NLP Edward Chang

Past, Present, Future

Past 1950 → 2013

hand-engineering models based on heuristics

polysemy 一词多义

Data Driven (supervised) 2013 – 2018

Data Driven (self-supervised learning) 2018 –

Frontier Research → multi-lingual

Frontier Research → multi-modal

NLP Tasks

Info/News Related Apps
- Info/news Recommendations
- News Summary
- Document Classification
- Question Answering
- Named Entity Recognition
- Sentiment Analysis
- Dialogue (Amazon Alexa Prize)

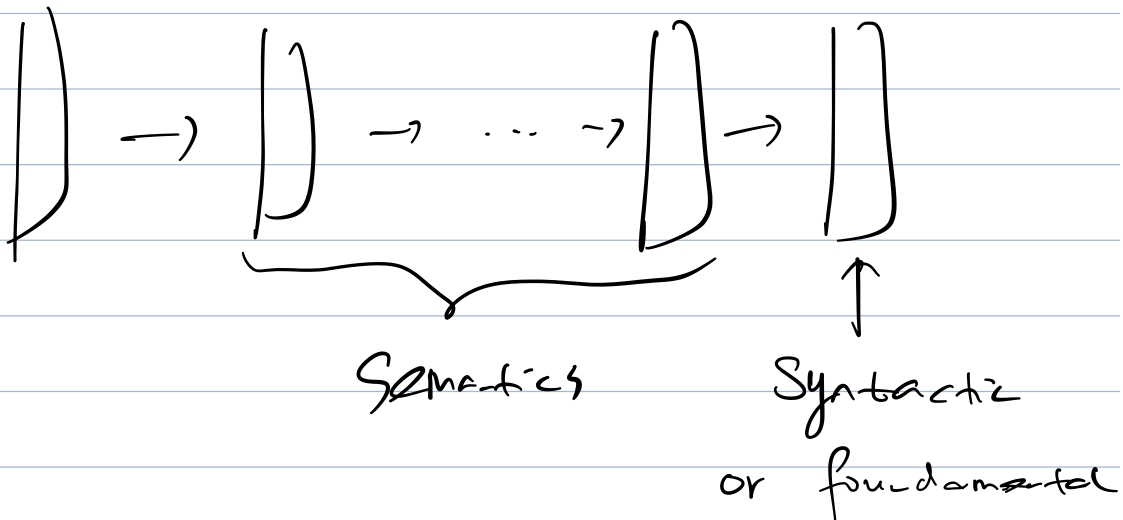
Goals/Metrics
- Optimize Accuracy/Relevance
  - Model and Data
- Minimize Training Time (and Cost)
  - For information freshness
- Reduce Serving Latency (and Cost)
- Improve Engineering Productivity
  - Zero-shot learning
  - Fine-tuning
  - Multi-task training

Language:

✓ Writing / recording makes thoughts presented over space and time, and hense knowledge accumulated.

✓ Novel ideas → invention

✓ invention → knowledge



features

Vision / Language

1958 David Hubel / Torsten Wiesel

cat's cortex $\Longrightarrow$ biophysical model

2010 / 2013 Visualize Vision

✓ Words (Linguistics)

→ discrete symbols

→ one hot representation

① too many words → vector dim high

② Similarity? Synonym, polysemy

Preprocessing :

Preprocessing
using human heuristics and linguistic principles

- Tokenization
  - What is a term, token and type
  - E.g., "a rose is a rose is a rose", 3 types 8 tokens
- Normalization
  - Car, car, CAR
- Stemming (or reweighting e.g., TF-IDF)
  - Removing articles (i.e., the, a, an)
- Annotation
  - play/Verb, play/Noun, Tokyo/Place, Trump/Person or Place
- Similarity (Distance Function)
  - Between words, sentences, paragraphs, and documents

Similarity: Word Co-occurrence Matrix

- Which objects (e.g., context units) to put into rows (or columns)?
- Which features (e.g., words) to put into columns (or rows)?
- Which values (numbers) to put into cells?

- How to interpret the matrix?

Word Document Matrix for "Bag" of Words (sparse)

Upper left corner of a matrix derived from the training portion of this IMDB data release: http://ai.stanford.edu/~amaas/data/sentiment/.

