Discriminative keyword extraction

Márius Šajgalík
marius.sajgalik@stuba.sk
FIIT STUBA

Keywords

Metadata

Concise representation of text, images, etc.

Help to organize data

Discriminative keywords

Keywords can be used to **discriminate** between **categories**

Higher information value

Avoid generally important words

Our method of discriminative keyword extraction

1. Candidate phrase extraction
2. Computation of **distributed representation** for candidate phrases
3. Substitution of phrases by most similar words
4. Computation of **discriminative metric**
5. Keyword selection (ranking)
6. Computation of document vector

Distributed representation of words

Words mapped to **vectors**

- word vectors, word embeddings

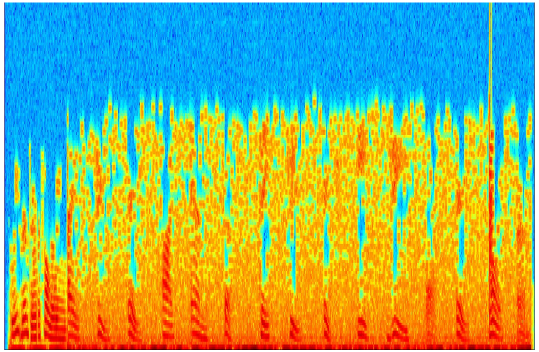
Distributed word **features**

Distributional hypothesis

> ***"Words are similar if they appear in similar context."***

Motivation: Why distributed representation?

audio



DENSE

images

DENSE

text

0	0	0	0.2	0	0.7	0	0	0	...	...

SPARSE

Properties of Distributed Representation

Multiple degrees of similarity

We can do **vector operations**

vector(”snow”) + vector(”ball”) ~= vector(”snow ball”)

Syntactic relations

vector(”biggest”) − vector(”big”) + vector(”small”) ~= vector(”smallest”)

Semantic relations

vector(”Paris”) − vector(”France”) + vector(”Germany”) ~= vector(”Berlin”)

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Table from: Mikolov et al. Efficient Estimation of Word Representations in Vector Space. ICLR, 2013.

word2vec

Open sourced tool from Google researchers

Learns **word embeddings** from **raw text**

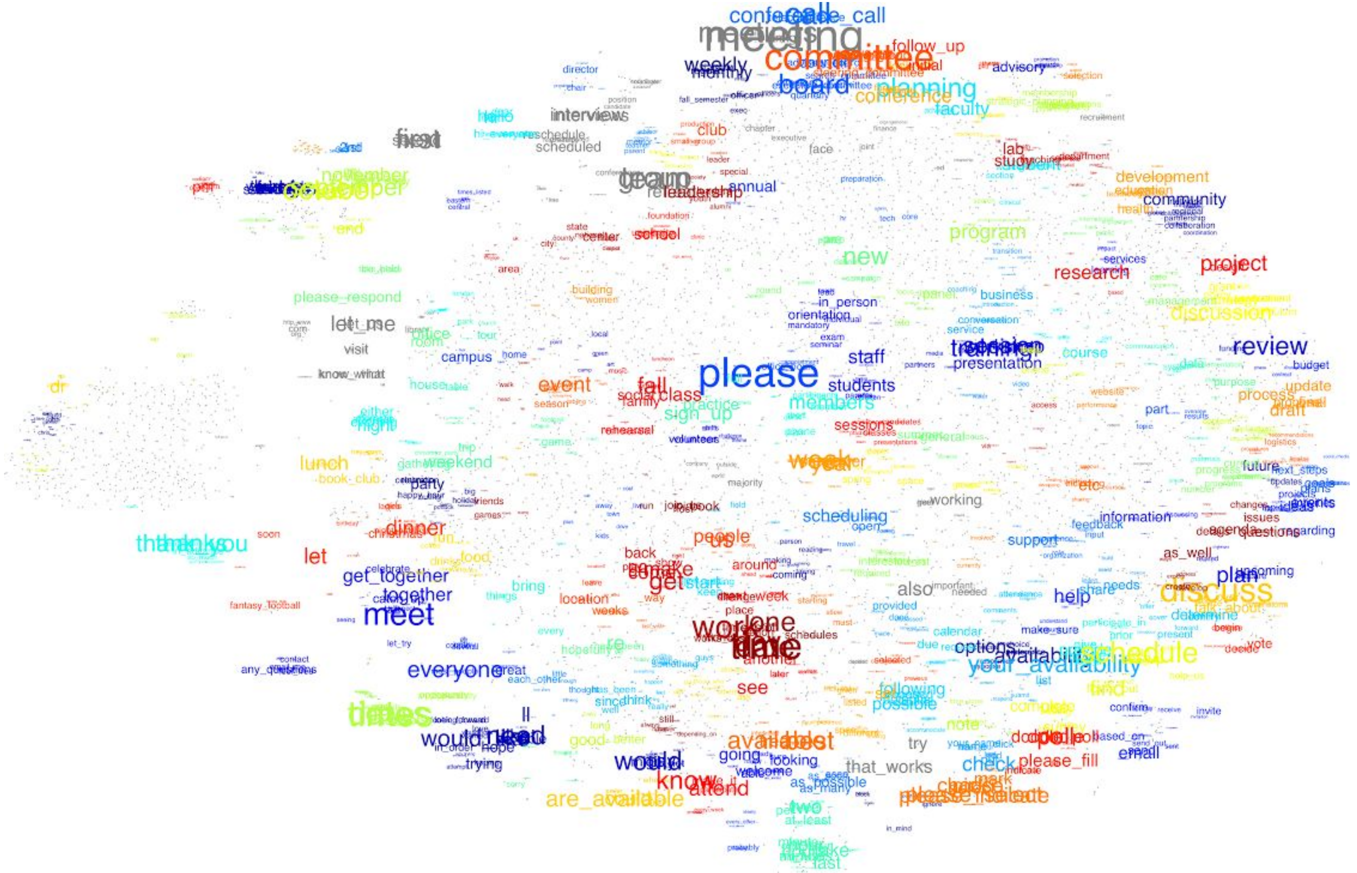
Efficient parallel implementation in C with pretrained English model

https://code.google.com/archive/p/word2vec/

Gensim - Python implementation for multiple tasks

http://radimrehurek.com/gensim/models/word2vec.html

Visualization of word embeddings with t-SNE



tutoring
ta
lab
math
english
division
department
study
teaching
internship
studies
nursing
law
section
medical
facilities
development
education
technology
community
network
health
regional
collaborative
global
partnership
housing
collaboration
coordination
search
clinical
stem
professional_development
international
engagement
career
fund
program
public
governance
industry
staffing
impact
transition
services
internal
emergency
leads
research
learning
project
design
energy
care
consultation
coaching
programme
grant
action
track
business
based
policy
management
strategy
introduction
half_day
web
technical
standards
discussion
financial
conversation
service
history
security

Distributed representation of ???

