Representation learning for natural language

PhD Defense, Olof Mogren

Chalmers University of Technology

March 23, 2018

Image classification

Imagenet competition

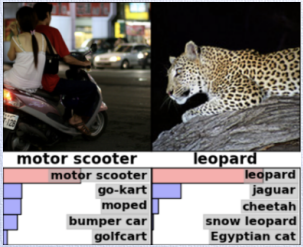
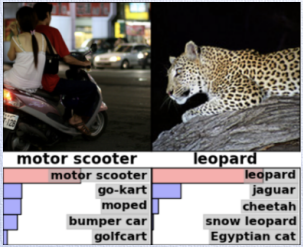
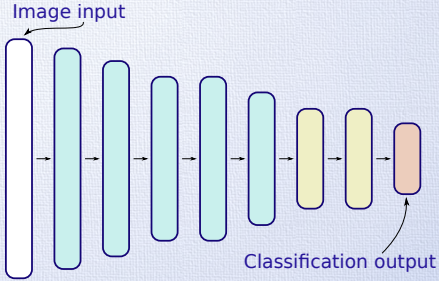


Image classification

Imagenet competition

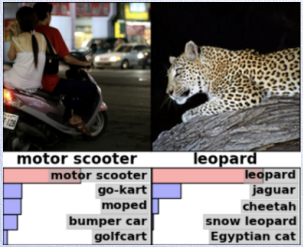


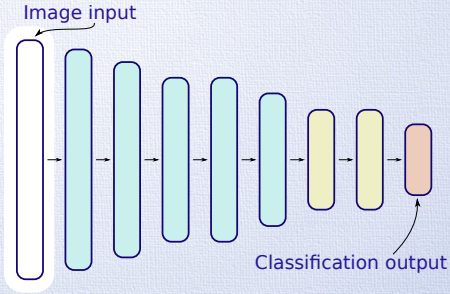


- 2012: Krizhevski, et.al.: Deep neural networks
- Output from each layer: vector

Image classification

Imagenet competition

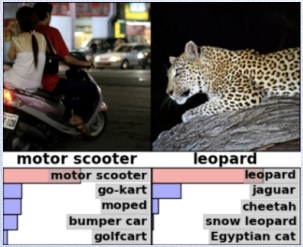


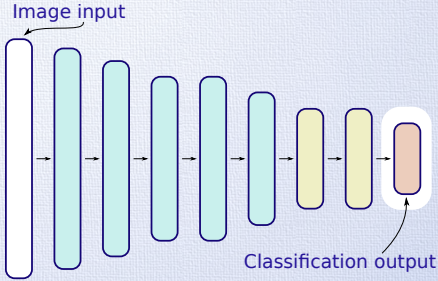


- 2012: Krizhevski, et.al.: Deep neural networks
- Output from each layer: vector

Image classification

Imagenet competition

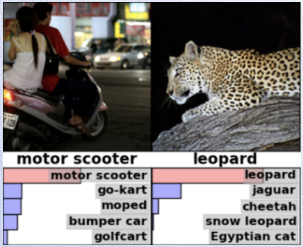


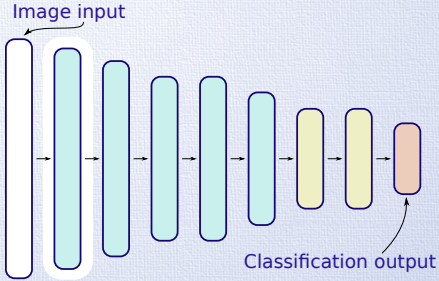


- 2012: Krizhevski, et.al.: Deep neural networks
- Output from each layer: vector

Image classification

Imagenet competition



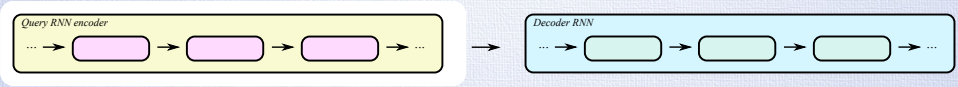


- 2012: Krizhevski, et.al.: Deep neural networks
- Output from each layer: vector

Sequence to sequence learning

- Sutskever, et.al (2014).
  - Encoder -> vector representation -> decoder
  - Basic architecture behind today's machine translation systems

Sequence to sequence learning



- Sutskever, et.al (2014).
  - Encoder -> vector representation -> decoder
  - Basic architecture behind today's machine translation systems

Sequence to sequence learning

- Sutskever, et.al (2014).
  - Encoder -> vector representation -> decoder
  - Basic architecture behind today's machine translation systems

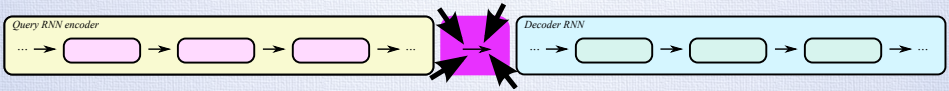
Sequence to sequence learning

- Sutskever, et.al (2014).
  - Encoder -> vector representation -> decoder
  - Basic architecture behind today's machine translation systems

Sequence to sequence learning

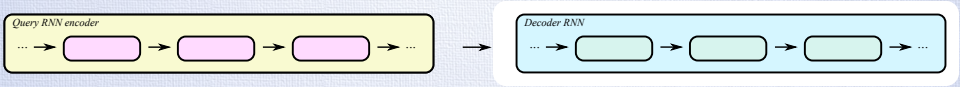
- Sutskever, et.al (2014).
  - Encoder -> vector representation -> decoder
  - Basic architecture behind today's machine translation systems

Sequence to sequence learning



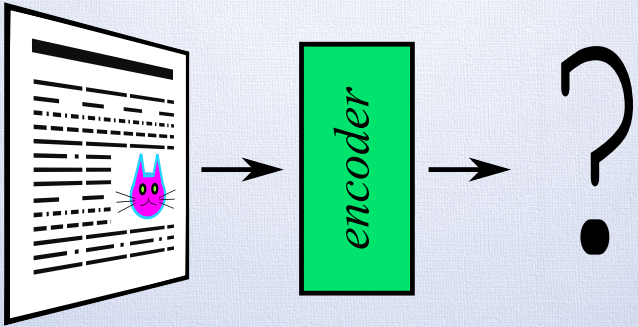
- Sutskever, et.al (2014).
  - Encoder -> vector representation -> decoder
  - Basic architecture behind today's machine translation systems

Sequence to sequence learning

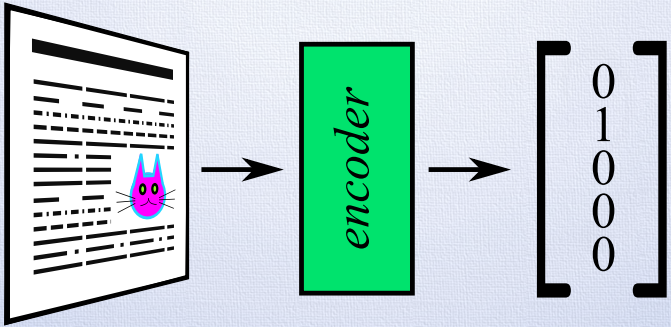


- Sutskever, et.al (2014).
  - Encoder -> vector representation -> decoder
  - Basic architecture behind today's machine translation systems

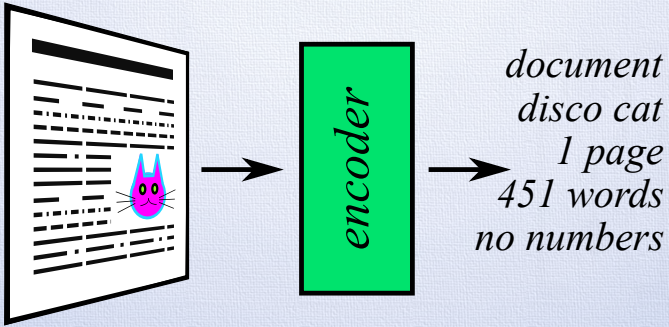
Representations



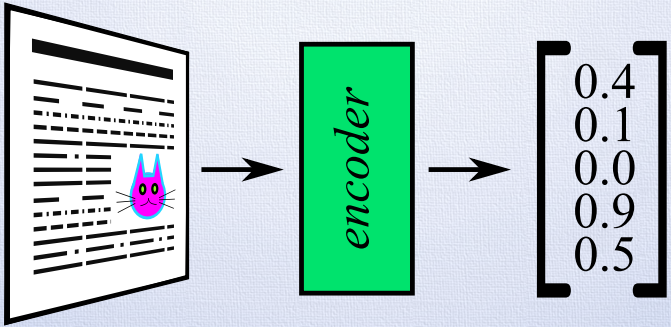
Representations



Representations



Representations



How can we learn the encoder?

How can we best utilize

learned representations for language?

What is encoded in the learned representations?

Outline

- Automatic summarization (**Paper I**, **Paper II**)
- Character RNNs
  - (Recognizing medical terms (**Paper III**))
  - Morphological analogies (**Paper IV**)
- Disentangling underlying factors of variation (**Paper V**)

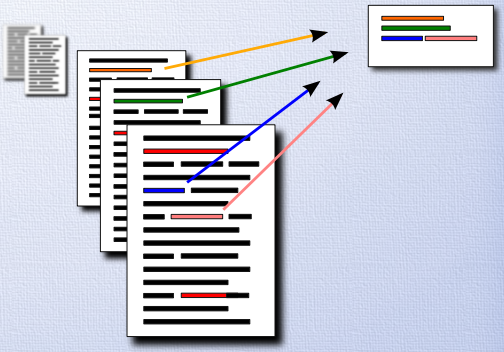
Natural language

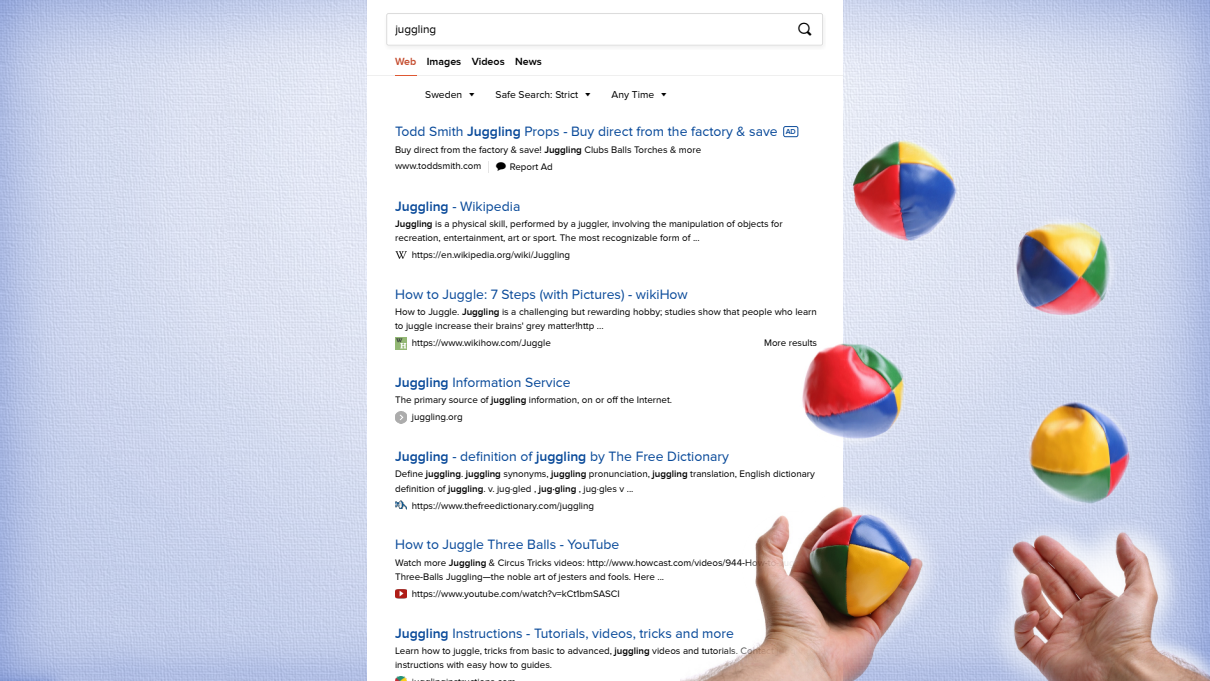
- Evolved through human interactions
- Spoken or **written**