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
against	0	0	0	1	0	0	3	2	3	0
age	0	0	0	1	0	3	1	0	4	0
agent	0	0	0	0	0	0	0	0	0	0
ages	0	0	0	0	0	2	0	0	0	0
ago	0	0	0	2	0	0	0	0	3	0
agree	0	1	0	0	0	0	0	0	0	0
ahead	0	0	0	1	0	0	0	0	0	0
ain't	0	0	0	0	0	0	0	0	0	0
air	0	0	0	0	0	0	0	0	0	0
aka	0	0	0	1	0	0	0	0	0	0

Cosine distance

Merely Co-occurrence :

Shortcomings of Merely Co-occurrence

- Counts only, no semantical relationship
  - Noun: Lexus, Benz, BMW
  - Verb: test, tested
  - Adjective: pretty, beautiful, attractive
- Synonymy: many ways to refer to the same object, e.g. car and automobile
  - leads to poor recall
- Polysemy: most words have more than one distinct meaning, e.g. fall, bank, model, chip
  - leads to poor precision

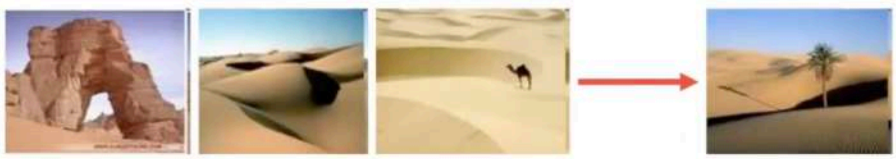
Recommendations

Market Basket Problem (Rakesh, 2000)

4/16/20

shelving or promotion so if you visited
Costco on several times you

Association-rule Based Prediction



To grow the base, we need association rules

- An association rule: $a, b, c \longrightarrow d$
- A Bayesian interpretation: $P(d \mid a, b, c) = \frac{N(a,b,c,d)}{N(a,b,c)}$
- The key is to count the occurrences (*support*) of itemsets $N(\ldots)$

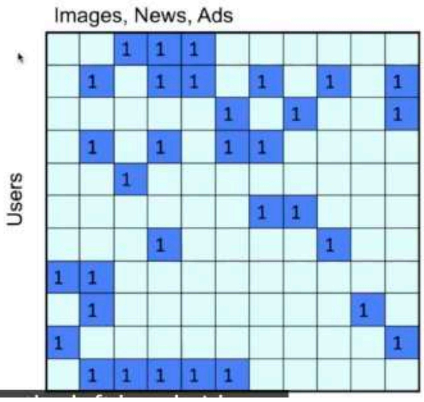
Recommender

A Collaborative Filtering Example

Based on *membership* so far, and *memberships* of others

↓

Predict further *membership*

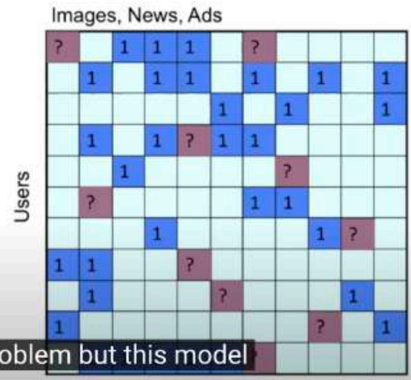


Collaborative Filtering

Based on *partially* observed matrix

↓

Predict *unobserved* entries



can solve this problem but this model suffers from

To grow the base, we need association rules

- An association rule: $a, b, c \longrightarrow d$
- A Bayesian interpretation: $P(d \mid a, b, c) = \frac{N(a,b,c,d)}{N(a,b,c)}$
- The key is to count the occurrences (*support*) of itemsets $N(\ldots)$

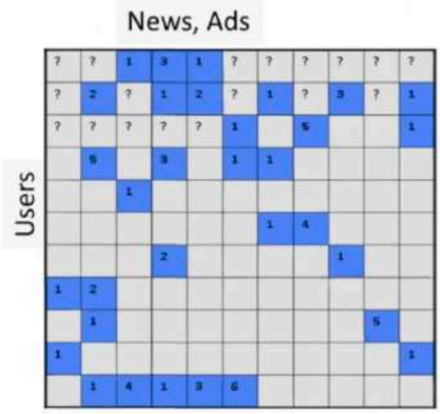
roblem Solved?