Web pages visited by users

Documents bookmarked by users

Products viewed/bought by customers

Our method of discriminative keyword extraction

1. Candidate phrase extraction
2. Computation of **distributed representation** for candidate phrases
3. Substitution of phrases by most similar words
4. Computation of **discriminative metric**
5. Keyword selection (ranking)
6. Computation of document vector

Discriminative metrics

Based on **categorisation** of text documents

Most of the metrics can be expressed by **ABCD statistics**

Frequency	word W	other words
in category CAT	A	B
in other categories	C	D

Discriminative metrics (2)

Metric name	Function expressed in terms of ABCD statistics
rf	$log(2 + \frac{A}{max(C,1)})$
tds	$\frac{A/(A+B)}{(A+C)/N}$
ig	$N \times \frac{A}{N} \times \frac{log(A \times N)}{(A+C)\times(A+B)} \times \frac{B}{N} \times \frac{log(B \times N)}{(B+D)\times(A+B)} \times \frac{C}{N} \times \frac{log(C \times N)}{(A+C)\times(C+D)} \times \frac{D}{N} \times \frac{log(D \times N)}{(B+D)\times(C+D)}$
gr	$-ig/ (\frac{A+B}{N} \times log(\frac{A+B}{N}) + \frac{C+D}{N} \times log(\frac{C+D}{N}))$
χ^2	$N \times \frac{(A \times D - B \times C)^2}{(A+D)(B+C)(A+B)(C+D)}$
idf	$log(\frac{N}{A+C})$

Our method of discriminative keyword extraction

1. Candidate phrase extraction (noun phrases)
2. Computation of **distributed representation** for candidate phrases
3. Substitution of phrases by most similar words (kNN in vector space)
4. Computation of **discriminative metric**
5. Keyword selection (just sorting)
6. Computation of document vector (a sum)

Computation of distributed representation for candidate phrases

Summation of vectors of individual terms

Terms - words and short phrases - entries in the dictionary

Dynamic programming to compute the vector of ambiguous phrases

Our method of discriminative keyword extraction (2)

Article about a protest in harbours

- category *ships* in Reuters-21578 dataset

pirates, *pirate, steamer, anchorage, Ship, sail, ferry, destroyer, Ships, destroyers*

Our method of discriminative keyword extraction (3)

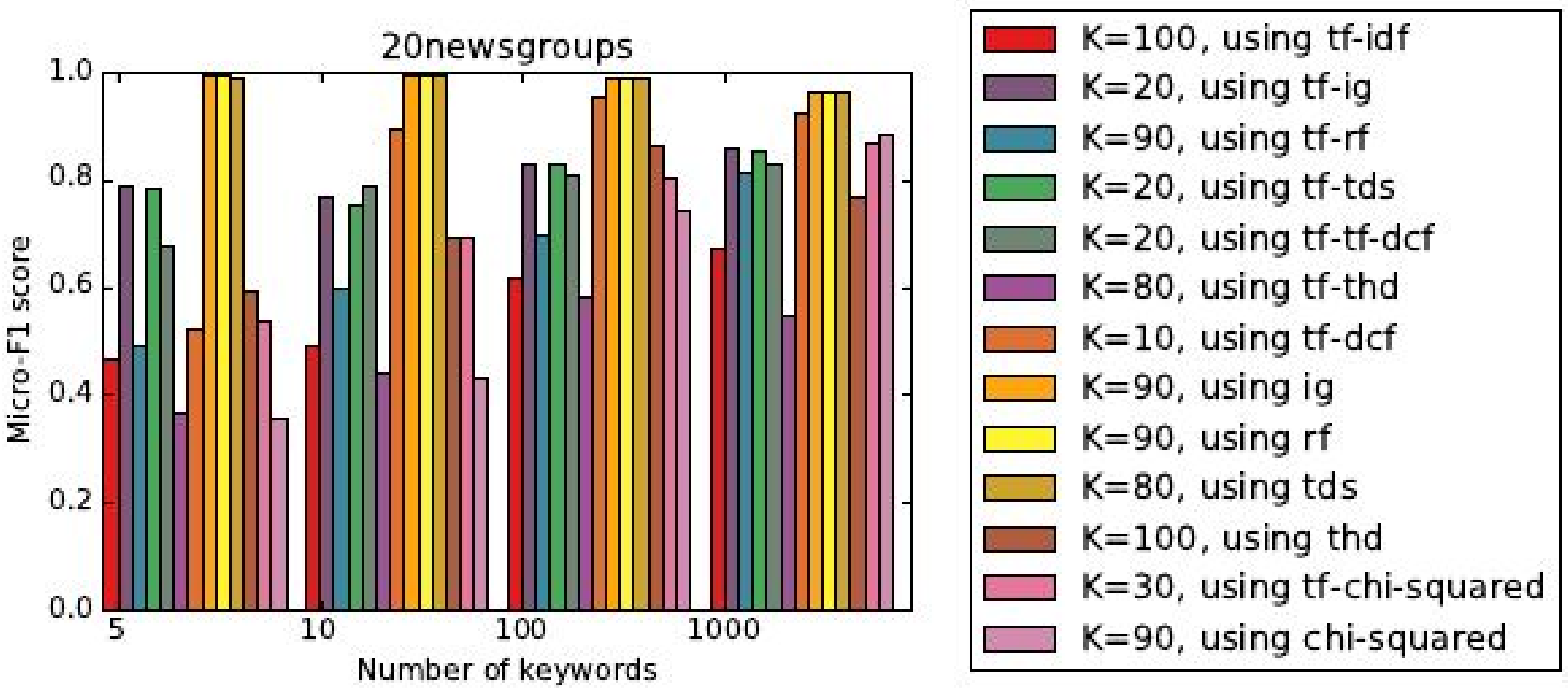
Manual evaluation

- Part of 20newsgroups dataset
- TF-RF baseline

Automatic evaluation

- 4 different datasets
- Analysis of influence of the individual parameters
    - Choice of discriminative metric
    - Number of extracted words