- Collections of documents
- Documents
- Sentences
- Words
- Characters

Automatic document summarization

- Textual summary of input
- Brief
- Cover important topics




juggling
Web
Images
Videos
News
Sweden
Safe Search: Strict
Any Time
Todd Smith Juggling Props - Buy direct from the factory & save
AD
Buy direct from the factory & save! Juggling Clubs Balls Torches & more
www.toddsmith.com
Report Ad
Juggling - Wikipedia
Juggling is a physical skill, performed by a juggler, involving the manipulation of objects for recreation, entertainment, art or sport. The most recognizable form of ...
https://en.wikipedia.org/wiki/Juggling
How to Juggle: 7 Steps (with Pictures) - wikiHow
How to Juggle. Juggling is a challenging but rewarding hobby; studies show that people who learn to juggle increase their brains' grey matter!http ...
https://www.wikihow.com/Juggle
More results
Juggling Information Service
The primary source of juggling information, on or off the Internet.
juggling.org
Juggling - definition of juggling by The Free Dictionary
Define juggling, juggling synonyms, juggling pronunciation, juggling translation, English dictionary definition of juggling. v. jug·gled , jug·gling , jug·gles v ...
https://www.thefreedictionary.com/juggling
How to Juggle Three Balls - YouTube
Watch more Juggling & Circus Tricks videos: http://www.howcast.com/videos/944-Ho
Three-Balls Juggling—the noble art of jesters and fools. Here ...
https://www.youtube.com/watch?v=kCt1bmSASCI
Juggling Instructions - Tutorials, videos, tricks and more
Learn how to juggle, tricks from basic to advanced, juggling videos and tutorials. Co
instructions with easy how to guides.

Summary

Although the etymology of the terms **juggler** and **juggling** in the sense of **manipulating objects for entertainment** originates as far back as the 11th century, the current sense of **to juggle**, meaning **"to continually toss objects in the air and catch them"**, only originates from the late 19th century.

Some writers have called the terms **juggling** and **juggler** a **lexicographical nightmare**, and following the current dictionary definitions have stated "In the twenty-first century the term **juggler** is applied to that kind of **entertainer** who **throws up objects from one hand to another in a continuous rhythmical sequence without dropping them to the floor**.

Input statistics

- Input: 6,016 words, 301 sentences
- One topic: juggling
- 5 different sources: online tutorials and Wikipedia
- Summary length: 103 words

Extractive summarization

- Select sentences from input documents
- Maximize *coverage* and *diversity*
- Sentence similarity: word overlap
- E.g. Lin, Bilmes (2011)

Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Extractive summarization

- Select sentences from input documents
- Maximize *coverage* and *diversity*
- Sentence similarity: word overlap
- E.g. Lin, Bilmes (2011)

Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Extractive summarization

- Select sentences from input documents
- Maximize *coverage* and *diversity*
- Sentence similarity: word overlap
- E.g. Lin, Bilmes (2011)

Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Extractive summarization

- Select sentences from input documents
- Maximize *coverage* and *diversity*
- Sentence similarity: word overlap
- E.g. Lin, Bilmes (2011)

Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Extractive summarization

- Select sentences from input documents
- Maximize *coverage* and *diversity*
- Sentence similarity: word overlap
- E.g. Lin, Bilmes (2011)

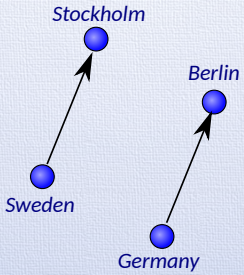
Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Word embeddings

- Trained using co-occurrences in text
- Trained on large datasets



$$v(Stockholm) - v(Sweden) \approx v(Berlin) - v(Germany)$$

king

- (’kings’, 0.71)
- (’queen’, 0.65)
- (’monarch’, 0.64)
- (’crown_prince’, 0.62)

queen

- (’queens’, 0.74)
- (’princess’, 0.71)
- (’king’, 0.65)
- (’monarch’, 0.64)

Stockholm

- (’Stockholm_Sweden’, 0.78)
- (’Helsinki’, 0.75)
- (’Oslo’, 0.72)
- (’Oslo_Norway’, 0.68)

Can neural word embeddings

make automatic summaries more relevant?

Embedding similarity

Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

(+)

Embedding similarity

Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Embedding similarity

Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Embedding similarity

Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Embedding similarity

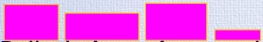
Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Embedding similarity

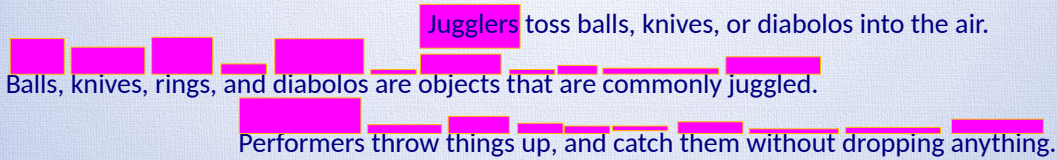
Jugglers toss balls, knives, or diabolos into the air.



Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Embedding similarity



(+)

Summarization with word embeddings

- In **Paper I**, we leverage the power of neural word embeddings
- Sentence embedding: sum of word vectors

$$emb(s) = \sum_{w \in s} w$$

- Compare sentences using cosine similarity:

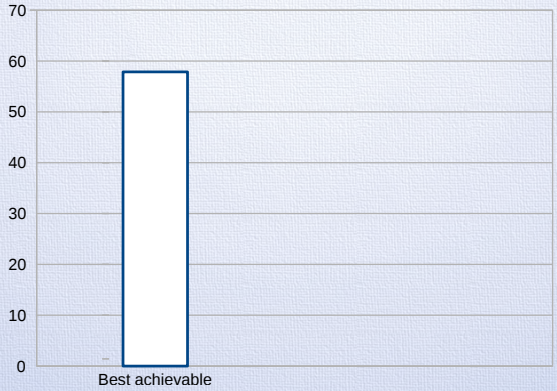
$$sim(s_1, s_2) = \frac{emb(s_1) \cdot emb(s_2)}{|emb(s_1)| \cdot |emb(s_2)|}$$

- Greedy sentence selection (Lin & Bilmes 2011): $\mathcal{L}(S) + \lambda \mathcal{R}(S)$.

***Paper I** is joint work with Mikael Kågebäck, Nina Tahmasebi, and Devdatt Dubhashi.*

Paper I: first work ever to use word embeddings for summarization!

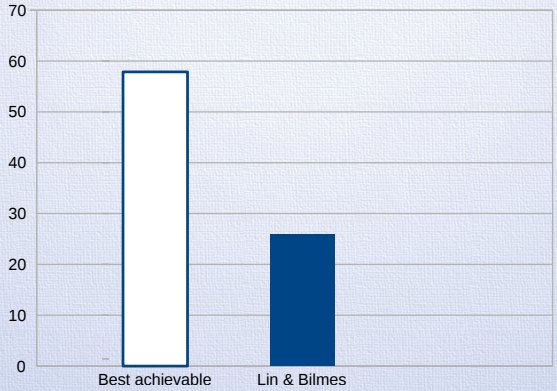
Results: Summarization using word embeddings



ROUGE-1 Recall
OPINOSIS:
Online user reviews

- ROUGE-1 counts word overlaps between reference summaries and model output
- Higher is better!

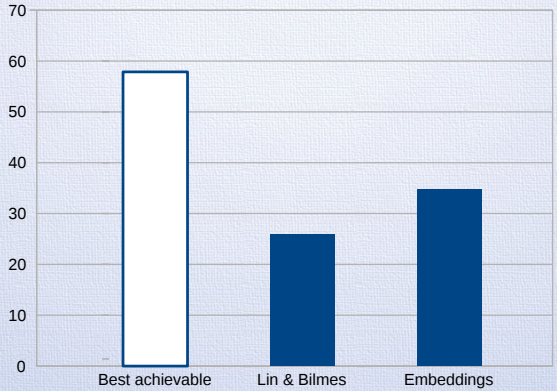
Results: Summarization using word embeddings



ROUGE-1 Recall
OPINOSIS:
Online user reviews

- ROUGE-1 counts word overlaps between reference summaries and model output
- Higher is better!

Results: Summarization using word embeddings



ROUGE-1 Recall
OPINOSIS:
Online user reviews

- ROUGE-1 counts word overlaps between reference summaries and model output
- Higher is better!

Can we capture more aspects of similarity to make even better automatic summaries?

Multiple similarities

Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Multiple similarities

Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Many performers have enjoyed a star billing.

Multiple similarities

Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Many performers have enjoyed a star billing.

The audience loved watching the show.

Multiple similarities

Jugglers toss balls, knives, or diabolos into the air.

Balls, knives, rings, and diabolos are objects that are commonly juggled.

Performers throw things up, and catch them without dropping anything.

Many performers have enjoyed a star billing.

The audience loved watching the show.

I'm not a professional, she said, I just like juggling!

Aggregating similarities

- **Paper II** explores how to capture more aspects of similarity
- Word overlap score:

$$sim^{overlap}(s_i, s_j) = \frac{\sum_{w \in s_i} tf_{w,i} \cdot tf_{w,j} \cdot idf_w^2}{\sqrt{\sum_{w \in s_i} tf_{w,s_i} idf_w^2} \sqrt{\sum_{w \in s_j} tf_{w,s_j} idf_w^2}}$$

- Word embedding score:

$$sim^{embeddings}(s_i, s_j) = \frac{emb(s_1) \cdot emb(s_2)}{||emb(s_1)|| \cdot ||emb(s_2)||}$$

- Sentiment similarity score:

$$sim^{sentiment}(s_i, s_j) = 1 - \left| \frac{sentiment(s_i)}{|s_i|} - \frac{sentiment(s_j)}{|s_j|} \right|$$

where "sentiment" is positive or negative

***Paper II** is joint work with Mikael Kågebäck and Devdatt Dubhashi.*

Multiplicative aggregation

$$sim(s_1, s_2) = \prod_l sim^l(s_1, s_2)$$

sim^l is word overlap scores, word embedding scores and sentiment scores.