- **No.**

- Problem #1 Strength of Interest?
- Problem #2 Matrix Sparsity
- Problem #3 Semantics
- Problem #4 Cold start, no information of a new user to match interest

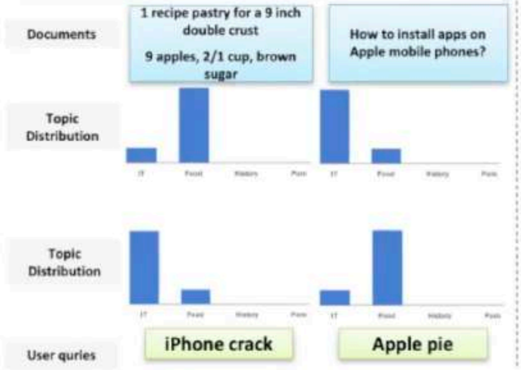
LSA

Latent Semantic Analysis



- Not all interest levels are equal
- Encoding interest strength with probability

- Search
  - Construct a latent layer for better for semantic matching
- Example:
  - iPhone crack
  - Apple pie



19 sorry about 2005 and 2013 is the

LDA: Latent Dirichlet Allocation

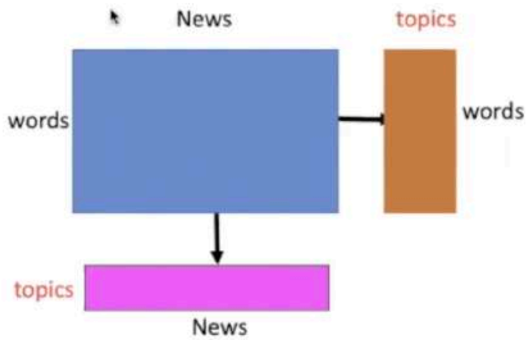
David Blei, Andrew Ng and Michael I. Jordan , 2003 (citation > 26,320)

Inputs:

1. Training data: news as bags of words
2. Parameter: the number of topics

Outputs:

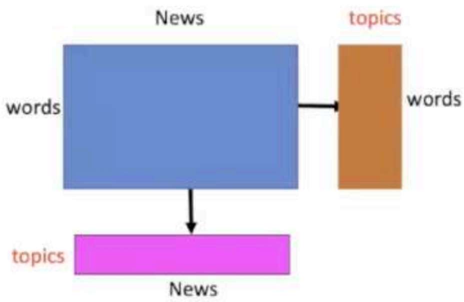
1. A co-occurrence matrix of topics and news
2. A co-occurrence matrix of topics and words



Polysemy

Polysemy

Issues



- Polysemy
  - Coarse grained, document level
  - I visited a *bank* along the *bank*.
  - Please *chip* in to buy a pizza and *chips* for group lunch.
- Synonym
  - Fall ≈ Autumn
  - Two words similar in semantics should be close in the feature space
  - How to precisely quantify similarity?

Guiding Context Hypotheses
Linguists, 1954 - 57

John Rupert Firth, (1957, 'A synopsis of linguistic theory')
"You shall know a word by the company it keeps."

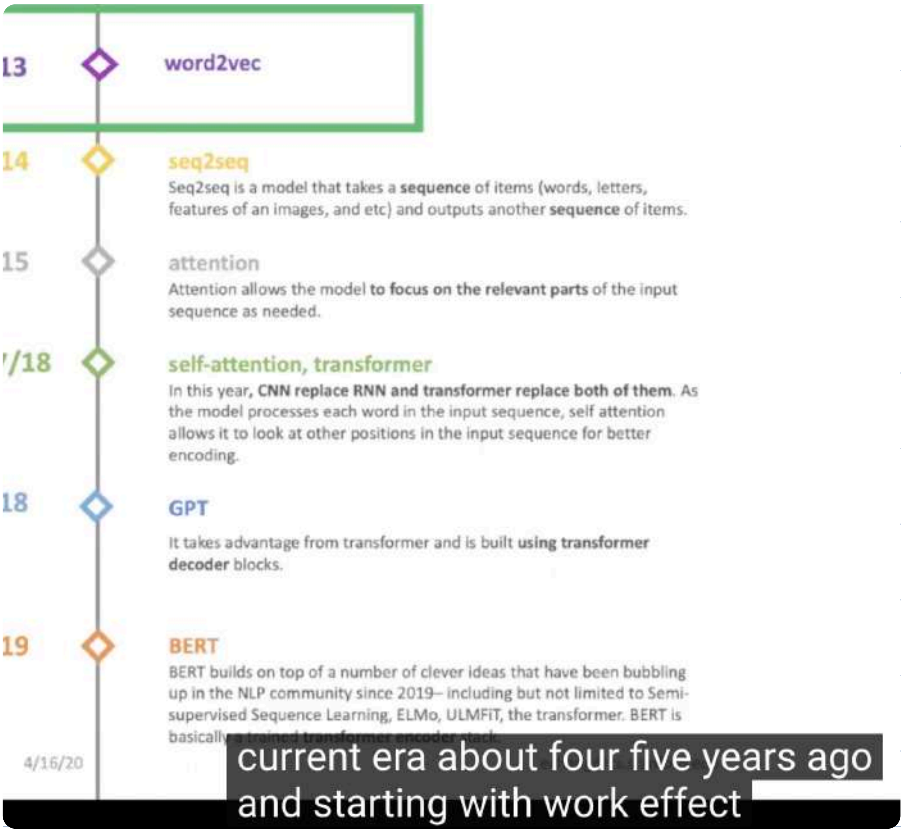
Firth (1957, 'A synopsis of linguistic theory')
"the complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously."