Our method of discriminative keyword extraction (4)



Our method of discriminative keyword extraction (5)

+ Capturing **abstract** concepts
+ **Discriminative** representation
+ **Justification** by distributed representation

Our method of discriminative keyword extraction (5)

+ Capturing **abstract** concepts
+ **Discriminative** representation
+ **Justification** by distributed representation

+ Weakly language dependent
+ **Scalability**
+ **Multiple** concepts mixing together
- Multiple concepts **mixing** together

Modelling user interests

User model as a tripartite graph (Mika, 2007)

$$G = (V, E)$$

$$V = U \cup W \cup D$$

$$E = \{(u, w, d) \mid u \in U, w \in W, d \in D\}$$

Users as categories

Using our method of discriminative keywords extraction

Modelling user interests (2)

Discovering value from community activity on focused question answering sites: a case study of stack overflow

question, answer, ask, answering, yes, query, ponder, rephrase, clue

Hybrid Web Recommender Systems

recommend, propose, autocompletion, recommended, predefine, recommendation, consider, websearch, inferencing, recommender

Context-aware query classification

contextualisation, contextualization, relevance, disambiguate, contextualise, contextual, contextualized, context, disconfirm

Modelling user interests (3)

Evaluated on 2 datasets

Annota

- Bookmarked research articles in digital library

Brumo

- Web browsing logs

Modelling user interests (4)

