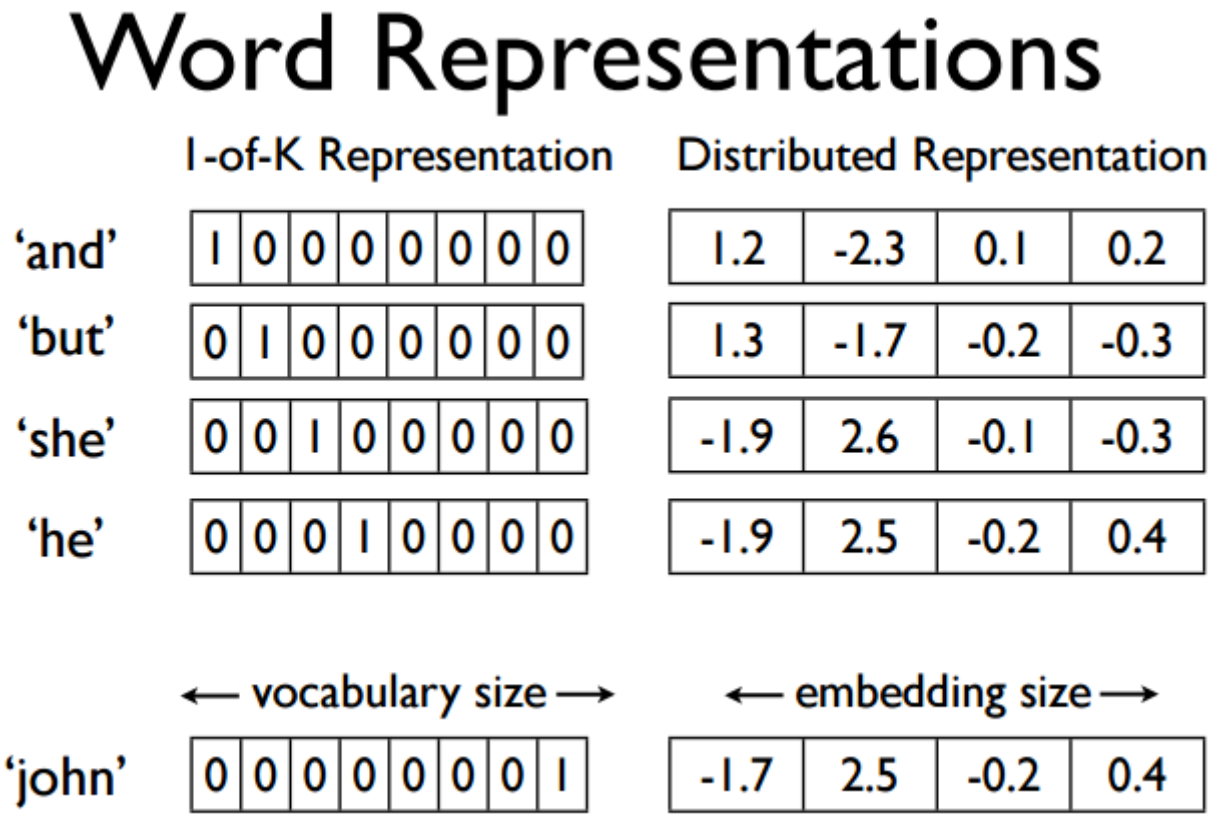

img_00324_00389

Semantic Representation

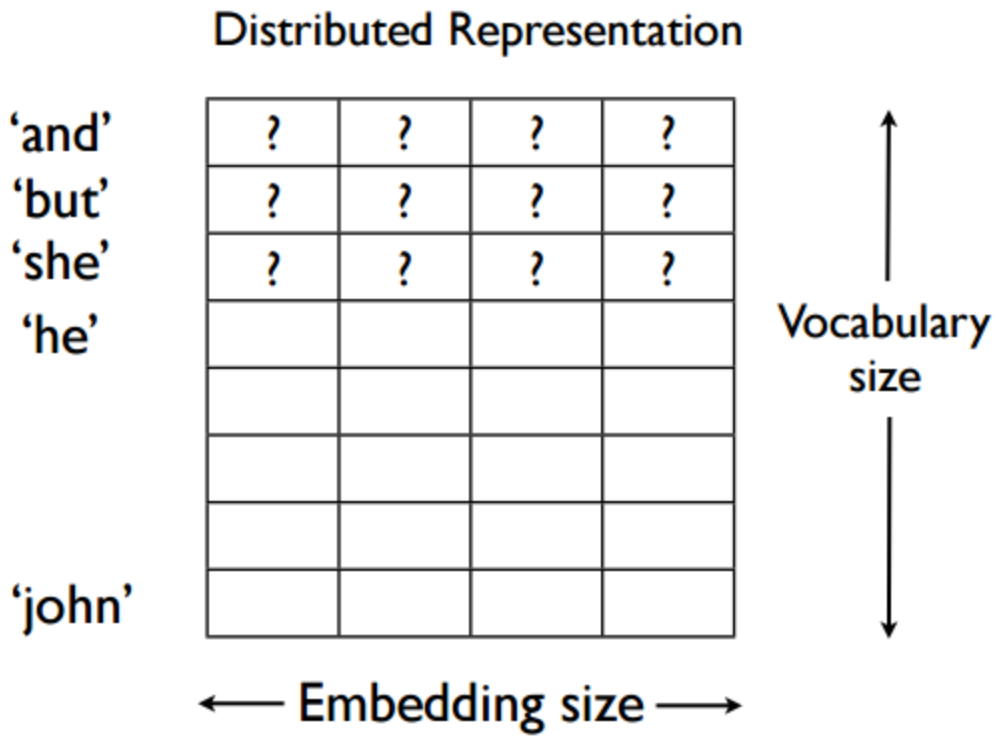
Distributed vector representations of words

- Each word is encoded as a vector of floats
  - vec queen= (0.2, -0.3, .7, 0, … , .3)
  - vec woman = (0.1, -0.2, .6, 0.1, … , .2)
- length of the vectors = dimension of the word representation
- key concept of word2vec: words with similar vectors have a similar meaning (context)