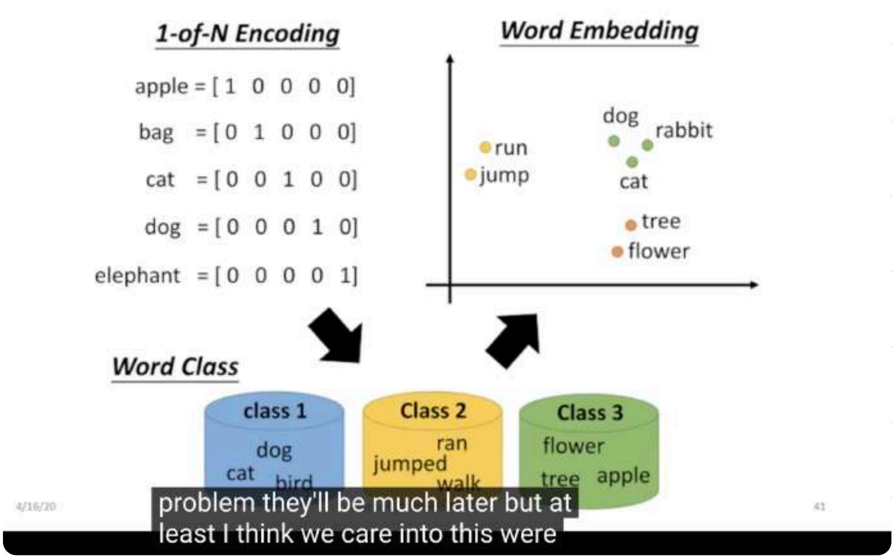
Zellig Harris (1954, 'Distributional structure')
"distributional statements can cover all of the material of a language without requiring support from other types of information."

Turney and Pantel (2010, 'From frequency to meaning')
"If units of text have similar vectors in a text frequency matrix, then they tend to have similar meanings."

seen many many years ago by linguistics
and

Word2vec


13
word2vec
14
seq2seq
Seq2seq is a model that takes a sequence of items (words, letters, features of an images, and etc) and outputs another sequence of items.
15
attention
Attention allows the model to focus on the relevant parts of the input sequence as needed.
/18
self-attention, transformer
In this year, CNN replace RNN and transformer replace both of them. As the model processes each word in the input sequence, self attention allows it to look at other positions in the input sequence for better encoding.
18
GPT
It takes advantage from transformer and is built using transformer decoder blocks.
19
BERT
BERT builds on top of a number of clever ideas that have been bubbling up in the NLP community since 2019– including but not limited to Semi-supervised Sequence Learning, ELMo, ULMFiT, the transformer. BERT is basically a trained transformer encoder stack.
4/16/20
current era about four five years ago
and starting with work effect

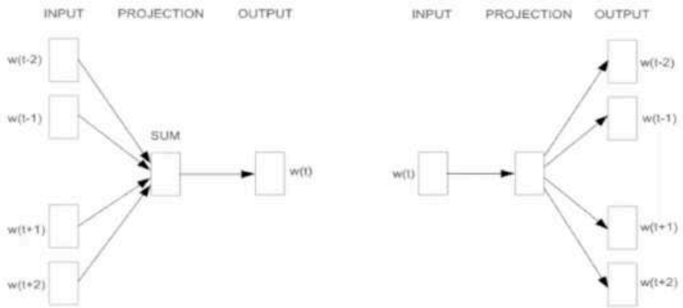

1-of-N Encoding
apple = [1 0 0 0 0]
bag = [0 1 0 0 0]
cat = [0 0 1 0 0]
dog = [0 0 0 1 0]
elephant = [0 0 0 0 1]
Word Embedding
run
jump
dog
rabbit
cat
tree
flower
Word Class
class 1
dog
cat
bird
Class 2
ran
jumped
walk
Class 3
flower
tree
apple
4/16/20
41
problem they'll be much later but at
least I think we care into this were