- (Dis)advantages of the method of discriminative keyword extraction

\+ **Personalised** keywords

\+ **Fast** automatic **evaluation** of classification

Diversion: Neural networks architectures

Neural networks as a black box

Traditional architectures of the last century

Mostly supervised learning

Neural networks as a white box

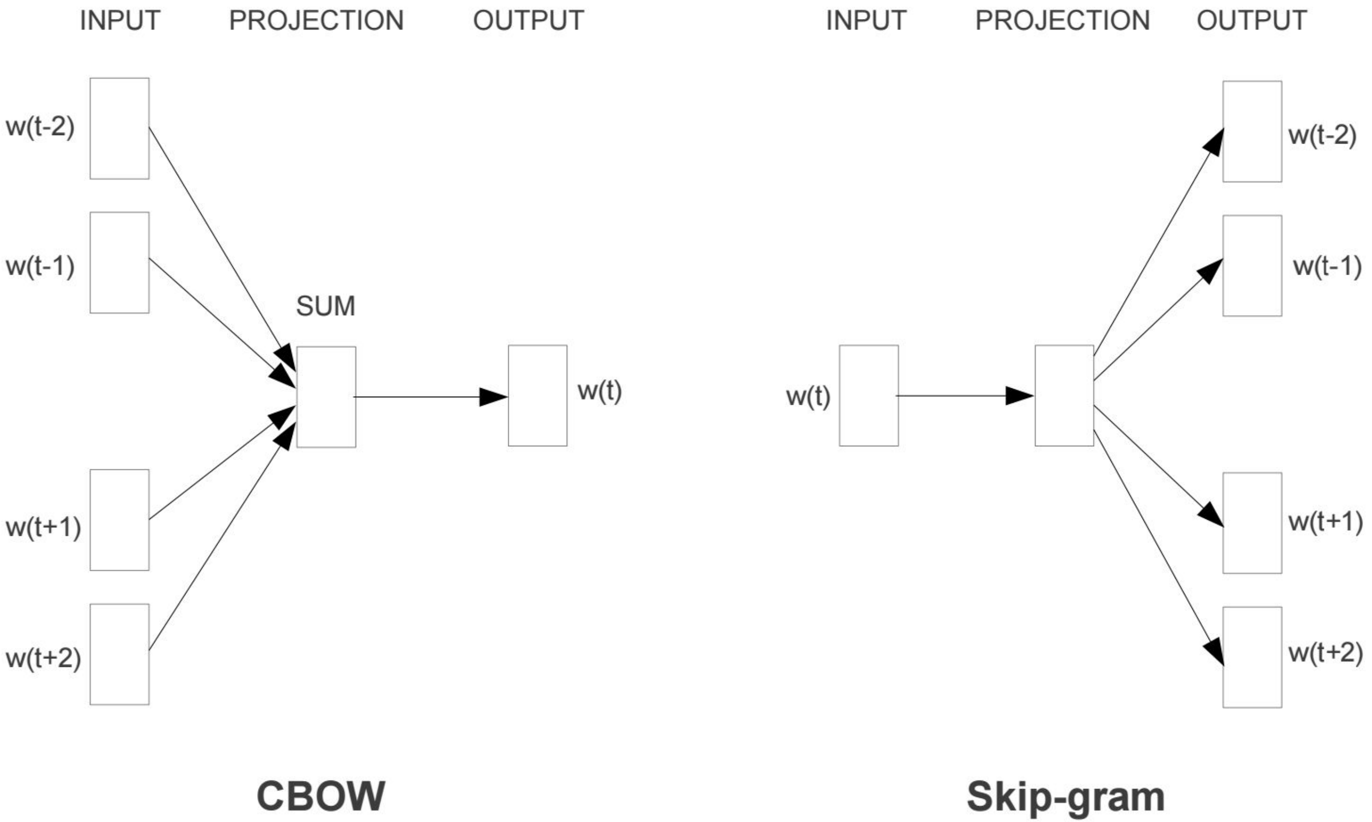
Unconventional architectures

Multiple outputs

Output in hidden layers

Network modularisation

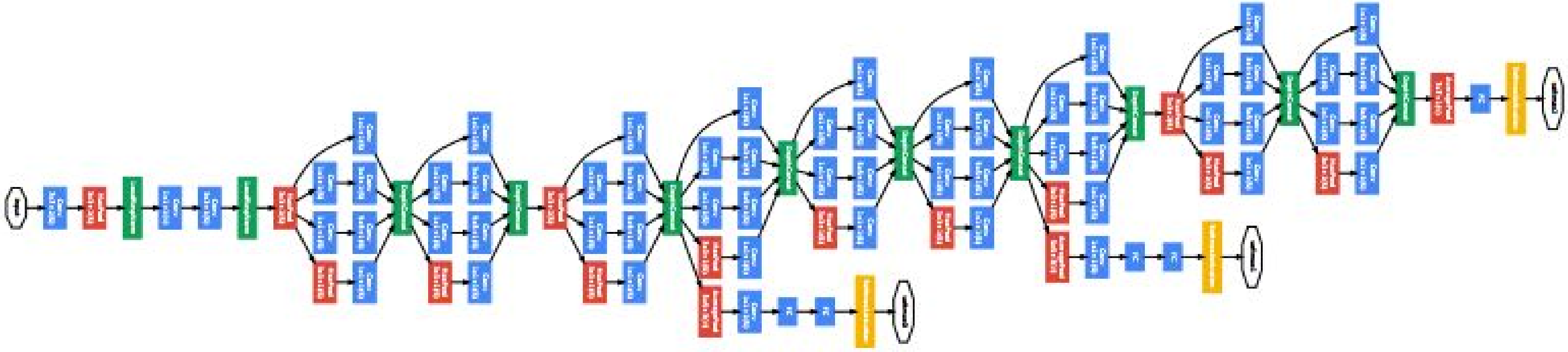
“word2vec” architectures - output in hidden layer