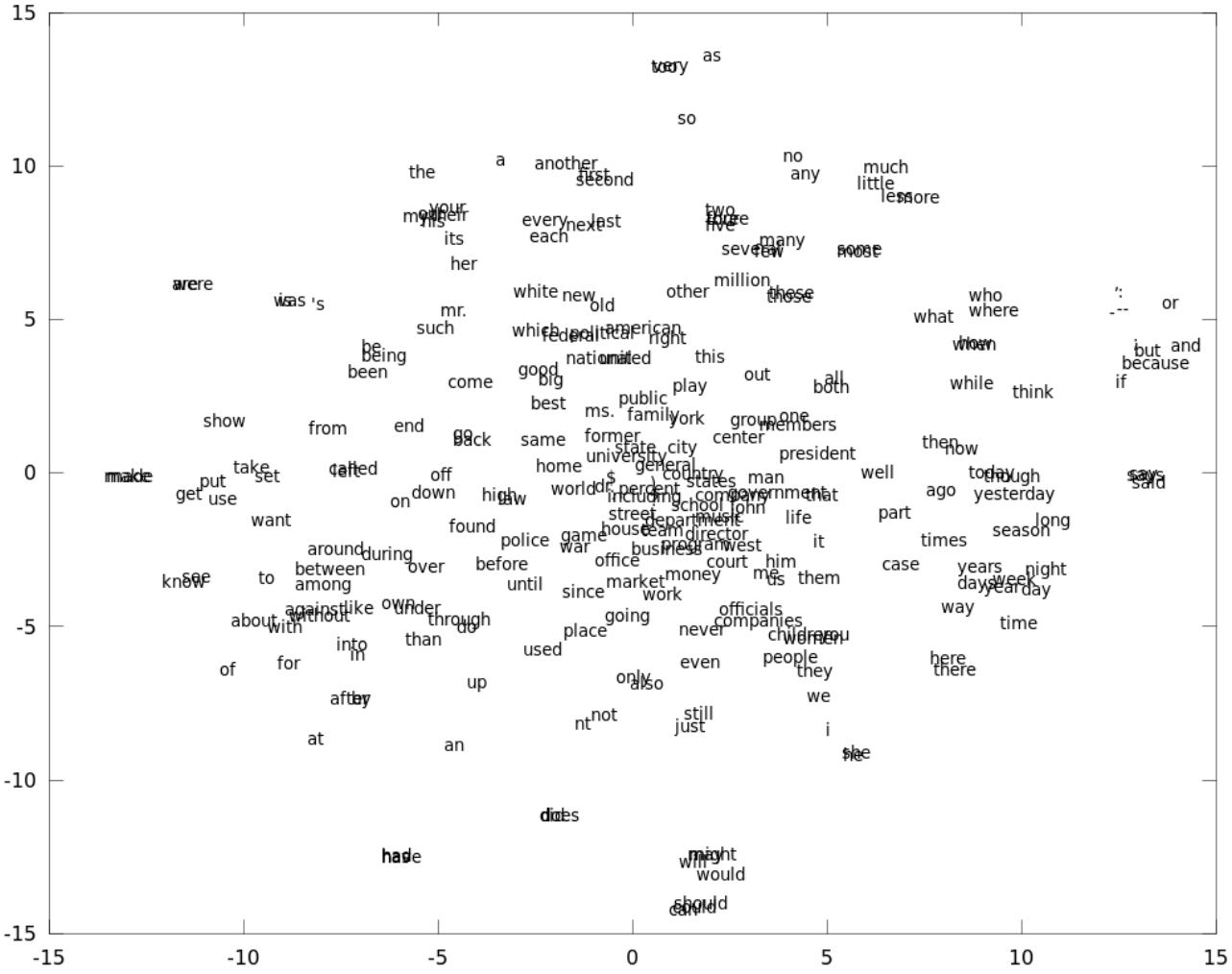
Distributed vector representations of words



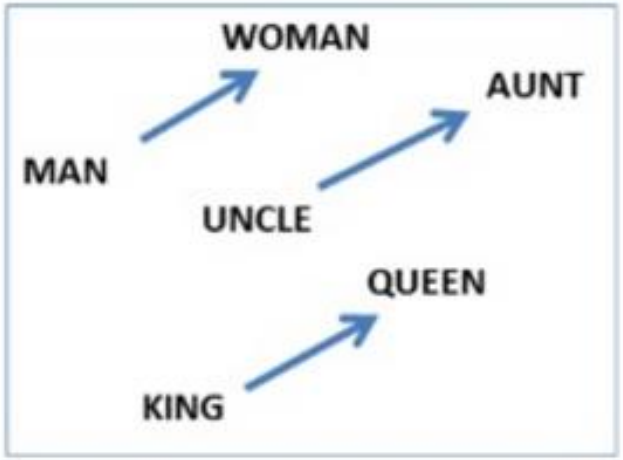
Distributed vector representations of words