:ount-based embedding → Predictive embedding Word2Vec

- Word2Vec is one of the most popular technique to learn word embeddings using shallow neural network. It was developed by Tomas Mikolov in 2013 at Google.
  - [ICLR13W] Efficient estimation of word representations in vector space (14,940 citations), 4/15/2020
  - [NIPS13] Distributed representations of words and phrases and their compositionality (18,759 citations), 4/15/2020

Represent the meaning of word – word2vec

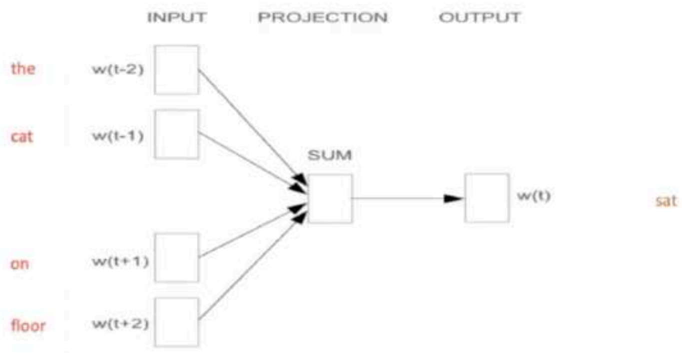
- Continuous Bag of Word (CBOW): use a window of word to predict the middle word
- Skip-gram (SG): use a word to predict the surrounding ones in window