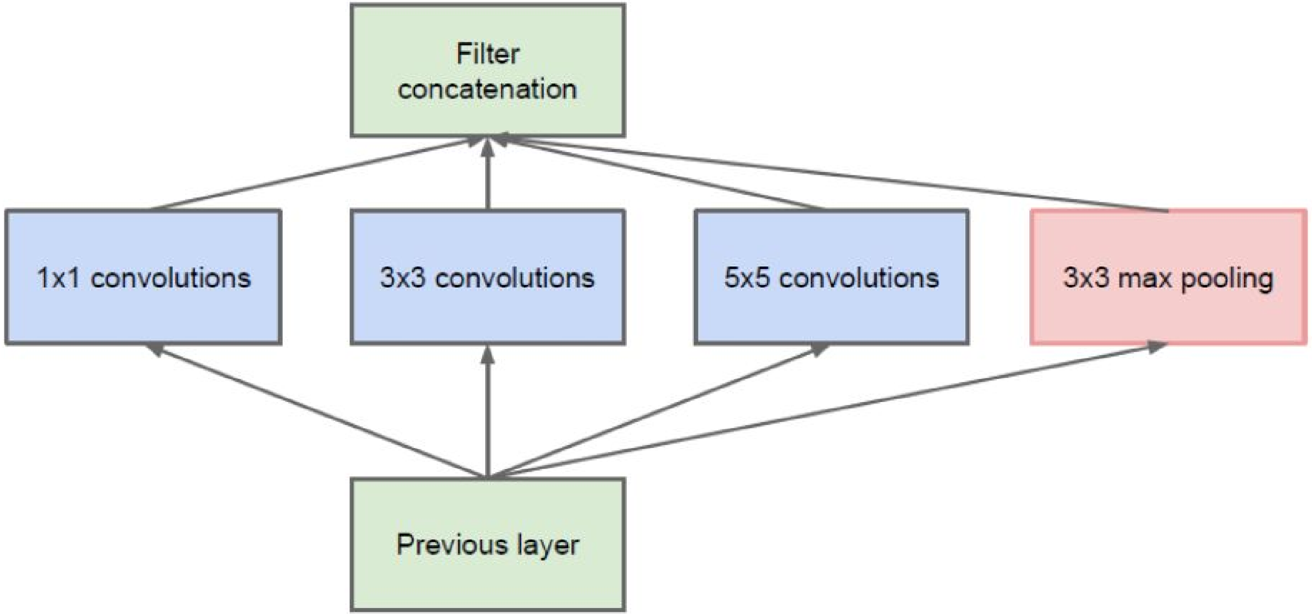
Going deeper with convolutions

Modularised

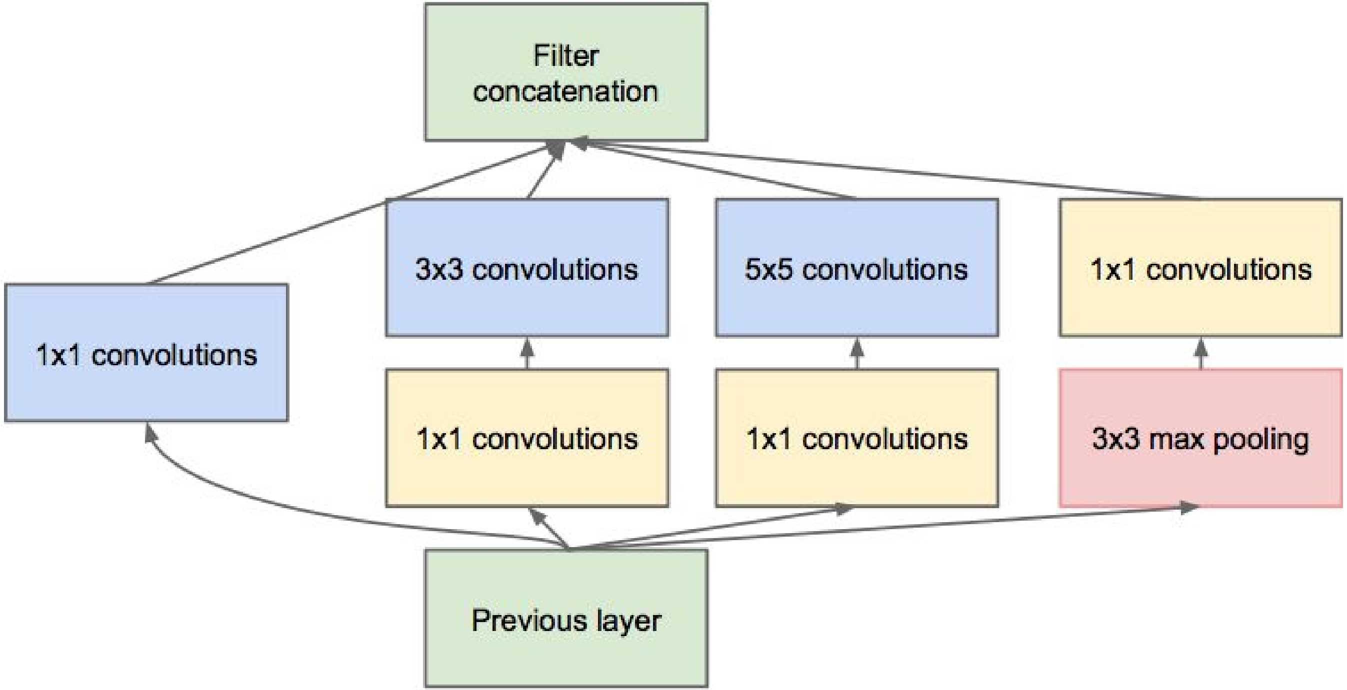
Multiple outputs - controlling hidden layers



Inception module - naive



Inception module - with dimensionality reductions



Rethinking the Inception Architecture

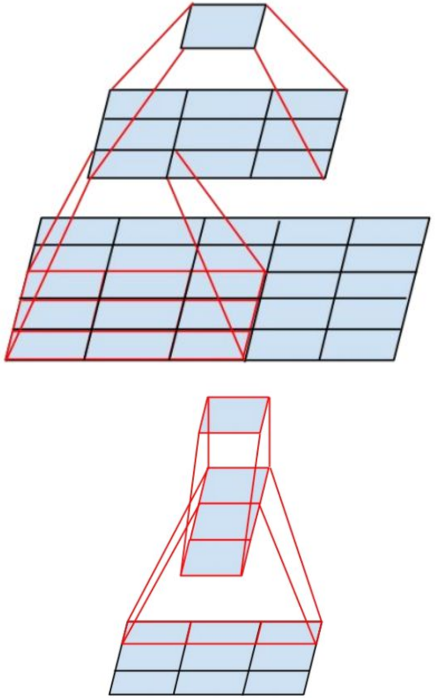
https://arxiv.org/pdf/1512.00567.pdf

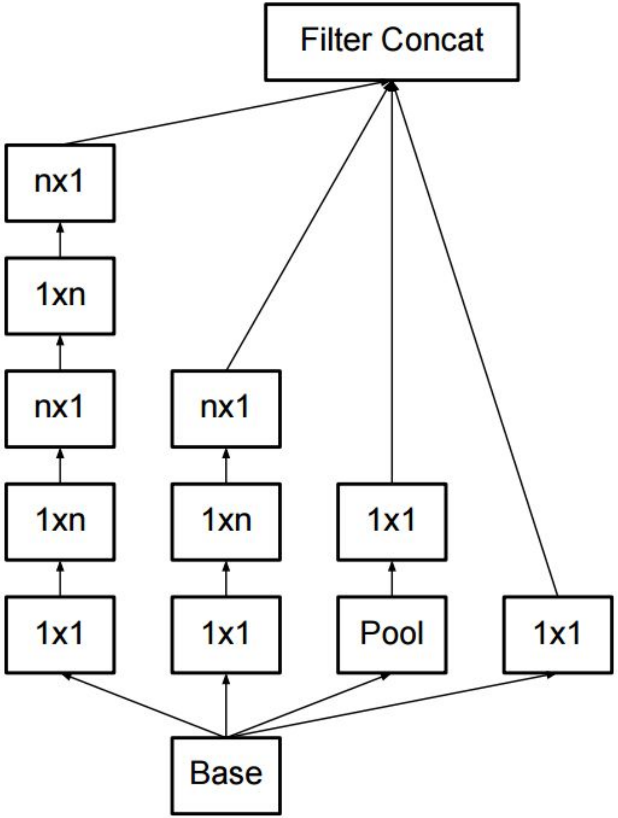
Factorization into smaller and asymmetric convolutions

Grid size reductions

Label-smoothing regularization

Rethinking the Inception Architecture (2)