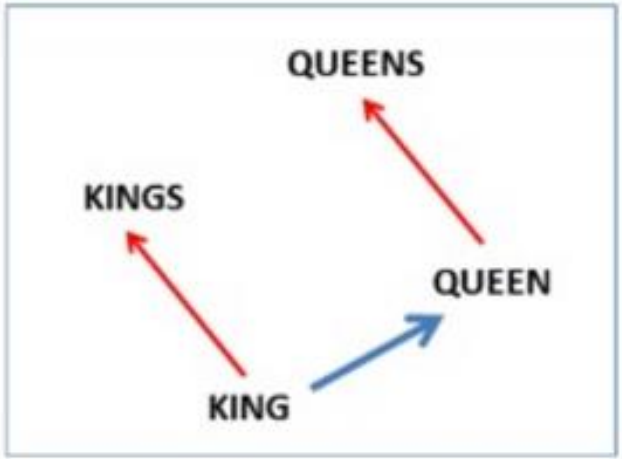


Distributed vector representations of words



Distributed vector representations of words

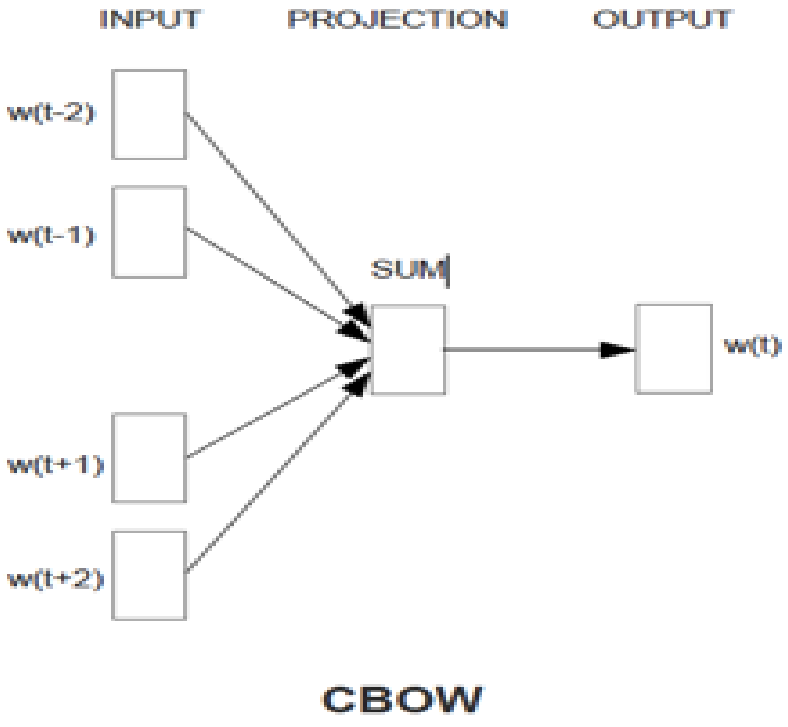




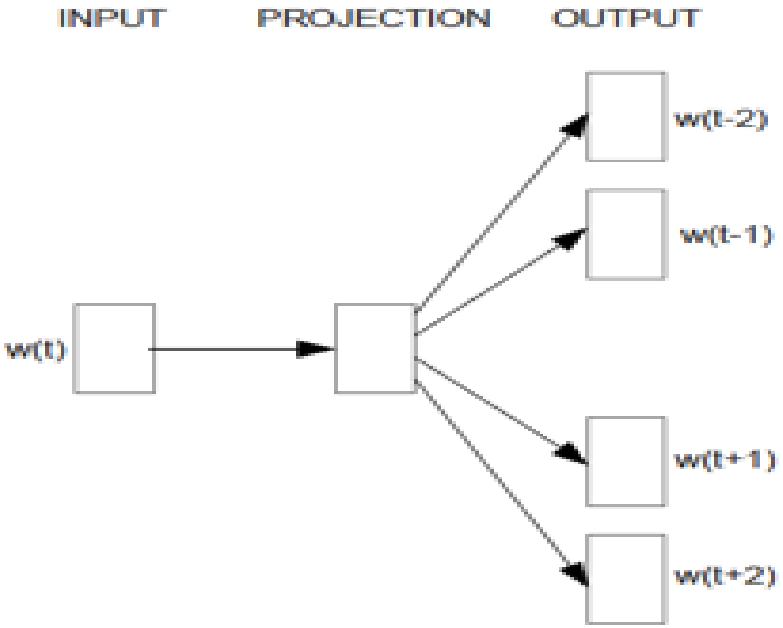
Rel

France : Paris	Italy : Rome	Japan : Tokyo	Florida : Tallahassee
big : bigger	small : larger	cold : colder	quick : quicker
Miami : Florida	Baltimore : Maryland	Dallas : Texas	Kona : Hawaii
Einstein : scientist	Messi : midfielder	Mozart : violinist	Picasso : painter
Sarkozy : France	Berlusconi : Italy	Merkel : Germany	Koizumi : Japan
copper : Cu	zinc : Zn	gold : Au	uranium : plutonium
Berlusconi : Silvio	Sarkozy : Nicolas	Putin : Medvedev	Obama : Barack
Microsoft : Windows	Google : Android	IBM : Linux	Apple : iPhone
Microsoft : Ballmer	Google : Yahoo	IBM : McNealy	Apple : Jobs
Japan : sushi	Germany : bratwurst	France : tapas	USA : pizza

Distributed vector representations of words



CBOW



Skip-gram

Embedding From Language Models (ELMo)

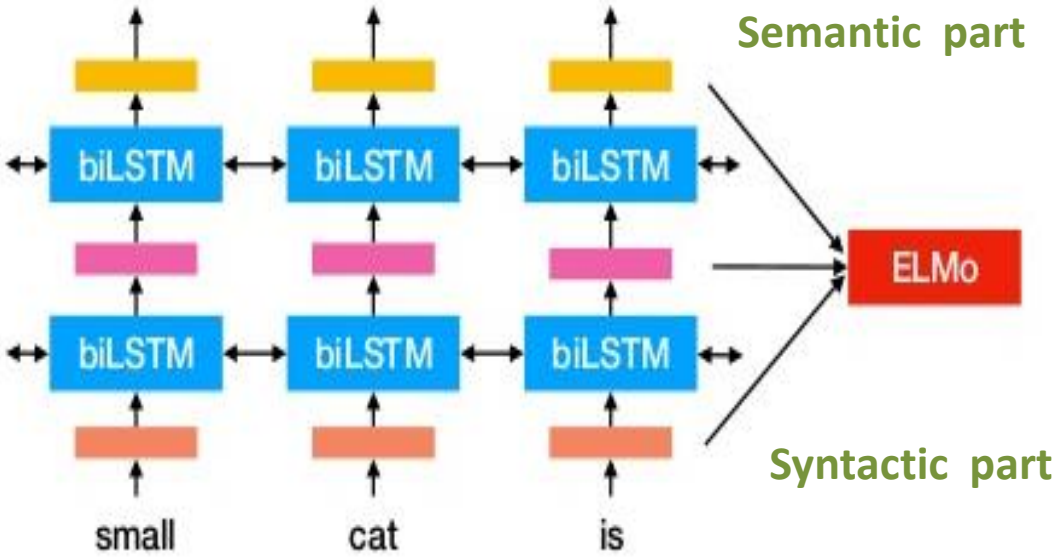
Learning high quality word representations should ideally model

1. Complex characteristics of word use (e.g., syntax and semantics)
2. How these uses vary across linguistic contexts (i.e., to model polysemy)

Can you help me

A can of soup

- **Deep contextualized word representation**
- **Each word is assigned a representation that is a function of the entire input sentence**



Proposed Framework

- *Assume we have a concept set c*

$$c=\{c_1,\ldots,c_N\}.$$

- *For a Video V we can calculate the vector*

$$c(v) = [p(c_1|v),\ldots,p(c_N|v)]$$

Where $p(c_i|v)$ represent the propobaility of having the concept c_i in the Video V

- *a(v):* denote the speech content of the video
- *o(v):* denote the text content of the video
- $= p(e|c(v))\ p(e|a(v))\ p(e|o(v))$

Proposed Framework

- *For a query e , we can calculate the probability that video V match query e by :*

$$p(e|v) = p(e|c(v), a(v), o(v))$$

$$= p(e|c(v))\, p(e|a(v))\, p(e|o(v))$$

With the independence assumption

Proposed Framework

- *If we have*
  - $\theta(e)$: *The mapping of the query e to the* distributional semantic space
  - $\theta_c(v)$: *The mapping of the video concepts content to the* distributional semantic space
  - $\theta_a(v)$: *The mapping of the video speech content to the* distributional semantic space
  - $\theta_o(v)$: *The mapping of the video text content to the* distributional semantic space

Proposed Framework

if we assume

- $p(e|a(v)) \propto s(\theta(e), \theta_a(v))$
- $p(e|o(v)) \propto s(\theta(e), \theta_o(v))$
- $p(e|c(v)) =$

$$\sum p(e|c_i)p(c_i|v) \propto \sum s\left(\theta(e), \theta_{(c_i)}(v)\right)p(c_i|v)$$

Similarity Measure

- *For multi words vectors in sentence or paragraph we pool the vectors*

$$X' = \sum_i x_i$$

- *For similarity measure use cosine similarity as proposed by Mikolov*

$$s_p(X, Y) = \frac{(\sum_i x_i)^T (\sum_j y_j)}{\left\|\sum_i x_i\right\| \left\|\sum_j y_j\right\|}$$

Preliminary Results

- Used TREC Video Retrieval Evaluation (TRECVid).