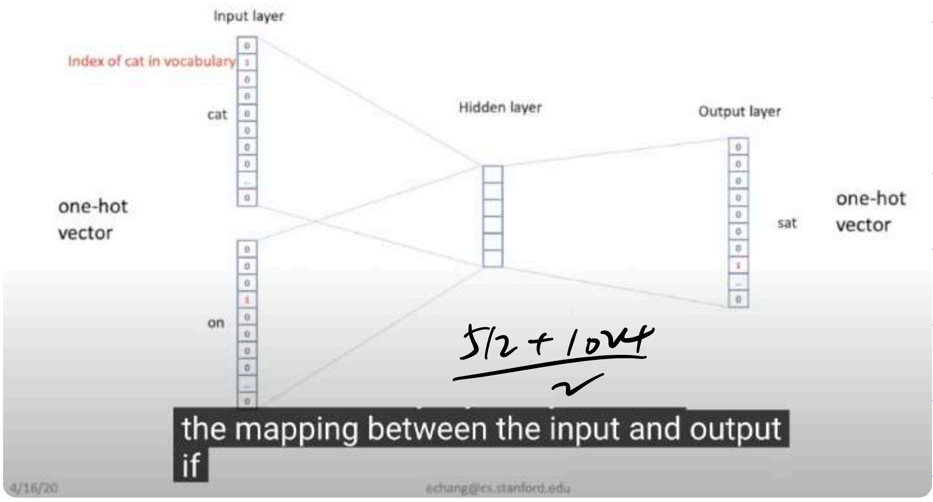


called self supervising learning we have five words

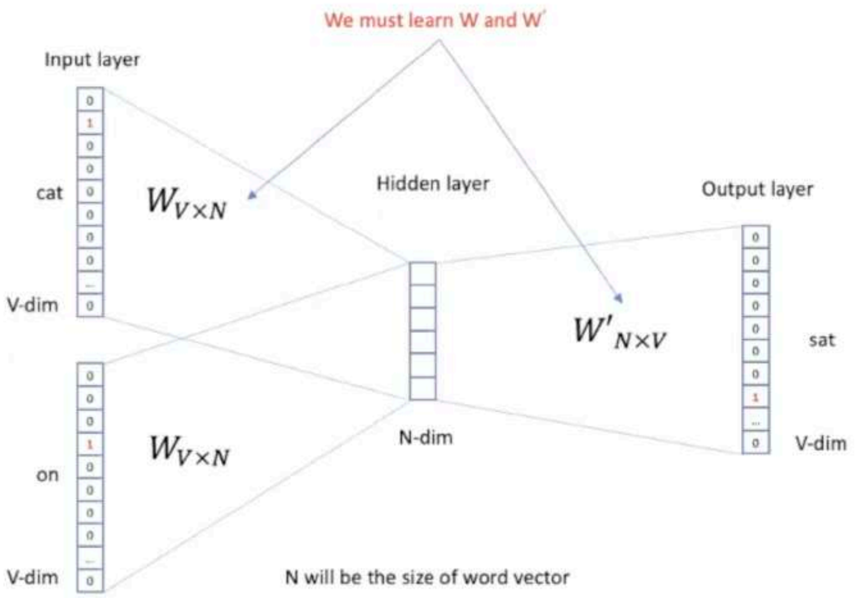
Word2vec – Continuous Bag of Word (CBOW)

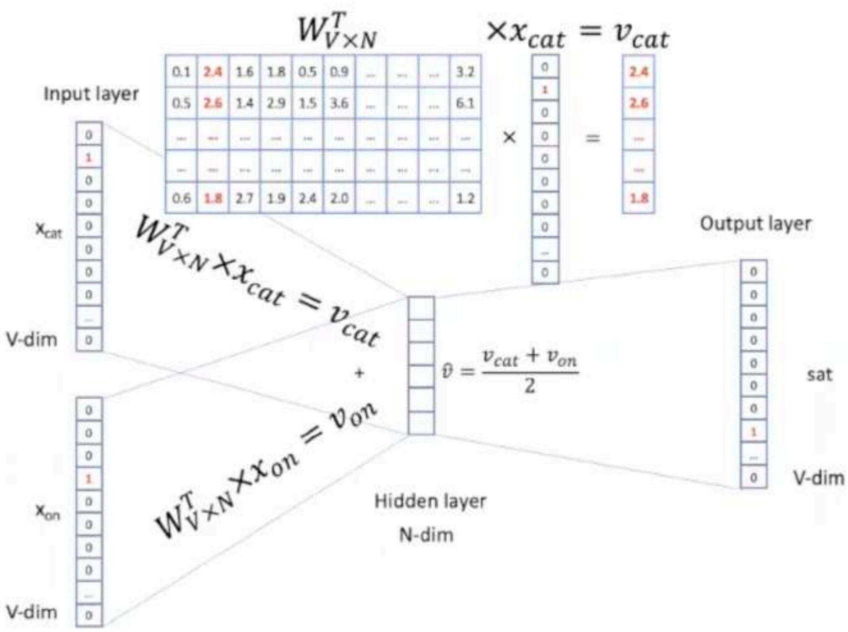
- E.g. "The cat sat on floor"
  - Window size = 2





de-facto number





CBOW vs. Skip-gram (Mikolov)

- **CBOW:** several times faster to train than the skip-gram, slightly better accuracy for the frequent words.

- **Skip-gram:** works well with small amount of the training data, represents well even rare words or phrases.

Some interesting results

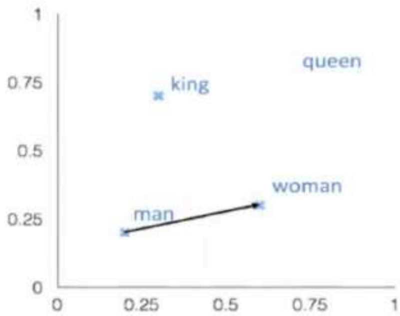
Word Analogies

Test for linear relationships, examined by Mikolov et al. (2014)

a:b :: c:? $\longrightarrow$ $d = \arg\max_{x} \frac{(w_b - w_a + w_c)^T w_x}{||w_b - w_a + w_c||}$

man:woman :: king:?

+	king	[0.30 0.70]
-	man	[0.20 0.20]
+	woman	[0.60 0.30]
	queen	[0.70 0.80]



Subword Consideration

1. Schütze (1993) pioneered subword modeling to improve representations by reducing sparsity, thereby increasing the density of connections in a VSM.
2. Subword modeling will also
   a. Pull morphological variants closer together
   b. Facilitate modeling out-of-vocabulary items
   c. Reduce the importance of any particular tokenization scheme

Subword :

Subword

Bojanowski et al. (2016) (the fastText team) motivate a straightforward approach:

1. Given a word-level VSM, the vector for a character-level *n*-gram *x* is the sum of all the vectors of words containing *x*.
2. Represent each word *w* as the sum of its character-level *n*-grams.
3. Add in the representation of *w* if available

A linguistically richer variant might use sequences of morphemes rather than characters.