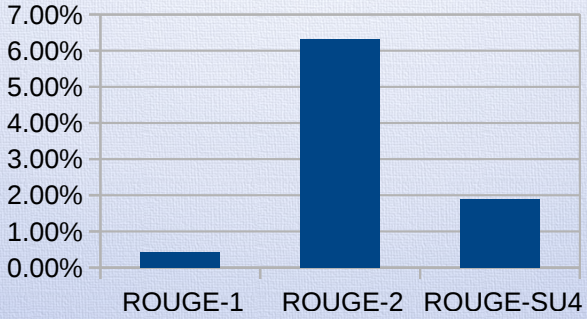
Similar sentences

"Oh, he’s gooood."
"He’s obviously been practising." (0.077)
"And he’s paying attention to what he’s doing!" (0.041)

"Since then, jugglers have been associated with circuses."
"Jugglers commonly feature in circuses, with many performers having enjoyed a star billing." (0.027)

Results: Summarization using aggregated similarity scores

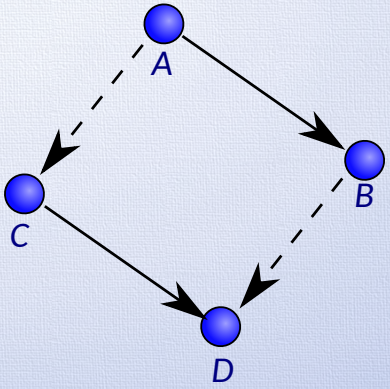
Relative improvement compared to original scores from Lin&Bilmes 2011. Multiplicative combination of two scores: TFIDF and Filtered (N,ADJ). Dataset: DUC 2004



(+) | (ROUGE)

Analogies

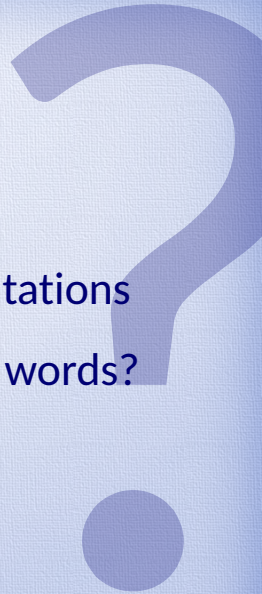
- *A* is to *B* as *C* is to *D*
- Few-shot learning



Morphological analogies

- *“see”* is to *“sees”* as *“eat”* is to what?
- Related tasks: morphological inflection/reinflection (source/target forms explicit):
  character-level RNNs (Kann, Schütze 2016)

Can character-level RNNs compute representations for morphological relations using only demo words?

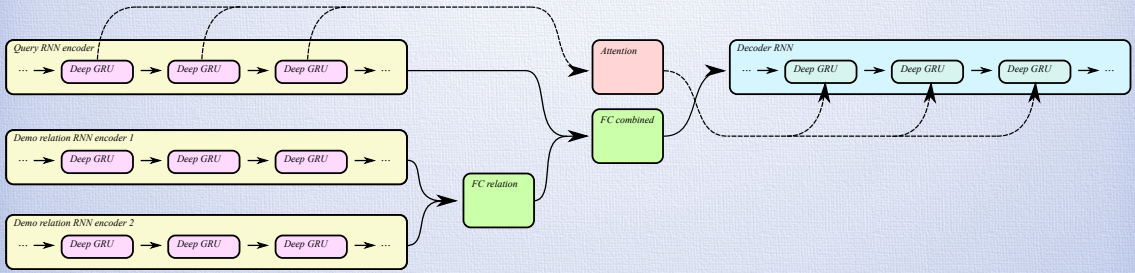


Morphological analogies

- In **Paper IV** we abstract away from character level syntax
- Morphological relations (“*see*” is to “*sees*”)

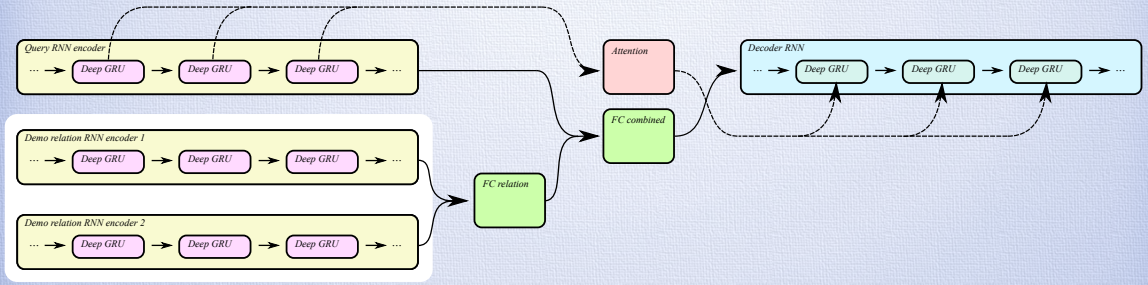
***Paper IV** is joint work with Richard Johansson.*

Proposed model



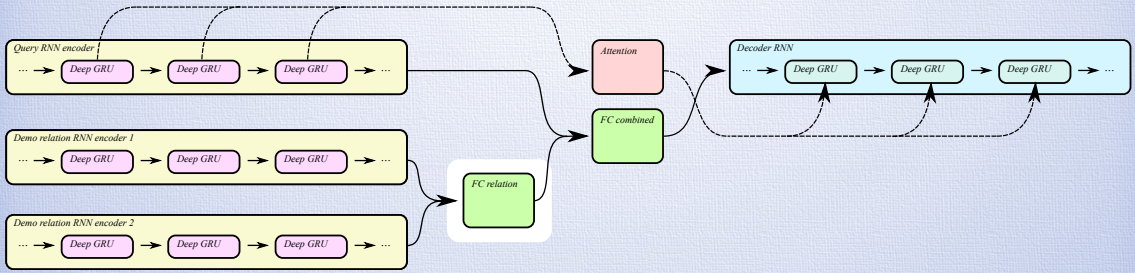
- Demo relation: *“see”*, *“sees”*
- Query word: *“eat”*
- Target word: *“eats”*

Proposed model



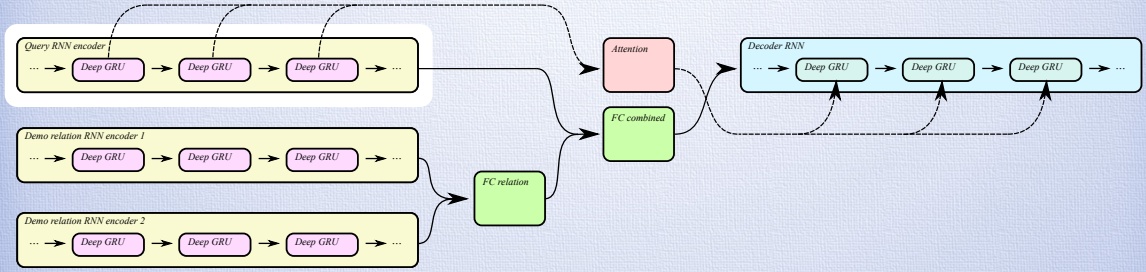
- Demo relation: *“see”*, *“sees”*
- Query word: *“eat”*
- Target word: *“eats”*

Proposed model



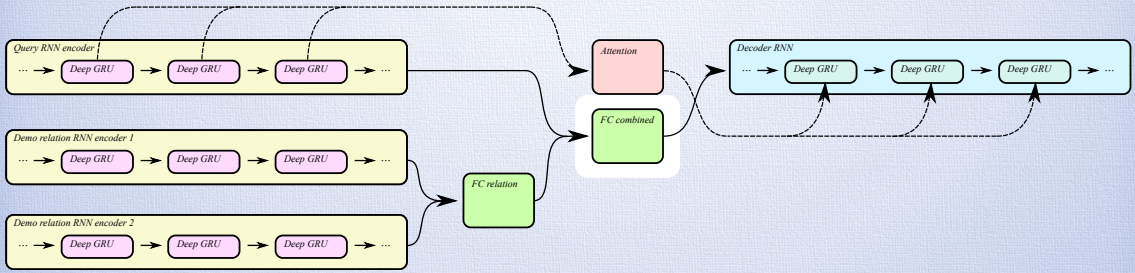
- Demo relation: *“see”*, *“sees”*
- Query word: *“eat”*
- Target word: *“eats”*

Proposed model



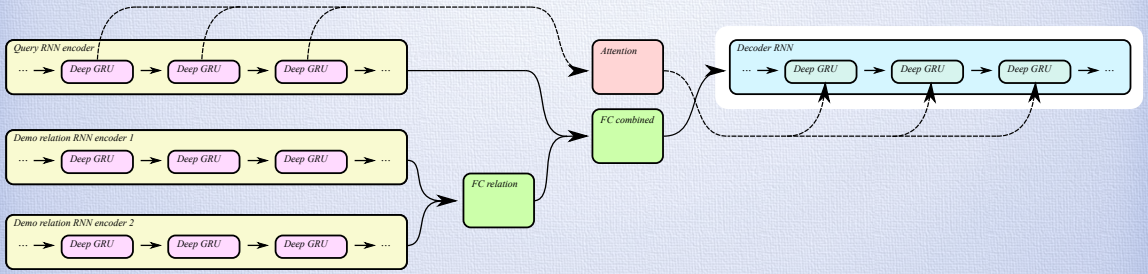
- Demo relation: *“see”*, *“sees”*
- Query word: *“eat”*
- Target word: *“eats”*

Proposed model



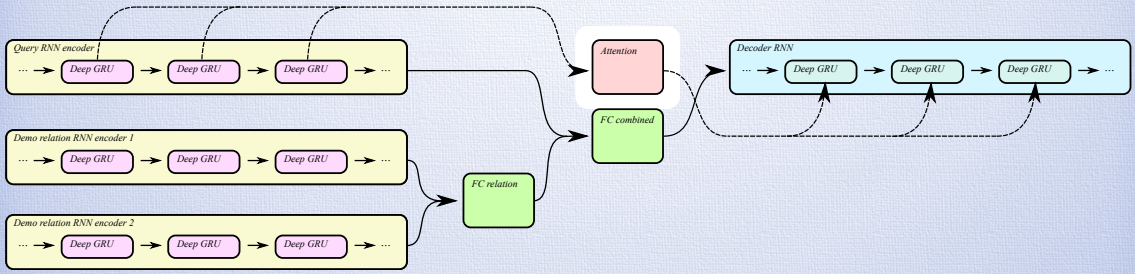
- Demo relation: *“see”*, *“sees”*
- Query word: *“eat”*
- Target word: *“eats”*