- A yearly challenge sponsored by NIST to promote research in content-based retrieval and analysis of videos.

Preliminary Results

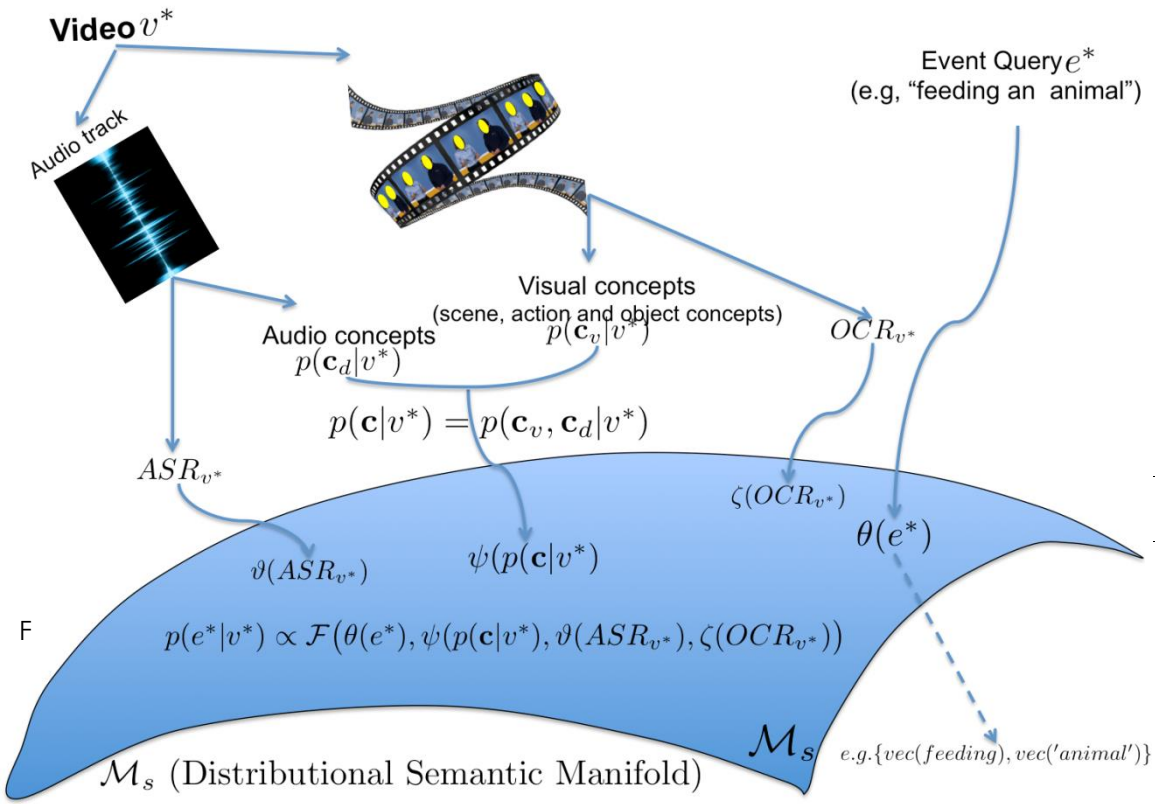
- Two distributional semantic models
  - Wikipedia model trained on 1 billion words, vocabulary size 120k, word vectors of 250 dimensions.
  - Google-News model trained on 100 billion words, vocabulary size 3 million words, word vectors of 300 dimensions.

Preliminary Results

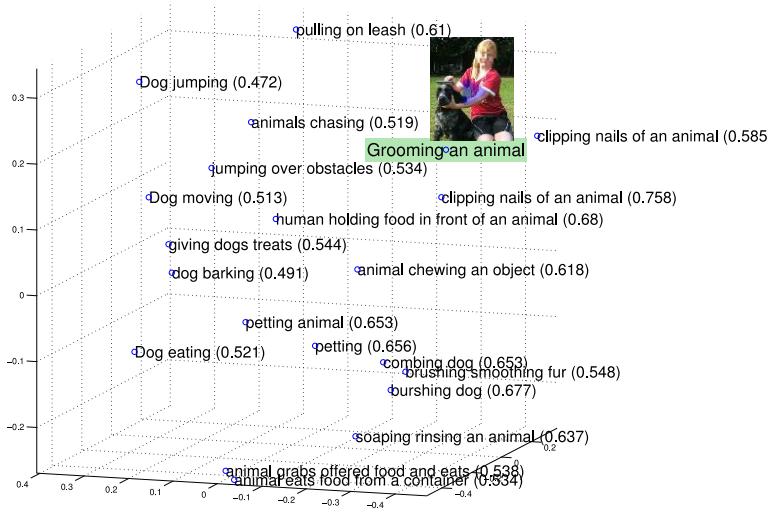
Used

- 1000 visual object concepts
- 500 scene concepts
- 500 action concepts
- Speech recognition
- Text detection / Transformation /Recognition

Preliminary Results



F



Preliminary Results

Modality	Approach	MAP	AUC
Visual concepts	Proposed Approach (event title query)	8.36%	0.834
Visual concepts	Dalton et al [10] (event title query)	3.4%	-
Visual concepts	Dalton et al [10] (Manually specified concepts)	7.4%	-
Visual concepts	Baseline - Overfeat [41] direct concept mapping	2.43%	
Speech	Proposed approach – Speech/GNews	4.23%	0.621
Speech	Baseline – matching	2.77%	0.567
Text	Proposed approach – OCR/GNews	4.81%	0.623
Text	Baseline - matching	1.8%	0.536
Text + Speech	Proposed approach	10.6%	0.67
Text+Speech+Visual	Proposed approach (All fused)	13.1%	0.830
Text+Speech+Visual	State-of-the-art [47]	12.6	0.730

Mean Average Precision (MAP) and Area under the ROC (AUC)

Arabic Word Vectors

Trained on 5.8 billion Arabic words

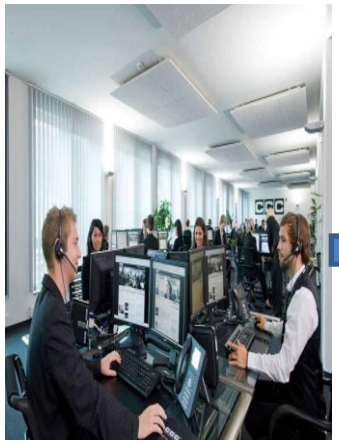
- Built three models for Arabic (CBOW, SKIP-G and GloVe)
- Used translation of Mikolov analogy test