Example with 4-grams

superbly becomes

[<w>sup, supe, uper, perb, erbl, rbly, bly</w>]

Some Tricks to deal with OOV

- I am working at DeepQ
- DeepQ [OOV]
- Working? Work too too many combinations (memory size)
- So, subword-based + character-based (only 26)

- Working = Work + ##ing
- Worked = Work + ##ed

Synonymy & Polysemy

Synonymy & Polysemy
after word2vec

$W_{V \times N}^{T}$

0.1	2.4	1.6	1.8	0.5	0.9	...	...	...	3.2
0.5	2.6	1.4	2.9	1.5	3.6	...	...	...	6.1
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
0.6	1.8	2.7	1.9	2.4	2.0	...	...	...	1.2

- Synonymy: many ways to refer to the same object, e.g. car and automobile
  - leads to poor recall

- Polysemy: most words have more than one distinct meaning, e.g. fall, bank, model, chip
  - leads to poor precision

Word2vec → Polysemy (Arora 2016)

- **Linear Algebraic Structure of Word Senses, with Applications to Polysemy**
  Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, Andrej Risteski, latest version 7 Dec 2018
- **Abstract**: Word embeddings are ubiquitous in NLP and information retrieval, but it's unclear what they represent when the word is polysemous, i.e., has multiple senses. Here it is shown that multiple word senses reside in linear superposition within the word embedding and can be recovered by simple sparse coding. The success of the method ---which applies to several embedding methods including word2vec--- is mathematically explained using the random walk on discourses model (Arora et al., 2016). A novel aspect of our technique is that each word sense is also accompanied by one of about 2000 discourse atoms that give a succinct description of which other words co-occur with that word sense. Discourse atoms seem of independent interest and make the method potentially more useful than the traditional clustering-based approaches to polysemy.

Word vectors that are polysemous

0.02	**0.03**	**0.9**	**0.02**	**0**	**0**	**0**	**0.02**	**0.01**	**0**
0.03	0.43	0.02	0	0	0	0.37	0.1	0.05	0
0.3	0.01	0.02	0.33	0.01	0	0	0.27	0.03	0.03
0.03	0.05	0.05	0.03	0.4	0.3	0.02	0.05	0.03	0.04

2017

Logical Remedy: Consider More Context

- Vector = f("word") Word2Vec
- Vector = f("word", "context")

- Seq2Seq (2014)
- RNN, LSTM (2015)
- Attention (2017,18)

Design Steps

Design Steps (Traditional vs. NN)

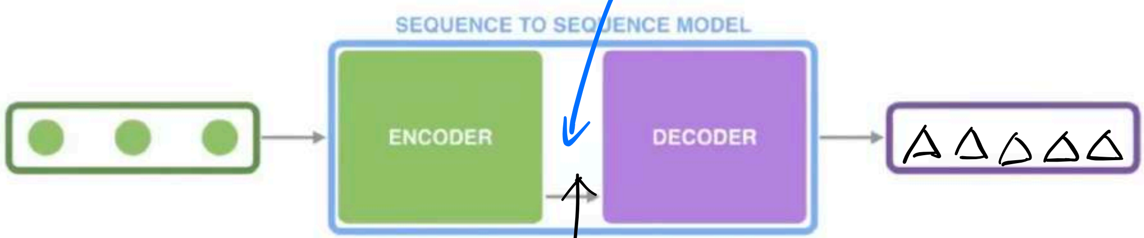
Traditional Models	**Neural Network Models**
Pre-processing	~~Pre-processing~~
Similarity (Co-occurrence matrix)	~~Similarity (Co-occurrence matrix)~~
Dimension reduction	~~Dimension reduction~~
~~Pretraining~~	Data-driven representation learning

Seq2seq model

- [EMNLP14] Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation (7,000+ citations)
- [NIPS14] Sequence to Sequence Learning with Neural Networks (8,000+ citations)
  - Proposed Seq2seq and GRU, a simplified LSTM algorithm
  - LSTM and GRU replaces RNN