Proposed model



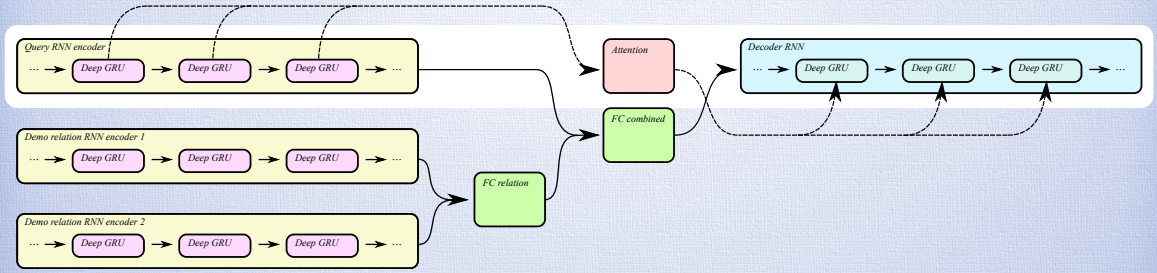
- Demo relation: *“see”*, *“sees”*
- Query word: *“eat”*
- Target word: *“eats”*

Proposed model



- Demo relation: *“see”*, *“sees”*
- Query word: *“eat”*
- Target word: *“eats”*

Proposed model



- Demo relation: *“see”*, *“sees”*
- Query word: *“eat”*
- Target word: *“eats”*

Correct (English):

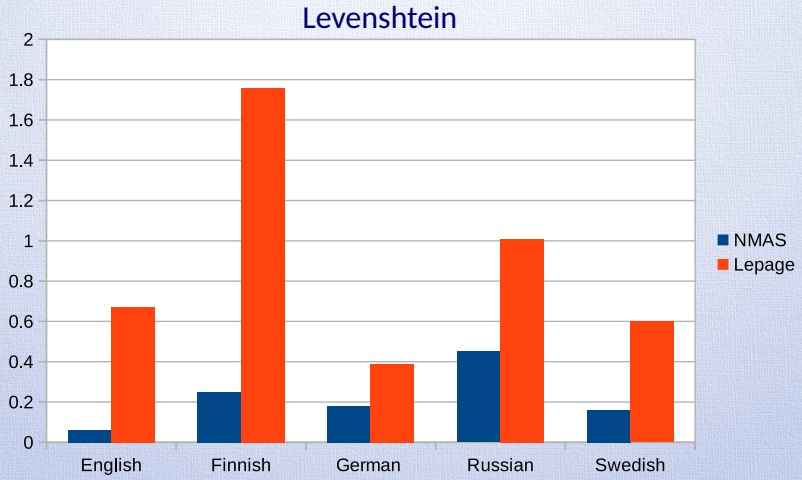
Demo word 1	Demo word 2	Query	Target	Output
misidentify	misidentifies	bottleneck	bottlenecks	bottlenecks
obliterate	obliterated	prig	prigged	prigged
ventilating	ventilates	disorganizing	disorganizes	disorganizes
crank	cranker	freckly	frecklier	frecklier
debauchery	debaucheries	bumptiousness	bumptiousnesses	bumptiousnesses

Incorrect (English):

Demo word 1	Demo word 2	Query	Target	Output
repackage	repackaged	outrun	outran	outrunned
misinformed	misinform	gassed	gas	gass
julep	juleps	catfish	catfish	catfishes
cedar	cedars	midlife	midlives	midlifes
affrays	affray	buzzes	buzz	buzze

Also evaluated on: Finnish, German, Russian, Swedish

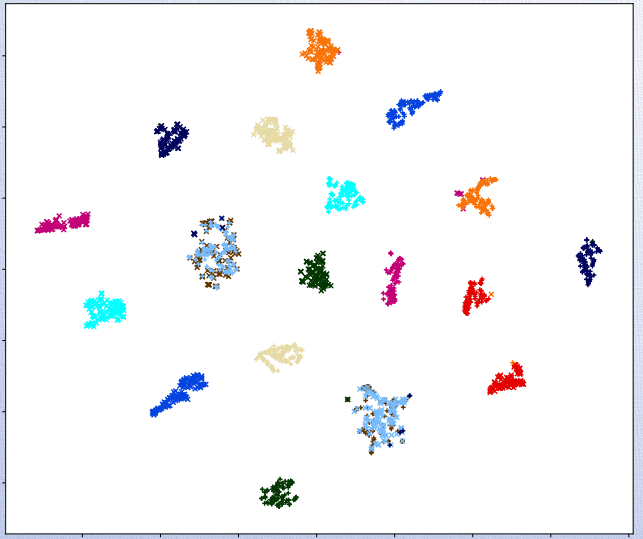
Results: Morphological analogies



(Levenshtein measures character-edits from prediction to correct target. Lower is better).

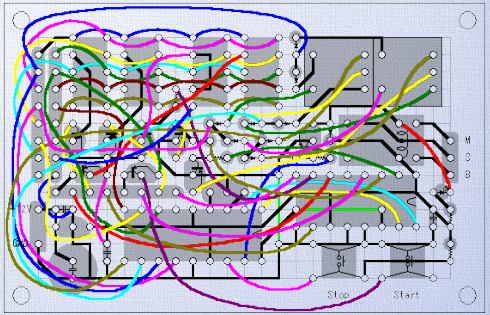
(+)

Relation embeddings



- (+) ADJ: Comparative-Positive
- (x) ADJ: Positive-Comparative
- (+) ADJ: Comparative-Superlative
- (x) ADJ: Superlative-Comparative
- (+) ADJ: Positive-Superlative
- (x) ADJ: Superlative-Positive
- (+) N: Plural-Singular
- (x) N: Singular-Plural
- (+) V: Progressive-3rd Pers. Sg. Presence
- (x) V: 3rd Pers. Sg. Presence-Progressive
- (+) V: Progressive-Past
- (x) V: Past-Progressive
- (+) V: Infinitive-Progressive
- (x) V: Progressive-Infinitive
- (+) V: Infinitive-3rd Pers. Sg. Presence
- (x) V: 3rd Pers. Sg. Presence-Infinitive
- (+) V: Infinitive-Past
- (x) V: Past-Infinitive
- (+) V: 3rd Pers. Sg. Presence-Past
- (x) V: Past-3rd Pers. Sg. Presence

Interpretability



- The learned analogy representations show some interpretability
- 2d visualization: an important tool
- Learned, distributed, representations are not always as easily explained
- *Disentangling underlying factors of variation* important for
  - interpretability
  - performance

Disentangled activations

In **Paper V** we penalize the sample covariance:

$$L_\Sigma = \frac{\sum_{i,j} |\mathcal{C}_{ij}|}{d^2}$$

$$\mathcal{C} = \frac{\sum_{i=1}^{N} (\mathbf{H} - \mathbf{1}_N \bar{\mathbf{h}})^T (\mathbf{H} - \mathbf{1}_N \bar{\mathbf{h}})}{N - 1}.$$

***Paper V** is joint work with Mikael Kågebäck.*

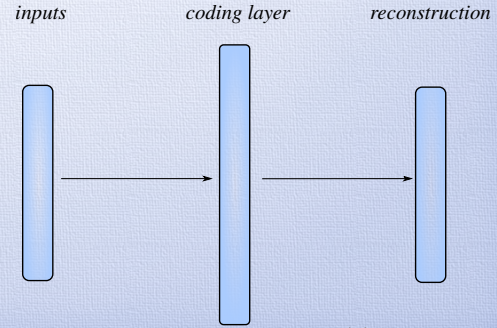
L_Σ

- Computationally inexpensive
- Promotes (linearly) independent representations
- Helps learning interpretable representations
- Helps you find the right layer size

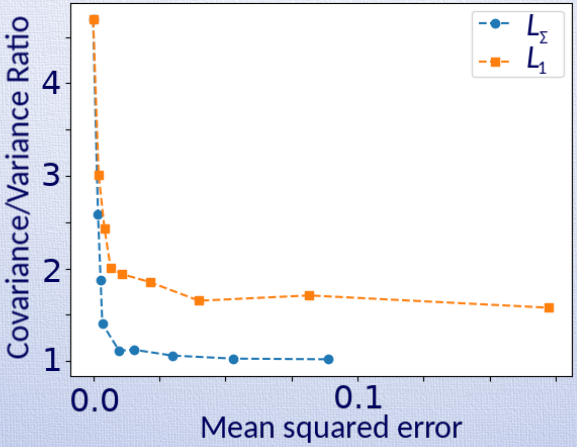
***Paper V** is joint work with Mikael Kågebäck.*

Experiment 1: Dimensionality reduction

- Random vectors z were sampled $\in \mathbb{R}^4$.
- Linear random transformation $x = \Omega z$
- Resulting vector: same information, but projected into $\mathbb{R}^8$
- Shallow autoencoder: 10-dimensional coding layer



Dimensionality reduction



Experiment 2 and 3

- Experiment 2: a model learns to disable unneccessary dimensions
- Experiment 3: real image data

(+)

In conclusion

- Word embeddings benefit extractive summarization **(Paper I)**
- Combining multiple sentence representations improve summarization **(Paper II)**
- Char-RNNs for morphological forms, transform words based on analogies specified as raw character-sequences. **(Paper IV)**
- L_Σ helps to learn disentangled activations. **(Paper V)**

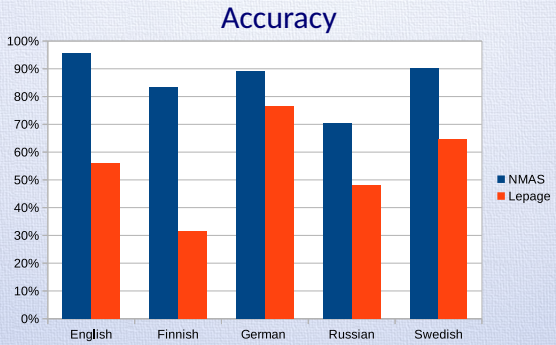
CHALMERS

http://**mogren.one**/

Appendix

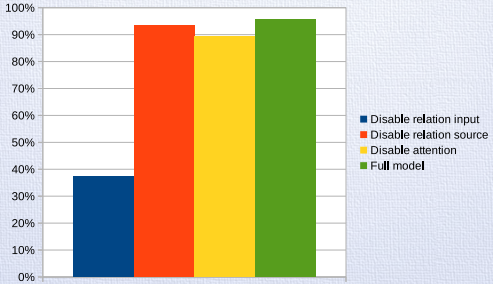
Artificial neural networks, Layers of abstractions, Recurrent neural networks, Microsoft Auto Summarize: The Iliad, Submodular optimization, Summarization evaluation, Overspecified regularized XOR details, CIFAR-10 autoencoder details, Analogies: prediction accuracy, Word embeddings: King Queen, Analogies: component contributions, The cat node, Perceptron, XOR, Convolutional neural networks, Sentiment RNN, Neural machine translation, Caption generation I, Caption generation II, Deep reinforcement learning, Back propagation, LSTM Recognizing medical terms in Swedish health records,

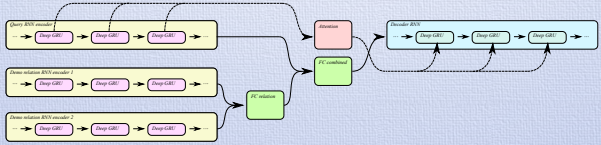
Results: Morphological analogies



(Exact match accuracy. Higher is better).