Type of relation ship	Word pair 1				Word pair 2			
Common capital city	Athens	أثينا	Greece	اليونان	Oslo	اوسلو	Norway	النرويج
Man-Woman	brother	شقيق	sister	شقيقة	grand son	حفيد	grand daughter	حفيدة
Superlative	bad	سيء	worst	اسوأ	big	كبير	biggest	اكبر
Plural nouns	bird	طائر	birds	طيور	car	سيارة	cars	سيارات

Arabic Word Vectors

Model	English SKIP-G300		English GloVe300		Arabic CBOW300		Arabic SKIP-G300		Arabic GloVe300	
Training words	300B		840B		5.8B		5.8B		5.8B	
	Cov.	Acc.	Cov.	Acc.	Cov.	Acc.	Cov.	Acc.	Cov.	Acc.
capital-common-countries	100	94.9	100	**100**	100	94.3	100	93.7	100	92.7
capital-world	100	93.1	100	**95.6**	100	74.7	100	77	100	80.4
currency	100	**37.8**	100	13.2	100	7.7	100	7.9	100	5.7
city-in-state	100	87.2	100	**87.4**	100	32.5	100	32.6	100	36.4
family	100	**95.3**	100	86.8	67.6	46.5	67.6	36.3	67.6	50.3
adjective-to-adverb	100	53.8	100	**65.6**	100	34.2	100	30.1	100	22
opposite	100	**57.6**	100	45.8	80	3.7	80	3.2	80	3.5
comparative	100	**97**	100	96.6	100	73.8	100	67	100	71.5
superlative	100	**95.4**	100	93.7	100	68.9	100	64.6	100	66.9
present-participle	100	96.7	100	**97.4**	93.9	46.1	93.9	42.1	93.9	30.5
national-ity-adjective	100	**95.7**	100	89.2	100	49.9	100	55.5	100	44.2
past-tense	100	**93.7**	100	91.9	100	44.7	100	41.6	100	43.5
plural	100	95.8	100	**96.5**	100	56.1	100	56.9	100	57.7
plural-verbs	100	89.5	100	**90.3**	100	80.1	100	75.5	100	72.2
TOTAL	100	**87.4**	100	86.2	98	54.3	98	53.6	98	53.5

Call Centers logs Mining



Automatic Speech Recognition

Calls Transcripts

- Key word spotting
- Call type classification
- Name Entity Extraction
- Sentiment analysis
- User mode recognition

Data Processing Modules

Analytics Info

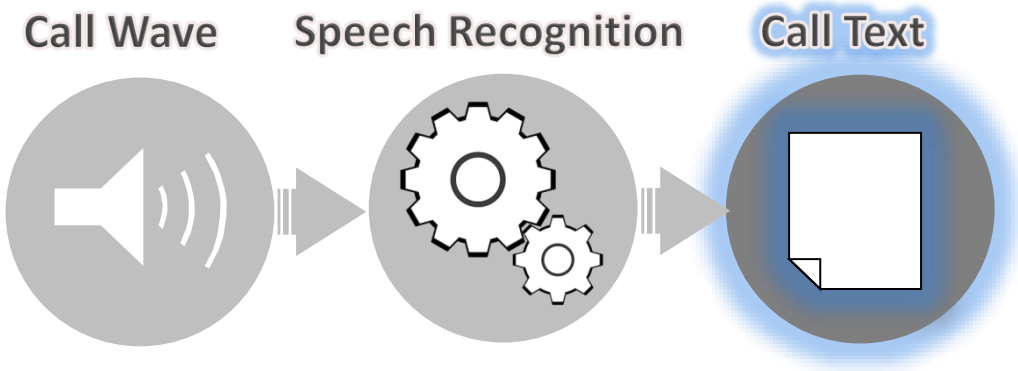
Back-End IVR System

Management Reports

Call Centers Analytics

Introduction

- The main goal of this module is to extract useful information from the ASR text transcripts of call-centers calls. In this module, there are several functions to achieve the required features of this system



Call Centers Analytics

Data preparation

- This phase aims to prepare words lists needed in the main functions.
- To extract these lists a word2Vector model is trained with large dialect Arabic corpus.
- This model is trained using over 70 million lines gathered from Arabic websites and produced a vocabulary of 1.7 million unique words.

Call Centers Analytics

Data preparation

- This task extracts a list of words relevant to the needed classes. Given the initial word lists, which are extracted manually.

- Lists are collected and manually revised for the following classes:

Ipod	*Phone*	*Buy Keywords*
Watch	*Tablet*	*Support Keywords*
TV	*PC*	*Complain Keywords*
Laptop	*Agent Evaluation*	

Call Centers Analytics

Text analyzer

Advertising Removal

- The input call needs to pass through a preprocessing unit to process the call before going through the rest of functions. The main function in the preprocessing unit is the advertisement removal function.

- The main goal of this function is to remove the advertising at the start of each call, which fills the waiting time.

Call Centers Analytics

Text analyzer

Call Centers Analytics

Text analyzer

Advertising Removal

Opening & Closing Clauses
Detects opening and closing expressions
(صباح الخير, شكرا لاتصالك , ...)

Agent Evaluation
Evaluates the performance of the agent during the call [0-9 range]

Call Type
Detects the type of the call (buying, complain, support)

Vendor Mentions
Counts how many times the name of the company and its synonyms mentioned

Competitors Mentions
Counts how many times the name of the competitor and its synonyms mentioned

Products Mentions
Detects and counts the mentioned products and accessories

Call Centers Analytics

Text analyzer

Call Type Classification

Scoring

Word list matching

$$score(C,W)=\frac{\sum_{i}^{n} match(C,W_i)}{m}$$

Word embedding score

$$score(C,W)=\frac{\sum_{i}^{n}\sum_{j}^{m} similarity(C_j,W_i)}{n}$$

Where C and W are the input call and the class word list and n is the length of the word list, and m is the length of input call. After getting the score for the three classes the output class considered to be the one that has the highest score.

Call Centers Analytics

Text analyzer

Results on real data

Agent Speech

مسا الخير خلود مع حضرتك مسا الخير يافندم اتشرف بالاسم أستاذه أمنيه أنا أساعد حضرتك سيتي إزاي اساعد حضرتك سكس للأسف أبل مش موفره قطع غيار في مصر بيحصل أكس اتشنج للجهاز حضرتك بالكامل للأسف أبل مش موفره الفتره دي في مصر بئالها فتره يافندم للأسف أنا معنديش علم موجوده في أماكن تانيه أو لأ تحت أمر حضرتك أي استفسار أخر شكرا لاتصال حضرتك بتريد لين ستورز

Opening (3)

- مسا الخير
- مع حضرتك
- أتشرف بالاسم

Closing (3)

- تحت أمر حضرتك
- أي استفسار آخر
- شكرا لاتصال حضرتك

Eval. (8.25/9)

- يا فندم
- حضرتك
- .....