PolyNet: A Pursuit of Structural Diversity in Very Deep Networks

https://arxiv.org/pdf/1611.05725v1.pdf

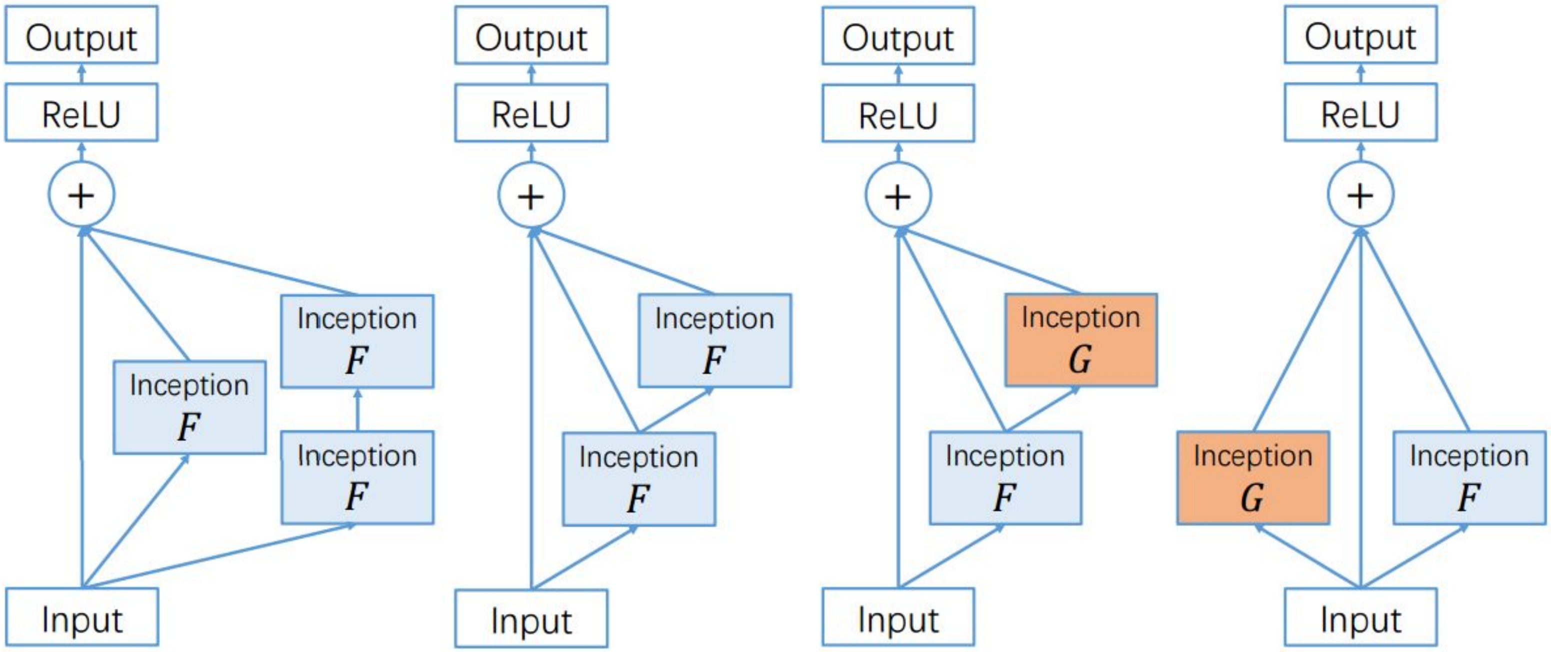
PolyInception modules

Initialization

Residual scaling

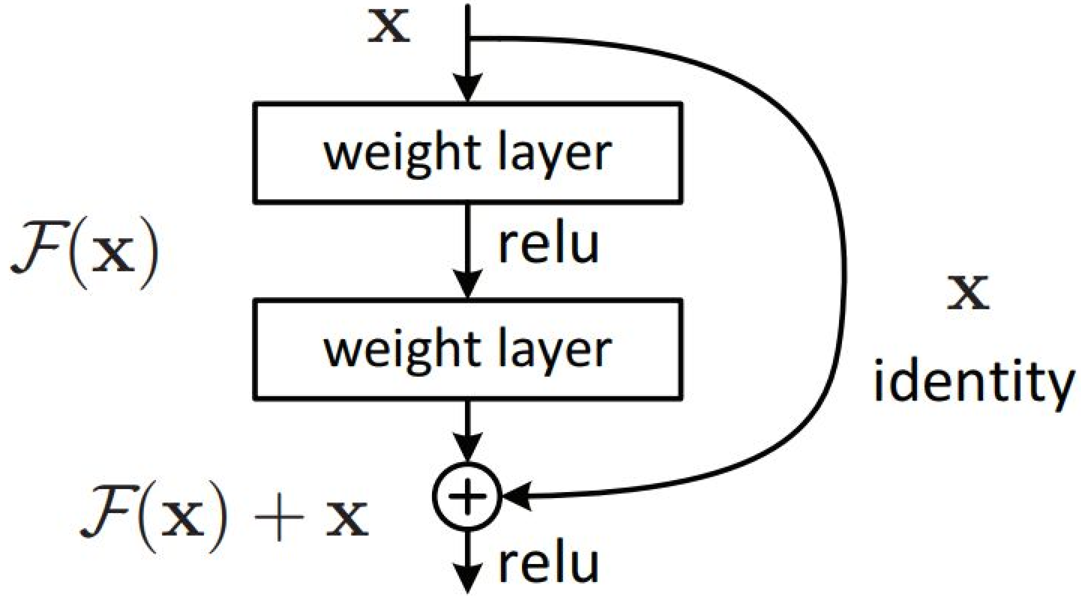
Stochastic paths

PolyNet: A Pursuit of Structural Diversity in Very Deep Networks (2)