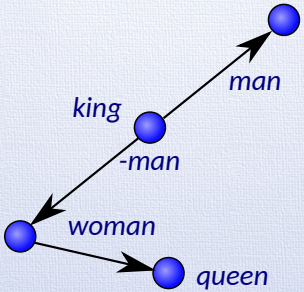
Analogies: component contributions





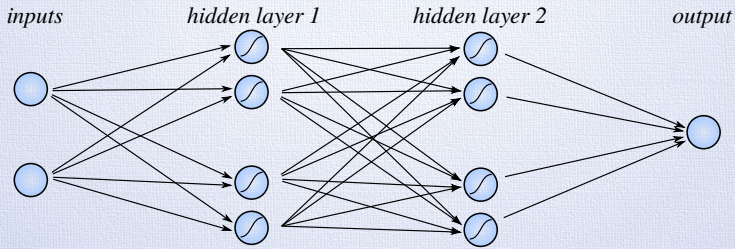
(-)

Word embeddings



$$v(king) - v(man) + v(woman) \approx v(queen)$$

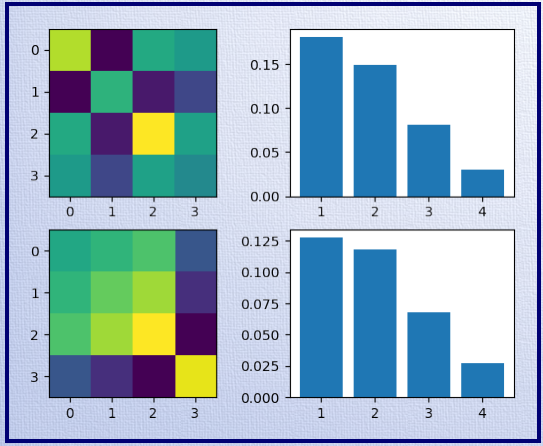
Experiment 2: Overspecified XOR



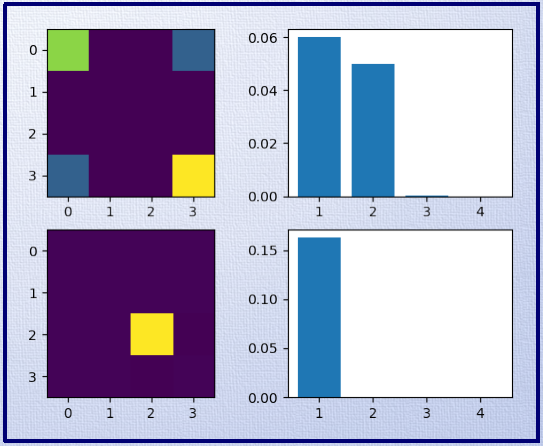
- Multi-layer perceptron (MLP)
- 2 hidden layers
- Each layer: 4 units
- (1 hidden layer, 2 units is sufficient to learn XOR)

Overspecified XOR: disabling activations

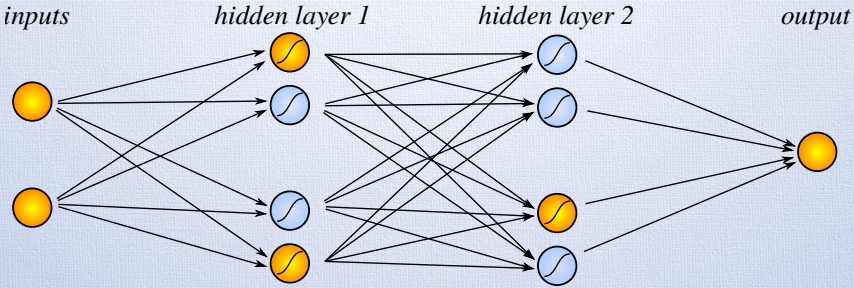
No regularization



L_Σ regularization

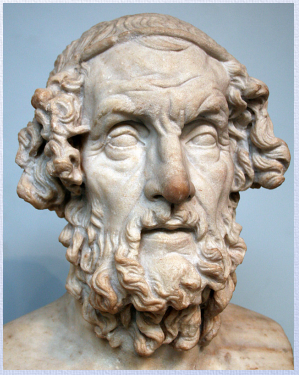


Overspecified XOR: disabling activations 2



Microsoft AutoSummarize

The Iliad by Homer

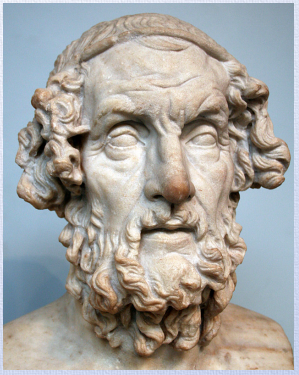


Microsoft AutoSummarize in Word 2008.
From a compilation of summaries of 100 most downloaded copyright free books by Jason Huff.

Microsoft AutoSummarize

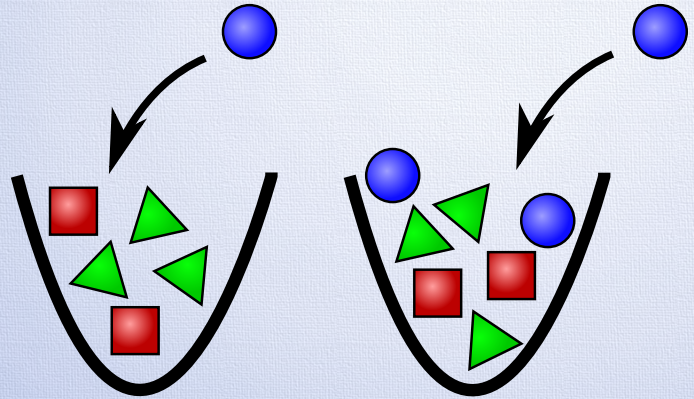
The Iliad by Homer

Gods! Gods! Gods!
“Hector! Gods! Gods!
“Hector! Gods!
“Gods! God!



Microsoft AutoSummarize in Word 2008.
From a compilation of summaries of 100 most downloaded copyright free books by Jason Huff.

Submodular optimization



$\mathcal{F}(bucket) =$
#distinct shapes

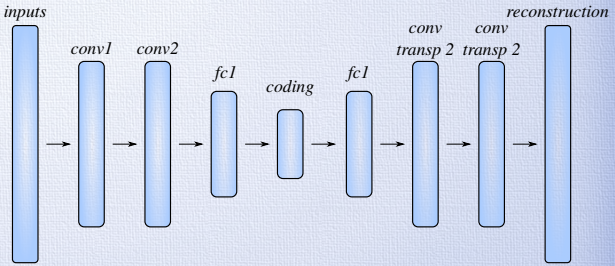
- Objective is a submodular set-function
- $(1 - \frac{1}{e}) \approx 0.632$ approximation

Evaluation

- “Generic” multi-document summarization:
  news articles
- DUC 2004 corpus
- 50 document sets, $\sim$ 10 docs each (149-590 sentences)
- Each document set - one specific news topic
- Gold-standard summaries
- Word overlap evaluation: ROUGE

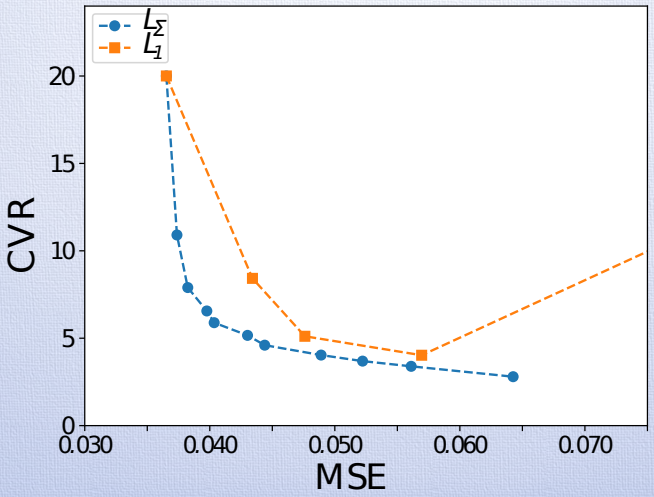
Experiment 3: CIFAR-10 autoencoder

- Convolutional autoencoder
- CIFAR-10: 32x32 color photographs (3072 dimensions)
- Encoder: two convolutional layers, two fully connected layers
- Coding layer with 84 dimensions
- ~500,000 parameters



(-)

CIFAR-10 autoencoder



Correct (Swedish):

Demo word 1	Demo word 2	Query	Target	Output
attackdykare	attackdykare	pillranden	pillrande	pillrande
attackdykare	attackdykare	pillrande	pillranden	pillranden
involverade	involverar	sprack	spricker	spricker
spricker	sprack	iriserar	iriserade	iriserade
sprack	spricker	iriserade	iriserar	iriserar
bensintank	bensintankar	arbetstid	arbetstider	arbetstider
slipkloss	slipklossar	stenåldershjärna	stenåldershjärnor	stenåldershjärno
underfundighet	underfundigheter	attackdykare	attackdykare	attackdykare
attackdykare	attackdykare	pillranden	pillrande	pillrande
attackdykare	attackdykare	pillrande	pillranden	pillranden
pillranden	pillrande	täckglas	täckglas	täckglas
pillrande	pillranden	täckglas	täckglas	täckglas

Back to English

Incorrect (Swedish):

Demo word 1	Demo word 2	Query	Target	Output
involverar	involverade	spricker	sprack	sprickte
otillbörligare	otillbörlig	hårömmare	håröm	hårömm
blekas	blek	likas	lika	lik
misstagas	missta	ledas	leda	led
krylla	kryllade	skräddarsy	skräddarsydde	skräddarsyde
dödförklara	dödförklarade	storgråt	storgrät	storgråtte

Back to English

Correct (German):

Demo word 1	Demo word 2	Query	Target	Output
erwürben	erwerben	abandonnierten	abandonnieren	abandonnieren
erwerben	erwürben	abandonnieren	abandonnierten	abandonnierten
abgeschmackten	abgeschmackteste	zirconiumhaltigen	zirconiumhaltigste	zirconiumhaltigste
abgeschmackteste	abgeschmackten	zirconiumhaltigste	zirconiumhaltigen	zirconiumhaltigen
gebärfähigen	gebärfähigstes	herzhaften	herzhaftestes	herzhaftestes
gebärfähigstes	gebärfähigen	herzhaftestes	herzhaften	herzhaften
Bleichstoffe	Bleichstoff	Transportvorgänge	Transportvorgang	Transportvorgang
Hufschläge	Hufschlag	Knäste	Knast	Knast
Knäste	Knast	Hufschläge	Hufschlag	Hufschlag