Call Centers Analytics

Text analyzer

- **Call Type**
  - Support
  - Buy
  - Complain
- **Company Mentions** — How many times the name of the company mentioned in the call
- **Products Mentions** — How many times the names of the products mentioned in the call

Results on real data

مسا النور لو سمحت كنت عايزه أعرف شاشة لاب أبل تلت ستاشر تسعه وستين ألفين وعشره تعمل كام طيب أوكيه مفيش مشكله لأ هو ده اللي معايا خلاص أوك أتفضل تمام زيرو حداشر أربعه خمسه تمام تمام أوكيه ميرسي

S 1 2

فالنور باشمهندس خلود بعد إذن حضرتك ممكن أعرف أي فون سيفن ميه ميه تمنيه وعشرين عامل كام اللي عادي طيب السكس إس ميه خمسه تمنيه تمنيه وعشرين ستميه تسعه وأربعين ماشي شكرا يا فندم تاني يافندم كا عامل كام السكس إس بلس سبعتاشر ماش شكرا شكرا لأ مأنا بس البيركس

B 0 3

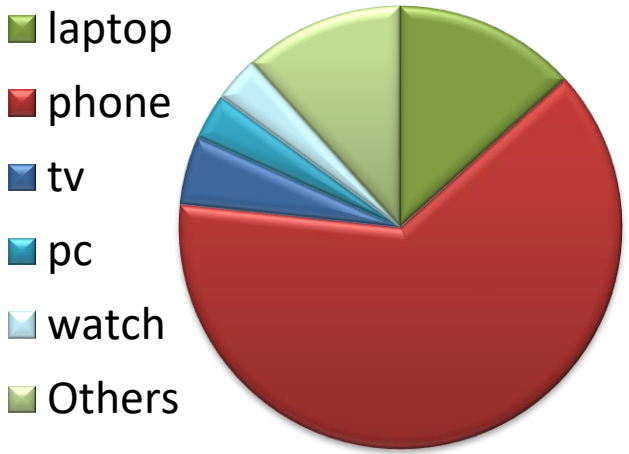
مسا النور الآي سي الشحن بتاع الموبايل موبايل آيفون سكس كل ماحط للبتاع ينور عليكي كده خلاص كوره نحط بتاع ينور أه عملتها والله بس أكتر من مره بس ماحصلش حاجه المازر بورد أو حاجه فيها يعني بكام في الرينج بدفع حاجه يعني آه هو بيتخطفوا الموبايل في يعني وأنا واقف ولا بياخدوا وبعد كده كلموني وكده آه شكرا

C 0 4

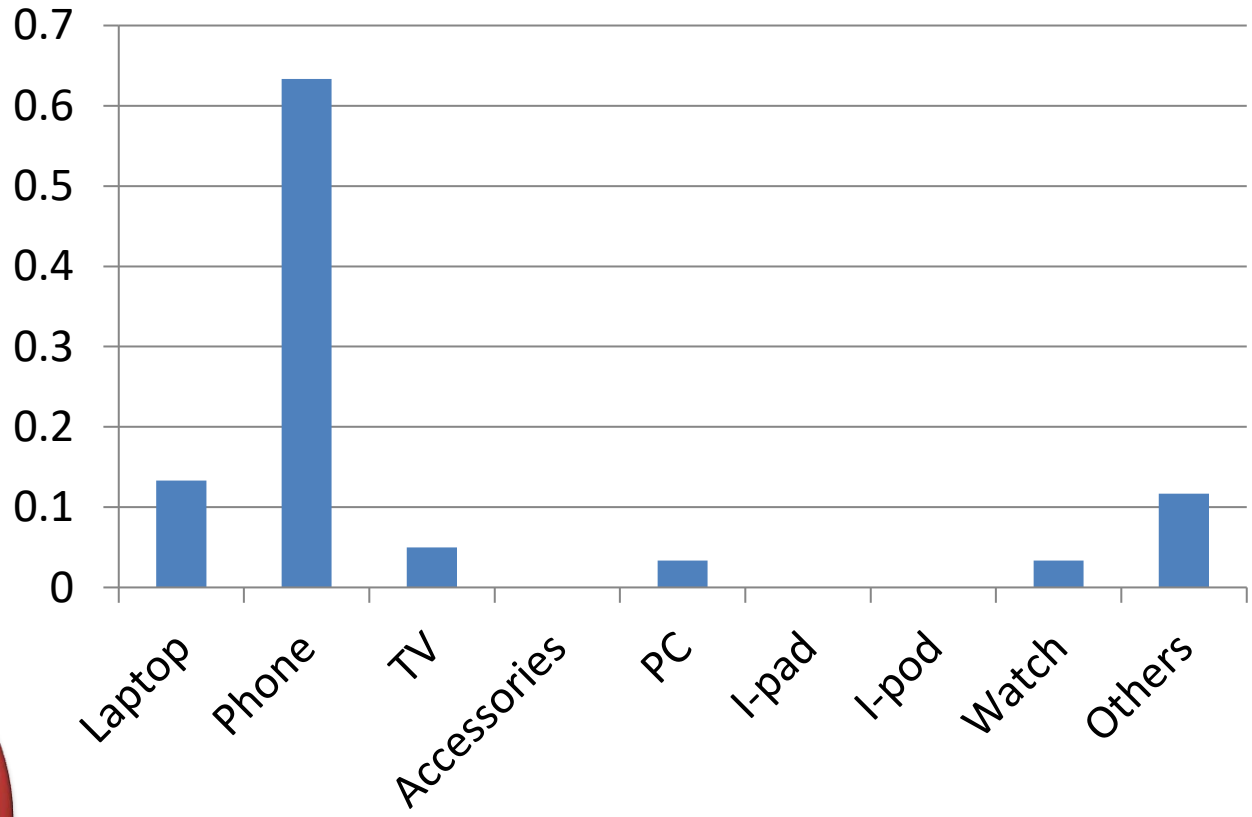
Call Centers Analytics

Mentioned Products

Product Name	Mentions (%)
Laptop	13.3%
Phone	63.3%
TV	05.0%
Accessories	00.0%
PC	03.3%
I-pad	00.0%
I-pod	00.0%
Watch	03.3%
Others	11.7%