(a) *poly-2* (b) *poly-2* (c) *mpoly-2* (d) *2-way*

ResNet - a universal module



Unsupervised keyword extraction

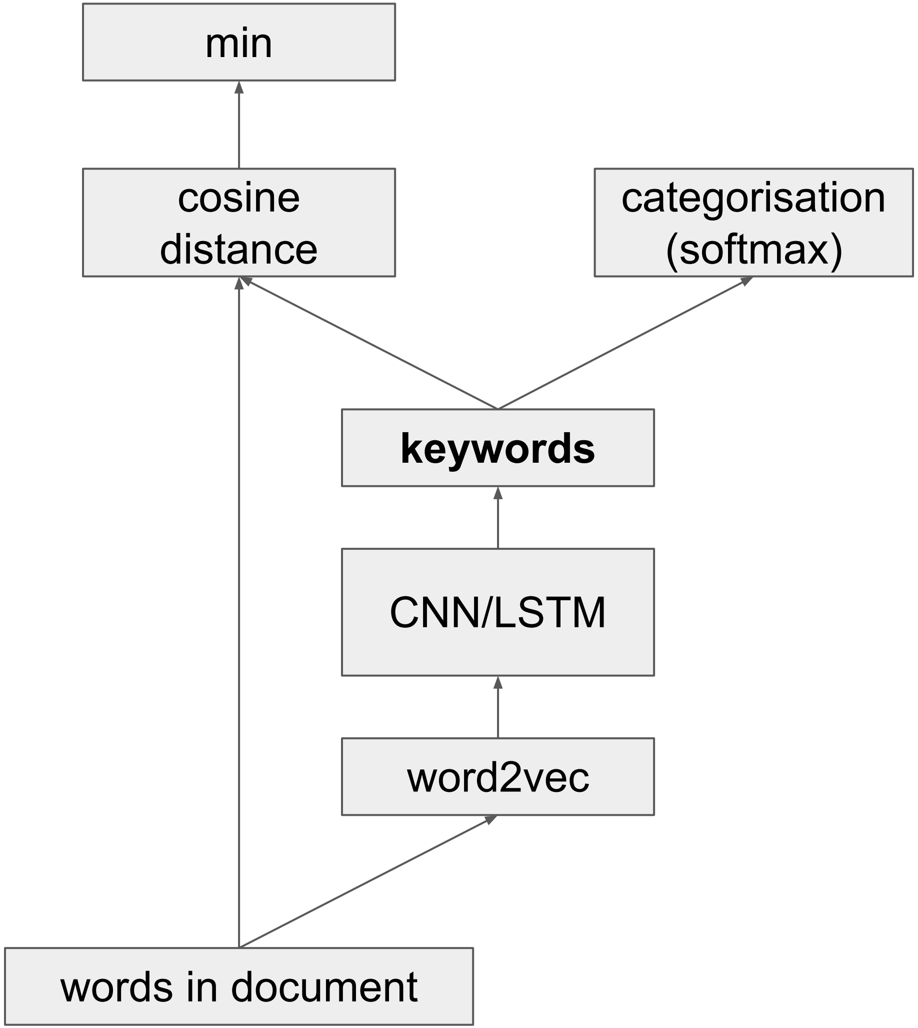
Unsupervised keyword extraction

The real **output in hidden layer**

Two objectives

- We want feature vectors representing **keywords**
- We want **discriminative** feature vectors

Unsupervised keyword extraction



Unsupervised keyword extraction

Works for short texts

- Beware of duplicates

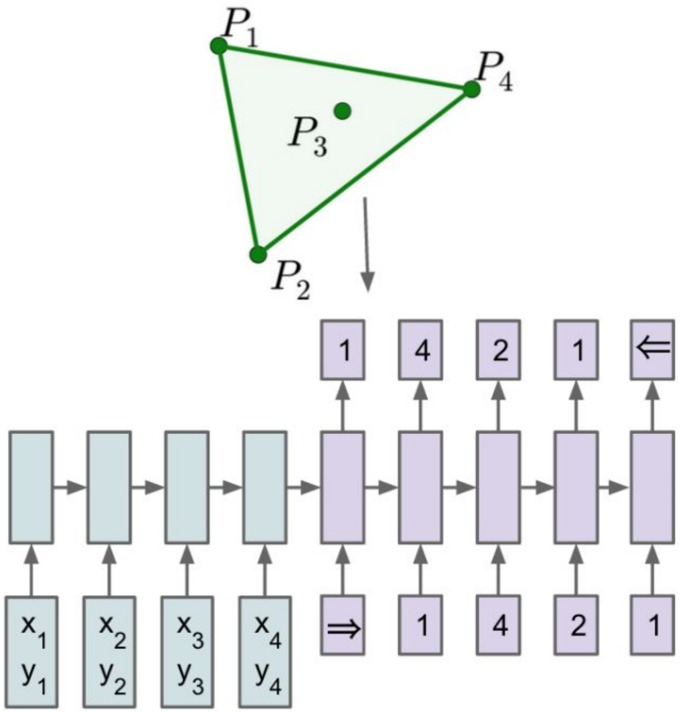
Standard categorisation datasets are challenging