Back to English

Incorrect (German):

Demo word 1	Demo word 2	Query	Target	Output
angeschlagenste	angeschlagener	schmälste	schmaler	schmäler
angeschlagener	angeschlagenste	schmaler	schmälste	schmalste
Zugriffe	Zugriff	Sandstürme	Sandsturm	Sandstürm
Zugriff	Zugriffe	Sandsturm	Sandstürme	Sandsturme
Bleichstoff	Bleichstoffe	Transportvorgang	Transportvorgänge	Transportvorgange
Hufschlag	Hufschläge	Knast	Knäste	Knaste
Knast	Knäste	Hufschlag	Hufschläge	Hufschlage
Äste	Astes	Aufwände	Aufwandes	Aufwändes
Astes	Äste	Aufwandes	Aufwände	Aufwande

Back to English

Correct (Finnish):

Demo word 1	Demo word 2	Query	Target	Output
pyöri	pyörimäisillään	seikkaile	seikkailemaisillaan	seikkailemaisillaan
assosiatiivisuudet	assosiatiivisuuksina	mahahapot	mahahappoina	mahahappoina
viestintäsatelliitit	viestintäsatelliittiin	ahneudet	ahneuteen	ahneuteen
paikossa	paikot	tasapuolisuudessa	tasapuolisuudet	tasapuolisuudet
seminormiin	seminormit	viestintäsatelliittiin	viestintäsatelliitit	viestintäsatelliitit
päälaenlohkoin	päälaenlohkot	hillopurkein	hillopurkit	hillopurkit
siamankeineen	siamankeina	trieereineen	trieereinä	trieereinä
falangeina	falangina	timantteina	timanttina	timanttina
glaukofaaneina	glaukofaani	mukavuusavioliittoina	mukavuusavioliitto	mukavuusavioliitto

Back to English

Incorrect (Finnish):

Demo word 1	Demo word 2	Query	Target	Output
konjugoimaisillaan	konjugoi	leimautumaisillaan	leimautui	leimaudui
augiitteina	augiitit	meininkeinä	meiningit	meininkit
päätueksi	päätukina	bulevardisportiksi	bulevardisportteina	bulevardisporttina
ortodoksikirkoksi	ortodoksikirkkoina	päätueksi	päätukina	päätuina
anodeissa	anodina	betonirampeissa	betoniramppina	betonirampina
puhelinlaskulta	puhelinlaskuna	tunnolta	tuntona	tunnona

Back to English

Correct (Russian):

Demo word 1	Demo word 2	Query	Target	Output
шлифуете	шлифуем	медлите	медлим	медлим
терпя	терпите	предпринимая	предпринимаете	предпринимаете
укоряете	укоряешь	спешиваетесь	спешиваешься	спешиваешься
путешествовали	путешествовать	проверяли	проверять	проверять
валяться	валяюсь	подлаживать	подлаживаю	подлаживаю
валяюсь	валяться	подлаживаю	подлаживать	подлаживать
собираюця	собирайся	разумеют	разумей	разумей
благодари	благодарили	покрывай	покрывали	покрывали

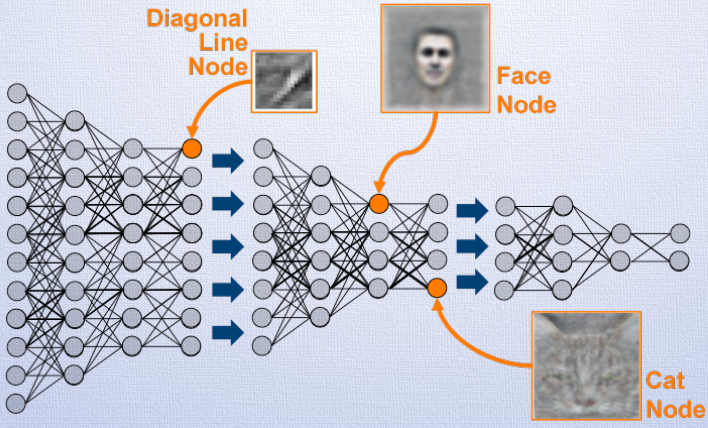
Back to English

Incorrect (Russian):

Demo word 1	Demo word 2	Query	Target	Output
подлаживать	подлаживаю	глядеть	гляжу	гляду
подлаживаю	подлаживать	гляжу	глядеть	гляжать
предупреждай	предупреждаешь	лечи	лечишь	лечёшь
подбирает	подбирай	пялица	пялься	пялись
испустя	испусти	проникая	проникай	проникить
делим	делящий	высимся	высящийся	высившийся
отплываем	отплывающий	делим	делящий	делищий

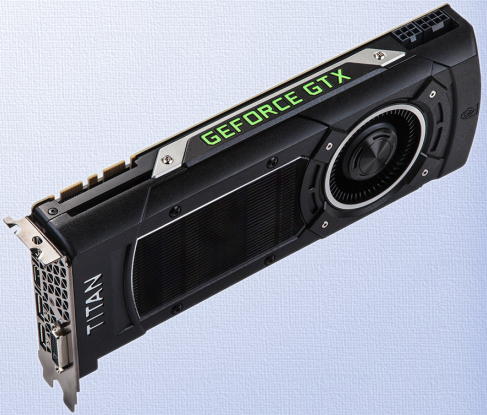
Back to English

Levels of abstractions: The cat node



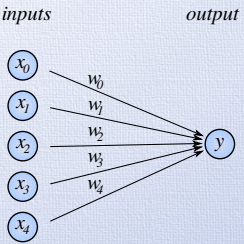
Why the success?

- Progress in model design and algorithms
- GPUs
- Interest from researchers and industry
- Practical use (See previous slide)
  Real applications at Google, Facebook, Tesla, Microsoft, Apple, and others!



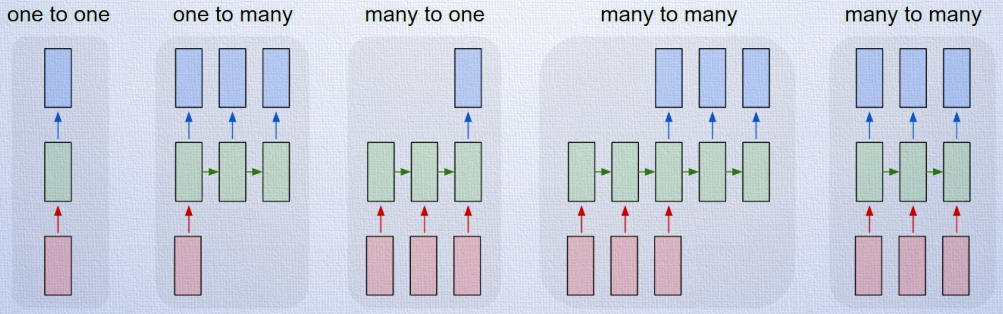
Perceptron

- 1957, Frank Rosenblatt
- $a = \sum_i w_i x_i + b$
- $y = I(a > 0)$
- Linear (binary) classification of inputs
- Can not learn any non-linear function (e.g. exclusive or, XOR)



(details)

Neural sequence models



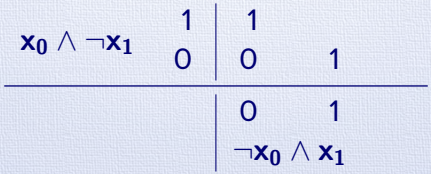
Andrej Karpathy

details

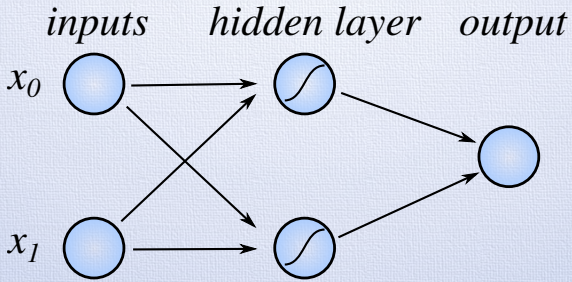
Modelling XOR 1/3

x_0	1	1	0
	0	0	1
		0	1
		x_1	

Modelling XOR, 2/3



Modelling XOR, 3/3



Convolutional neural networks

- Convolution filters; patches matching parts of input
- Successful e.g. for image recognition

Convolutional neural networks

- Convolution filters; patches matching parts of input
- Successful e.g. for image recognition

Convolutional neural networks

- Convolution filters; patches matching parts of input
- Successful e.g. for image recognition

Convolutional neural networks

- Convolution filters; patches matching parts of input
- Successful e.g. for image recognition

Convolutional neural networks

- Convolution filters; patches matching parts of input
- Successful e.g. for image recognition

Convolutional neural networks

- Convolution filters; patches matching parts of input
- Successful e.g. for image recognition

Convolutional neural networks

- Convolution filters; patches matching parts of input
- Successful e.g. for image recognition

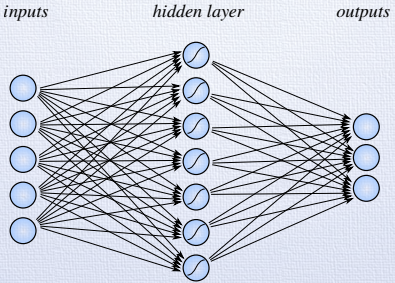
Convolutional neural networks

- Convolution filters; patches matching parts of input
- Successful e.g. for image recognition

Convolutional neural networks

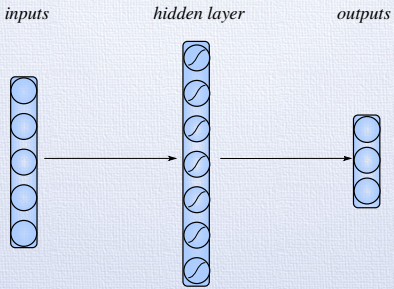
- Convolution filters; patches matching parts of input
- Successful e.g. for image recognition

End-to-end modelling



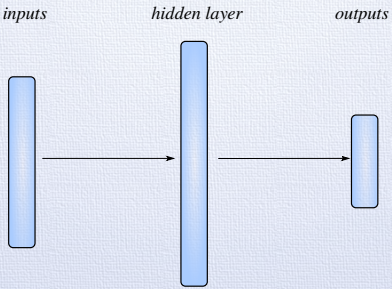
- Linear transformation: $\mathbf{a} = W\mathbf{x} + \mathbf{b}$
- Non-linear activation: $\mathbf{h} = g(\mathbf{a})$
- Output vector **h** can be viewed and used as a data representation