Results on real data

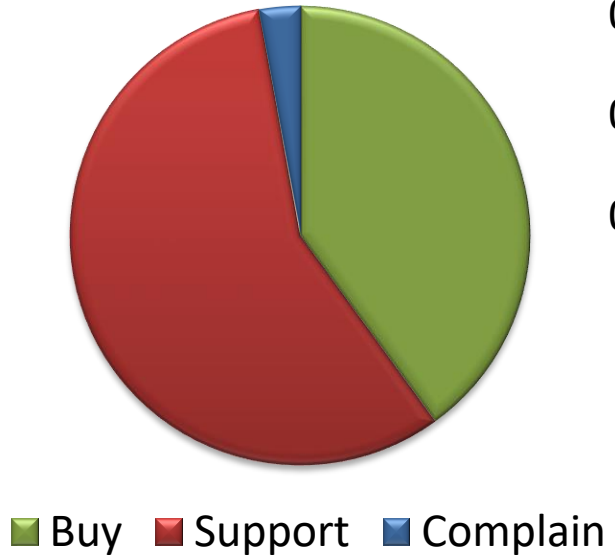


Phones represents 63% of customers' interests

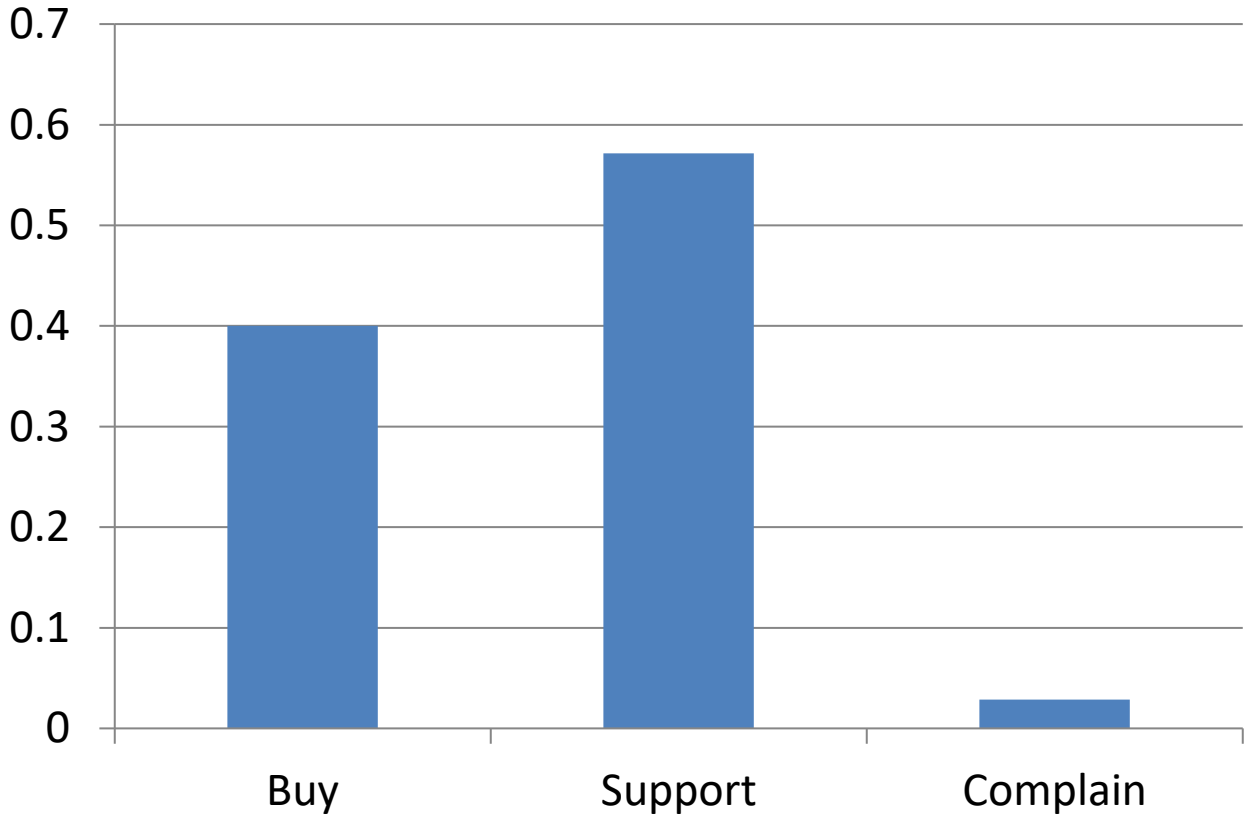
Call Centers Analytics

Call Types

Call Type	Frequency
Buy	40.0%
Support	57.2%
Complain	02.8%



Results on real data

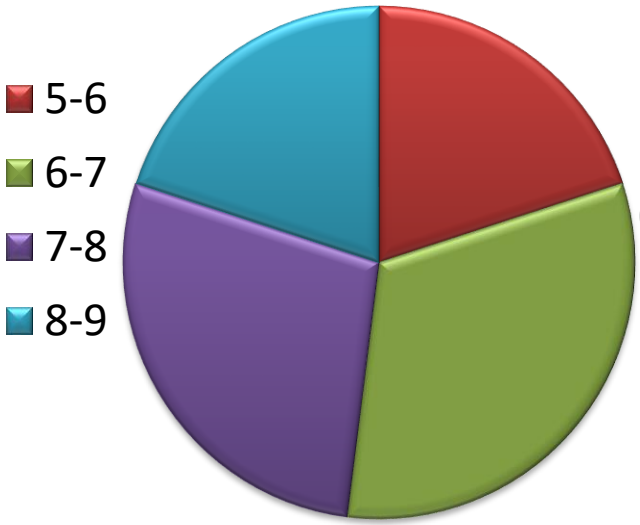


Support calls represents 57.2% of customers' calls

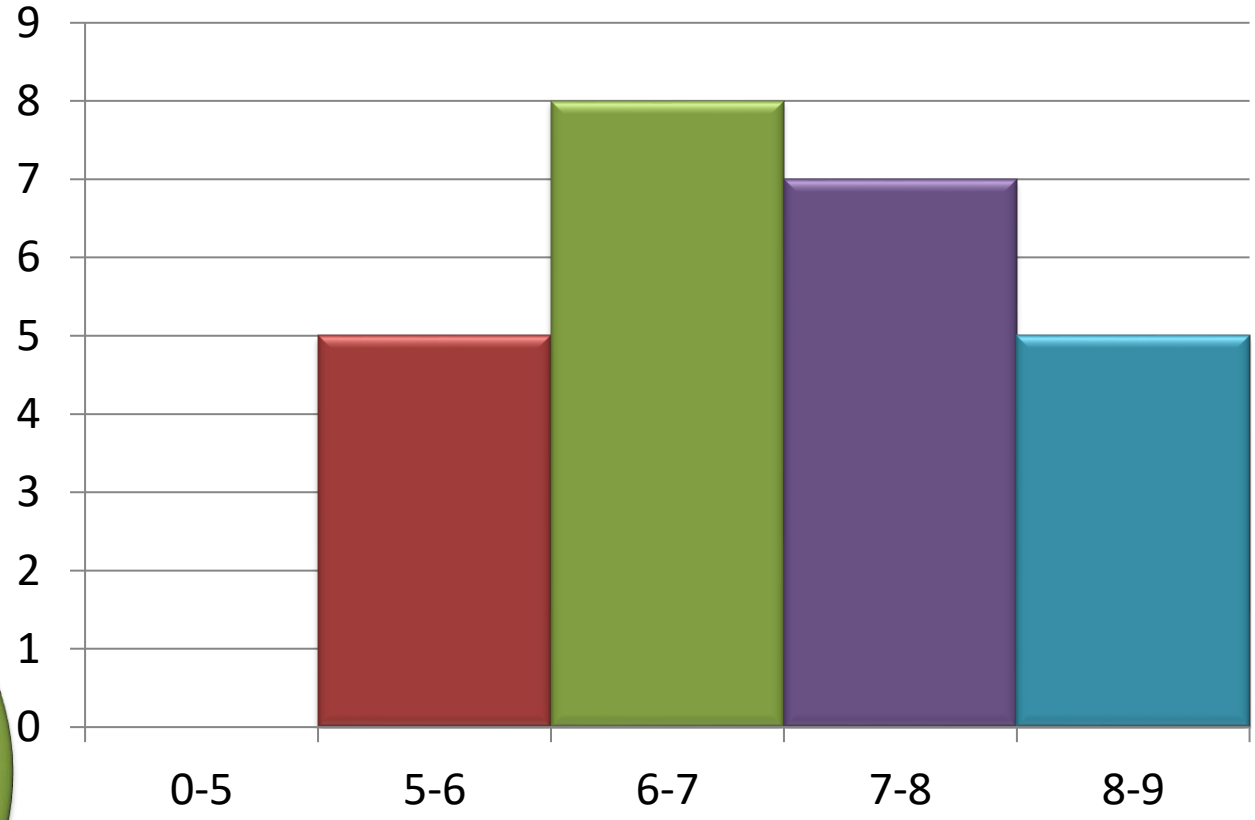
Call Centers Analytics

Agents Evaluation

Evaluation Ranges	Frequency
0-5	0
5-6	5
6-7	8
7-8	7
8-9	5



Results on real data



No Agent has an evaluation less than 5/9

Name Entities Recognition

Introduction

- Named Entity Recognition (**NER**) is a core task in Natural Language Processing. It is considered to be an important component for any intelligent system which works on text analysis and processing. NER task is about classifying the sentence words and entities into categories, such as **PERSON**, **ORGANIZATION** or **LOCATION**.

- We avoid any language-depended features as we want to apply the same architecture on MSA and Dialectal Arabic.

Name Entities Recognition

Data Collection and Annotation

Web Crawlers:
A crawler for a set of known news portals was built. The generated corpus consists of 991,014,528 words (11,325,363 unique words).

Normalization & Preprocessing:
After removing the contexts with non-Arabic words, we extracted the nine-word contexts of the most frequent trigrams. A set of three million contexts was generated. This normalized corpus passed to the linguists to apply the annotation process