We tried pointer networks

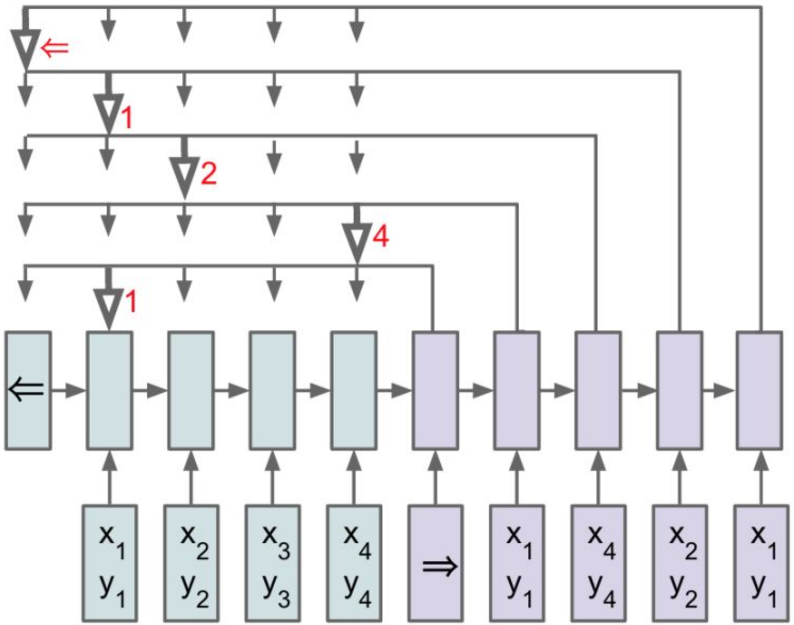
Investigating memory networks

Pointer networks

https://arxiv.org/pdf/1506.03134.pdf



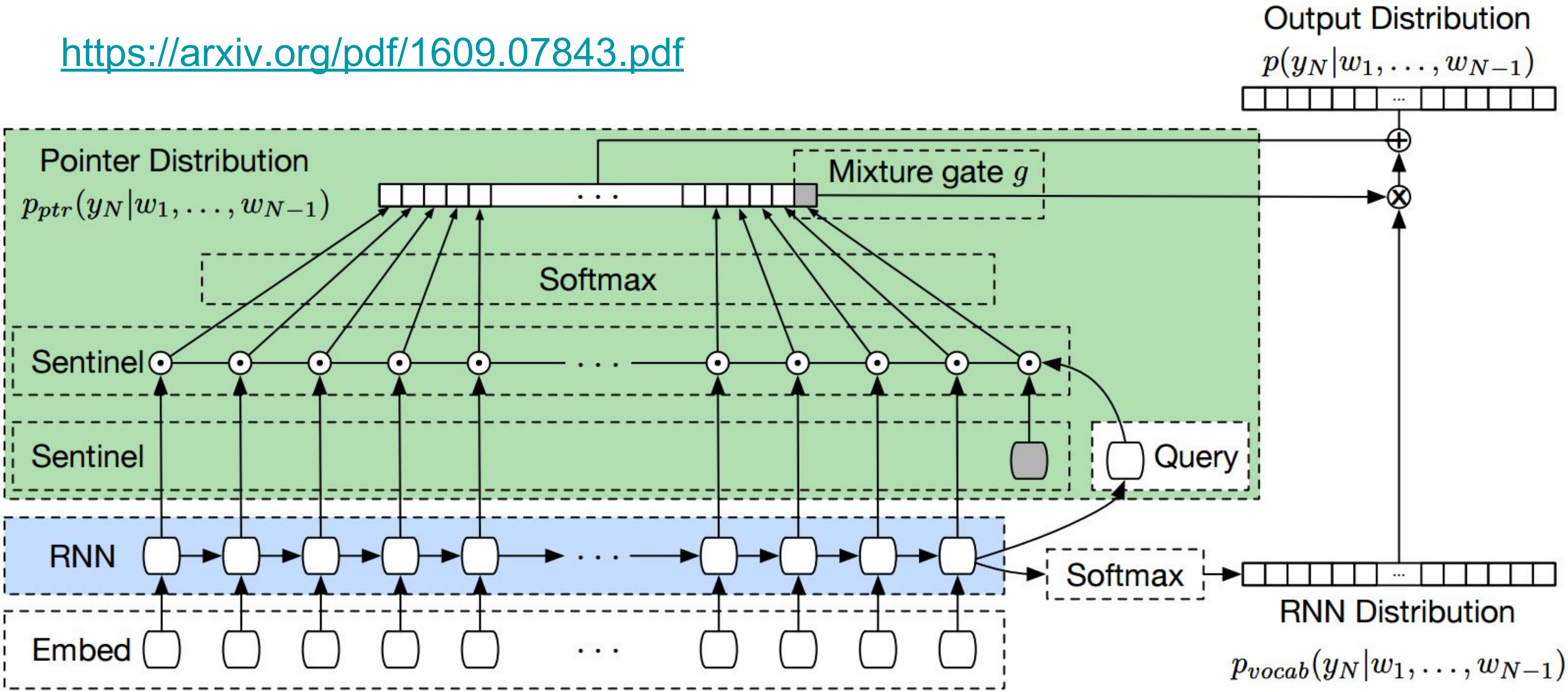
(a) Sequence-to-Sequence



(b) Ptr-Net

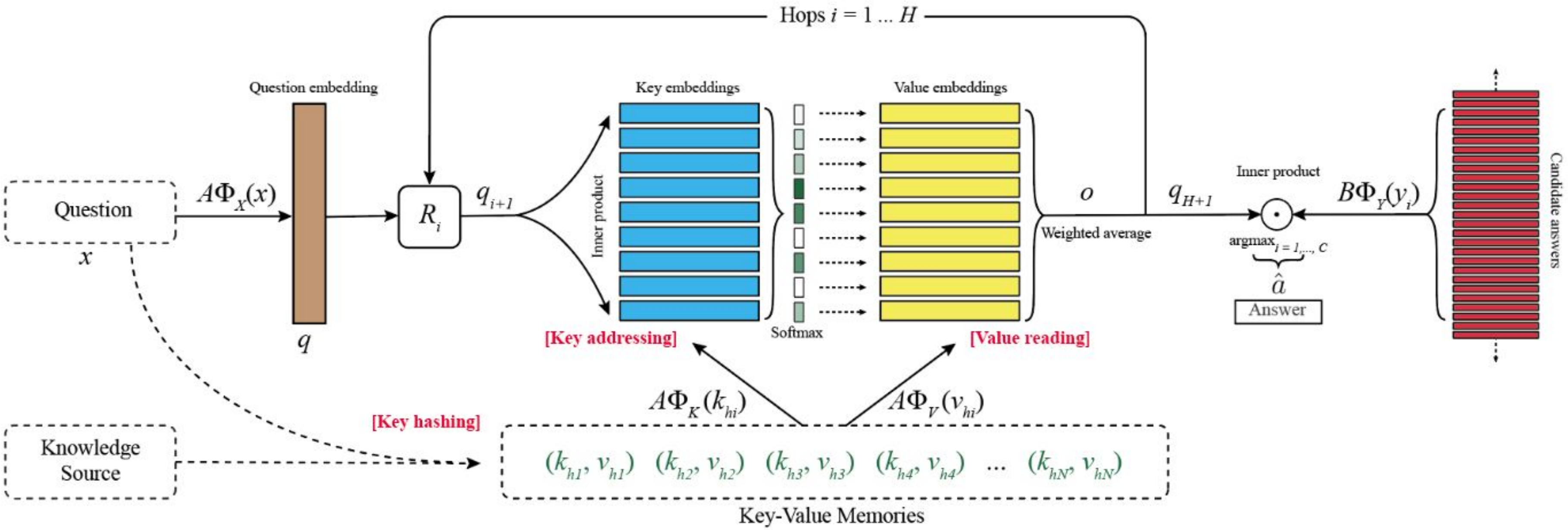
Pointer sentinel mixture model

https://arxiv.org/pdf/1609.07843.pdf



Memory networks

http://www.thespermwhale.com/jaseweston/icml2016/



Summary

Focus on **discriminative** features

Distributed representation

Neural networks + **unsupervised learning**

Neural network architectures as a **white box**