End-to-end modelling



- Linear transformation: $\mathbf{a} = W\mathbf{x} + \mathbf{b}$
- Non-linear activation: $\mathbf{h} = g(\mathbf{a})$
- Output vector $\mathbf{h}$ can be viewed and used as a data representation

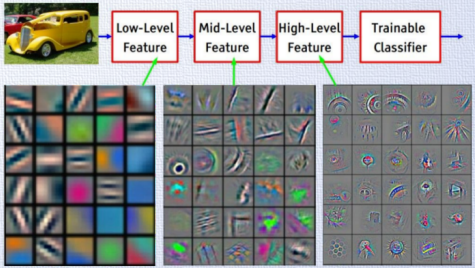
End-to-end modelling



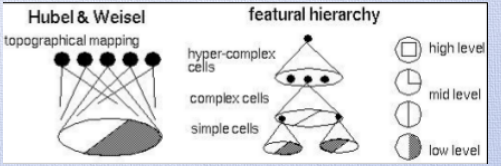
- Linear transformation: $\mathbf{a} = W\mathbf{x} + \mathbf{b}$
- Non-linear activation: $\mathbf{h} = g(\mathbf{a})$
- Output vector **h** can be viewed and used as a data representation

Levels of abstractions

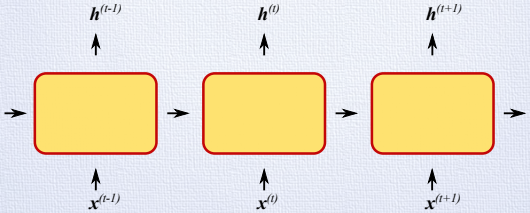
- Deep neural networks learn to compute hierarchies of representations
- Feature engineering not neccessary



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]



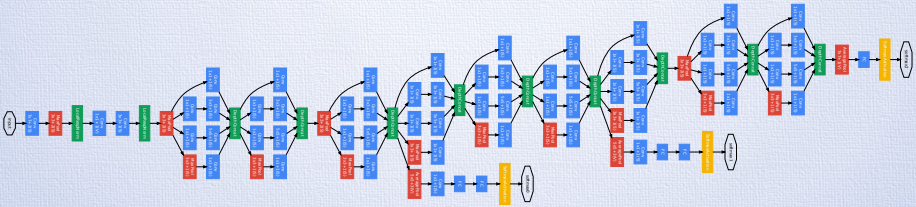
Recurrent neural networks (RNNs)



$$\mathbf{h}^{(t)} = \tanh(W\mathbf{x}^{(t)} + U\mathbf{h}^{(t-1)} + \mathbf{b})$$

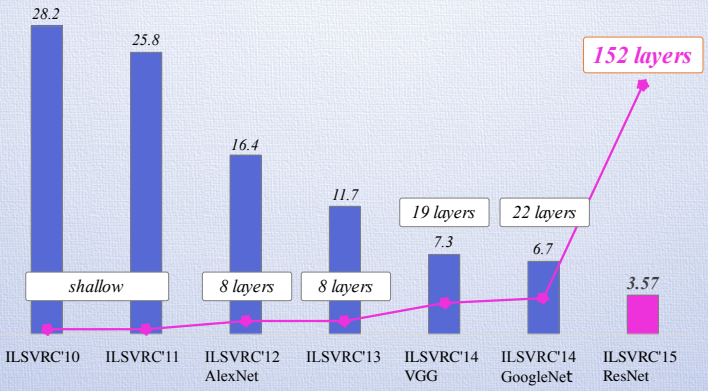
- Fixed-size vector representation for sequences
- Language input or output
- Words or characters

Deep learning for image processing



- Deeper and deeper
- 2014: GoogLeNet; 22 layers (illustration)
- 2015: Residual Nets; 152 layers
- “Surpassed” human performance in 2015

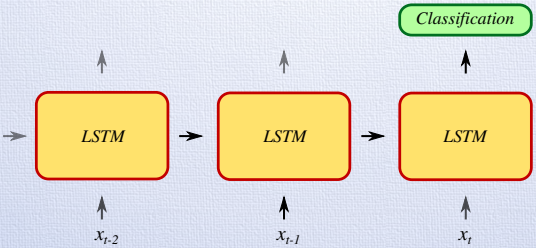
Deep learning for image processing



ImageNet Classification top-5 error (%)

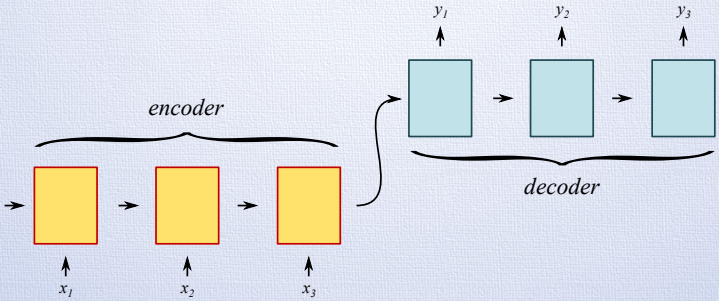
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Sentiment analysis



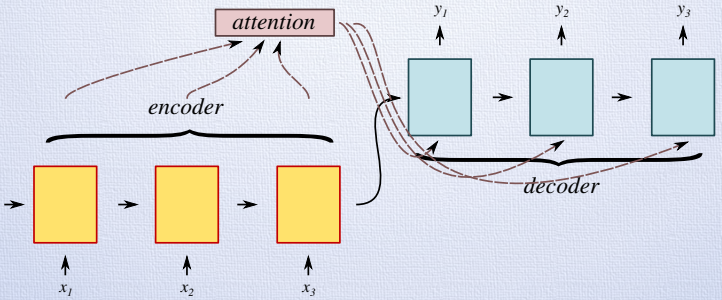
- Binary sequence classification

Machine translation



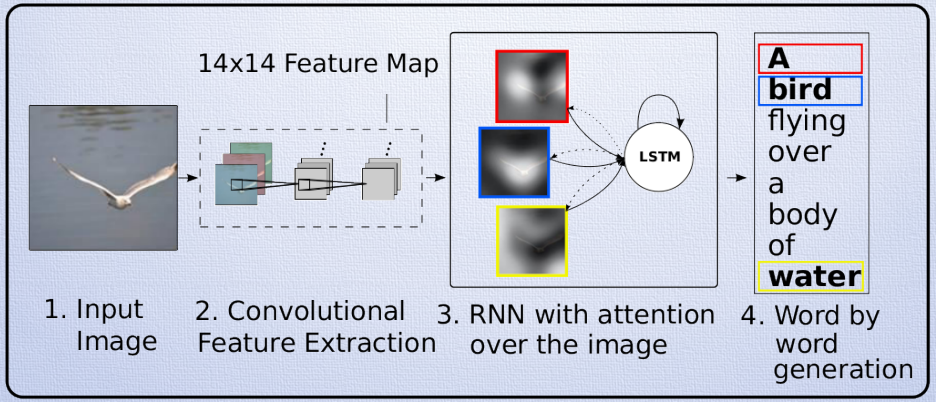
- “Sequence-to-sequence” learning
- Attention models

Machine translation



- “Sequence-to-sequence” learning
- Attention models

Caption generation



(+)

Deep reinforcement learning

- Learning a policy using an *infrequent* reward signal
- Deep Q-Learning: Model the “action-value” function
- Atari games.
- Alpha Go
- Autonomous driving

Q-Learning playing Atari Break-Out

Online Offline back to reinforcement learning

Attention visualization

A woman is throwing a frisbee in a park.

back to caption introduction

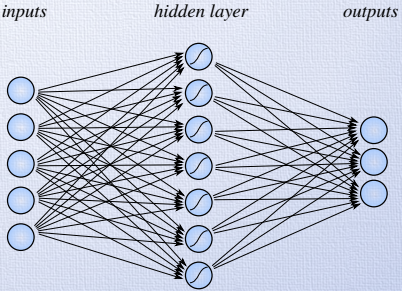
Attention visualization

A stop sign is on a road with a mountain in the background.

back to caption introduction

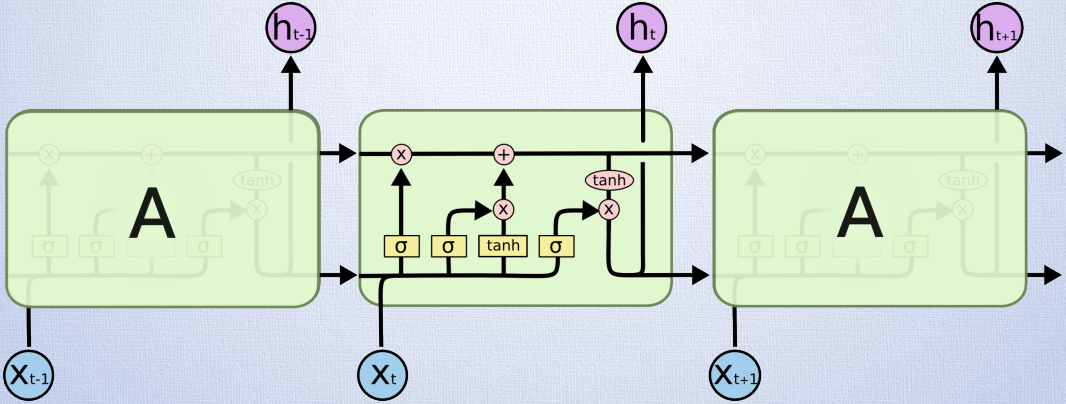
Learning

1. Forward pass (function application(s))
2. Compute error for output
3. Compute gradients (backpropagation)
   derivative of stacked layers: chain rule
4. Update weights (a small step)
   (minibatch stochastic gradient descent)