Annotation Tool:
The named entities annotation tool helps in two main tasks

- Named entities collection (gazetteers)
- Annotation of a clean corpus

Name Entities Recognition

Data Collection and Annotation

Annotation Tool:

Name Entities Recognition

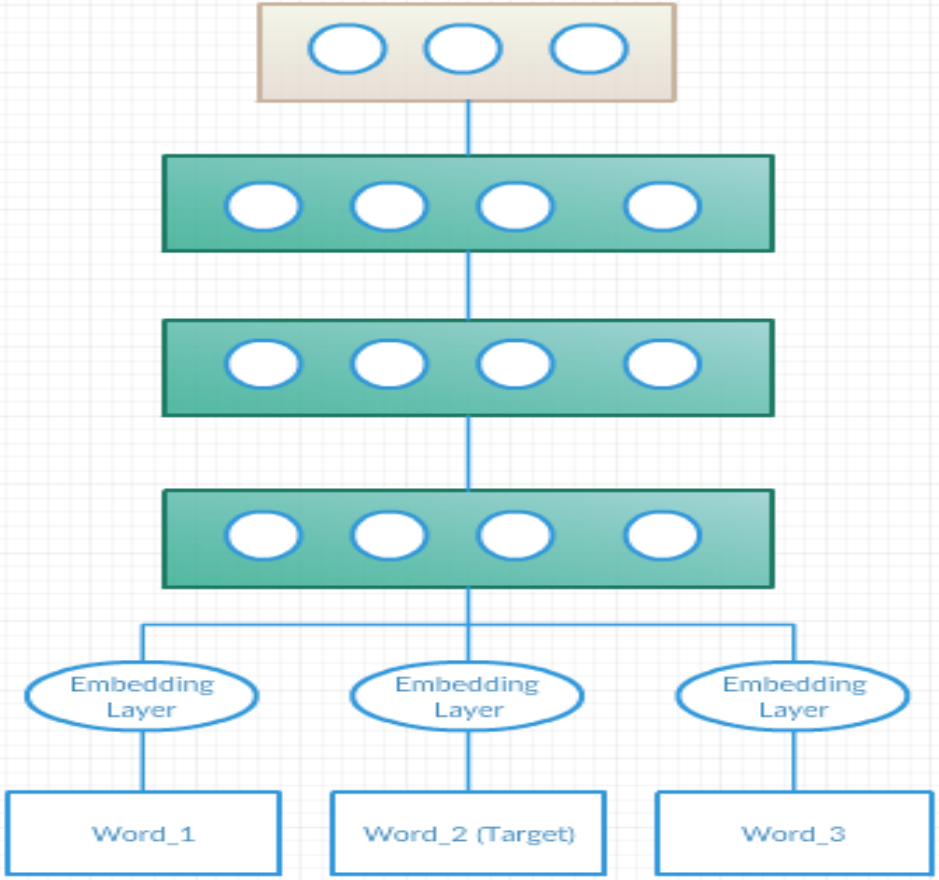
Features Extraction

We used Word2Vec technique as our Word Embedding model to build a look-up table for each unique word in our corpus. In general, there are 2 different Neural Networks architecture for Word2vec: Skip-gram Model and Continuous Bag of Words (**CBOW**), we used Skip-gram as according to the research community, Skip-gram gives better results and better performance. We trained the Word2vec using a huge corpus of Modern Standard Arabic.

Number of Sentence	**9137635**
Number of Unique Words	**1114811**

Name Entities Recognition

NER Architecture



Layer	Number of neurons
Input	300 + 300 + 300 = 900
Hidden Layer 1	1500
Hidden Layer 2	500
Hidden Layer 3	100
Softmax Layer	9

Name Entities Recognition

Results

We used open-source **ArabicNER** dataset along side our data collected from **Twitter** and **Wikipedia.** But our latest system is trained on our data only.

Total Number of Samples	**884102**
Number of Unique Words	**88458**

	Person	Location	Organization
F-score	86.2%	84.4%	81.2%

Demos

Try life demos on

www.rdi-eg.com

…… Thank You