back to deep learning

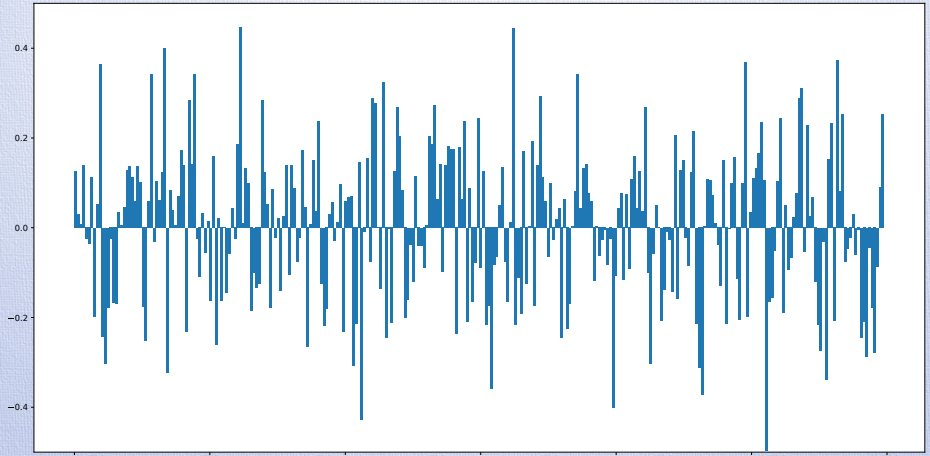
LSTM



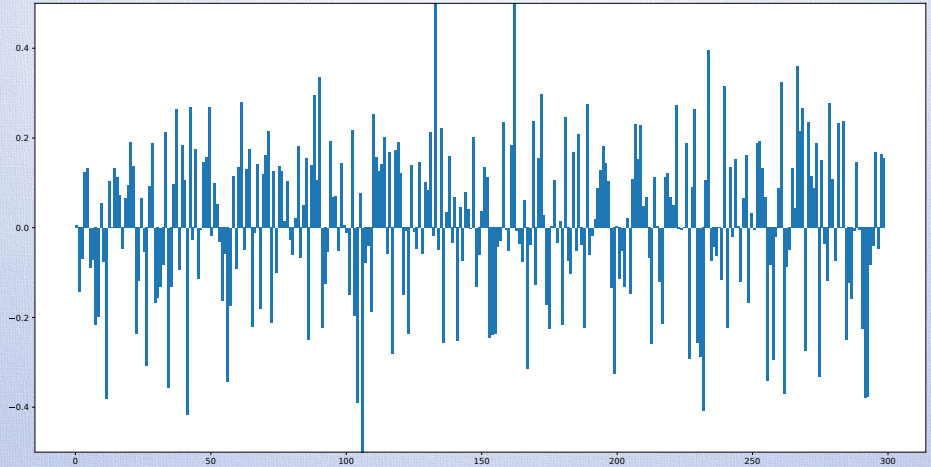
Christopher Olah

back to proposed model

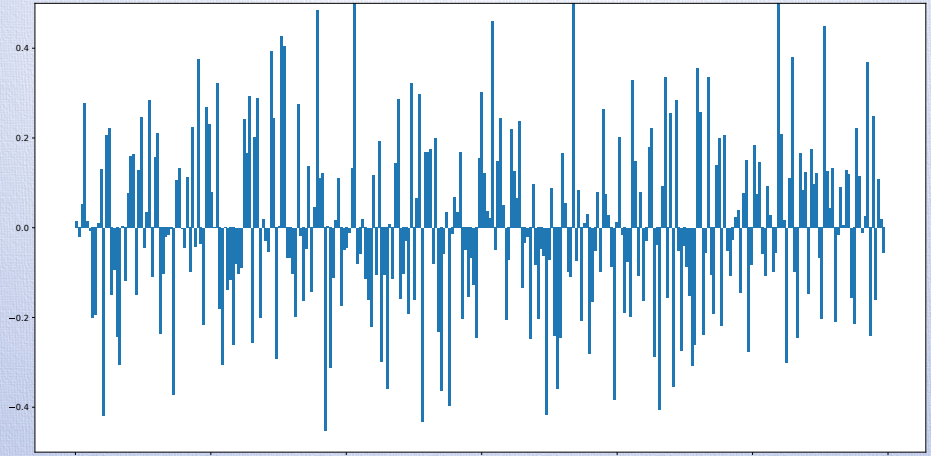
Word embedding: “King”



Word embedding: “Queen”



Word embedding: “Stockholm”



Recognizing medical terms in Swedish health records

- Named entity recognition (NER)
- Sequence tagging
  1. Disorders and findings
  2. Pharmaceutical drugs
  3. Body structure

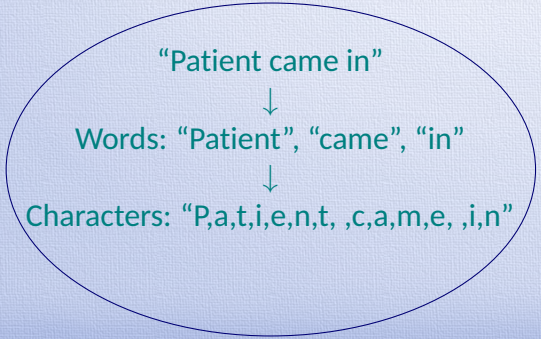
Example: *"Patient came in complaining of* ***abdominal pain****. Symptom started 2 weeks ago, sudden, usually lasts intermittently. He rates the* ***pain*** *as 8/10.* ***Pain*** *is located on the* ***periumbilical region****.* ***Pain*** *is described as* ***aching****,* ***shooting****,* ***squeezing****, and* ***throbbing****."*

Can character-level RNNs learn representations to solve medical named entity recognition?

Electronic health records

- Multi-word expressions
- One noun - several mentions
- Synonymy
- Hierarchy/Hyponymy
- Misspellings
- Redundancy
- Diverse writing style

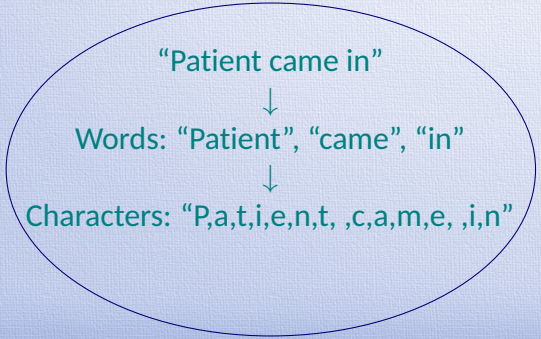
Character-based modelling



Character-based modelling

Traditional approaches:

- Sequences of word symbols



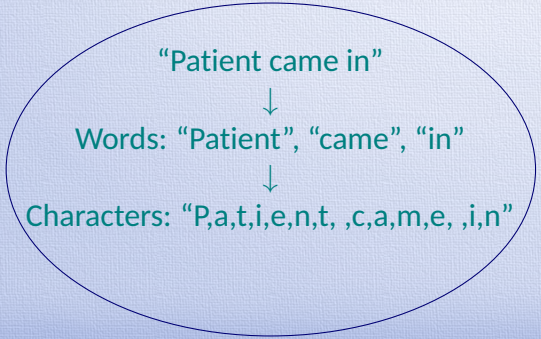
Character-based modelling

Traditional approaches:

- Sequences of word symbols

Our solution:

- Sequences of character symbols



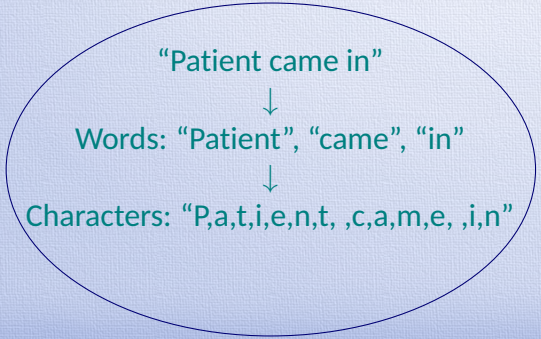
Character-based modelling

Traditional approaches:

- Sequences of word symbols
- BIO-coded tags
  (Begin, Internal, Other)

Our solution:

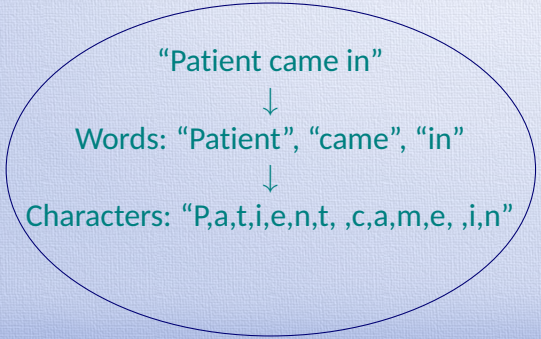
- Sequences of character symbols
- No BIO-tags needed



Character-based modelling

Traditional approaches:

- Sequences of word symbols
- BIO-coded tags
  (Begin, Internal, Other)
- Feature engineering

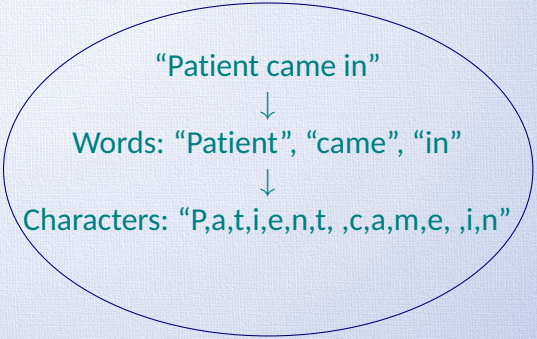


Our solution:

- Sequences of character symbols
- No BIO-tags needed

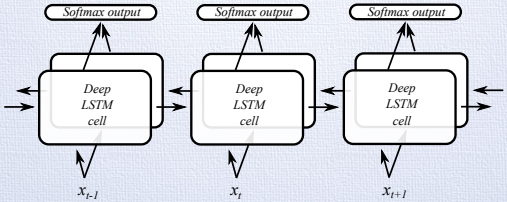
- Feature learning
- Character-level categorization

Character-based modelling



- No tokenization
- No out-of-vocabulary tokens
- Manageable alphabet
- Learns useful subword features:
  e.g.: suffixes, prefixes, capitalization, numericals
- Learns robustness and sensitivity to subword variation

The proposed model



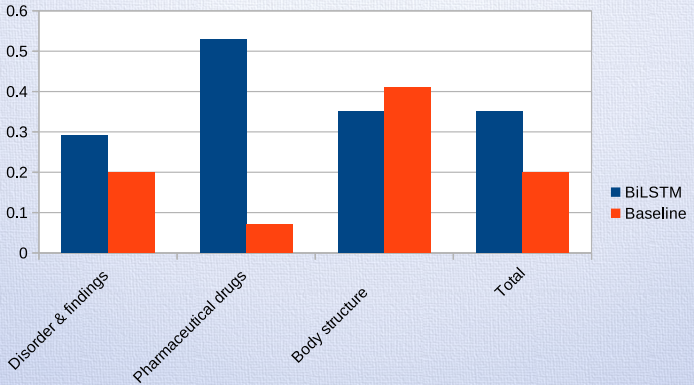
- Bidirectional deep RNN
- Character sequence as input, one output per character.
- Output postprocessing: word level majority vote

Paper III *is joint work with Simon Almgren and Sean Pavlov.*

Datasets

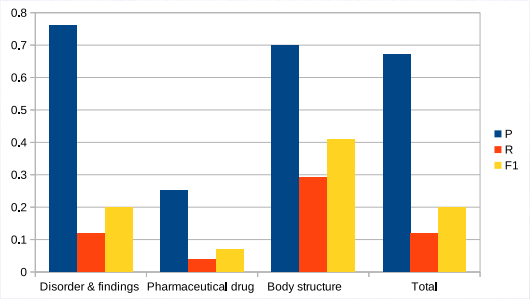
- Training, validation: data extracted from 1177 Vårdguiden and Wikipedia
  - Relatively high-quality text
  - Similar topic
- Evaluation: Stockholm EPR corpus
  - Electronic health records
  - Misspellings, redundancy, diverse writing style

Results: Recognizing medical terms



~75% relative improvement on total F-score.

Classification results



Proposed RNN performance.

(-)