Seq2seq model (white box) w/ context

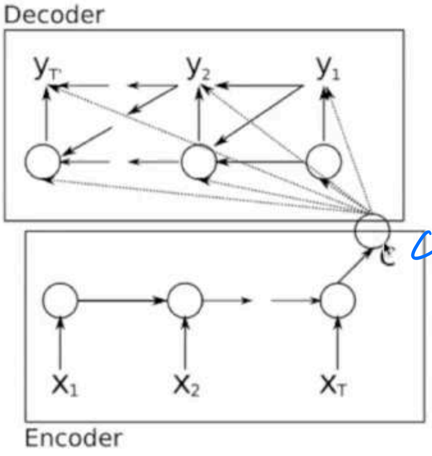
Context



final hidden state
to decoder

[EMNLP14] Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

(6,848 citations)



$$P(y_t|y_{t-1}, y_{t-2}, \ldots, y_1, \mathbf{c}) = g\left(\mathbf{h}_{\langle t \rangle}, y_{t-1}, \mathbf{c}\right)$$

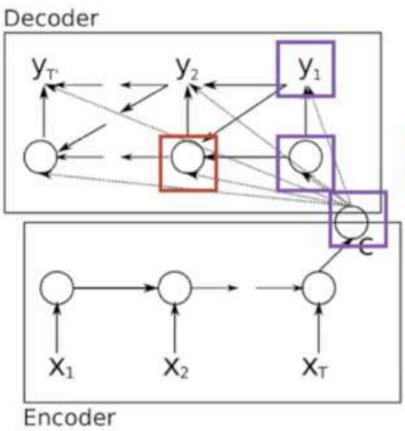
$$\mathbf{h}_{\langle t \rangle} = f\left(\mathbf{h}_{\langle t-1 \rangle}, y_{t-1}, \mathbf{c}\right)$$

context

input we produce this final token which
is contacts and we

[EMNLP14] Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

(6,848 citations)



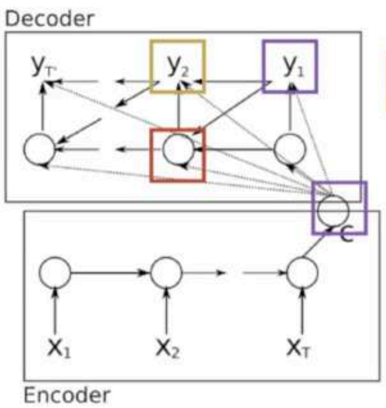
$$P(y_t|y_{t-1}, y_{t-2}, \ldots, y_1, \mathbf{c}) = g\left(\mathbf{h}_{\langle t \rangle}, y_{t-1}, \mathbf{c}\right)$$

$$\boxed{\mathbf{h}_{\langle t \rangle}} = f\left(\boxed{\mathbf{h}_{\langle t-1 \rangle}}, \boxed{y_{t-1}}, \boxed{\mathbf{c}}\right)$$

only papers on see there's no c2 and c1
just based on the final C and

[EMNLP14] Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

(6,848 citations)



$$\boxed{P(y_t|y_{t-1}, y_{t-2}, \ldots, y_1, \mathbf{c})} = g\left(\boxed{\mathbf{h}_{\langle t \rangle}}, \boxed{y_{t-1}}, \boxed{\mathbf{c}}\right)$$

$$\mathbf{h}_{\langle t \rangle} = f\left(\mathbf{h}_{\langle t-1 \rangle}, y_{t-1}, \mathbf{c}\right)$$

distance then we can no longer
understand or getting the information

LSTM Shortcomings

RNN + LSTM Shortcomings

- Long distance dependencies
  - Only the final hidden state of the sequence is passed to decoder, cannot address individual words
  - Vanishing gradients
- LSTM and RNN *cannot be parallelly trained* within training examples
  - A hidden state h_t depends on h_{t-1} and the input position
  - Sequential computation in nature, not parallelizable

In 2018, Two "Revolutions"

- Seq2Seq (2014)
- RNN, LSTM

2015

- Transformer using Attention (2017/18)
- Bidirectional Masked Language Model (2018)

he fall of RNN / LSTM

Eugenio Culurciello Follow
Apr 13, 2018 · 8 min read

"Drop your RNN and LSTM, they are no good!"

But do not take our words for it, also see evidence that Attention based networks are used more and more by Google, Facebook, Salesforce, to name a few. All these companies have replaced RNN and variants for attention based models, and it is just the beginning. RNN have the days counted in all applications, because they require more resources to train and run than attention-based models. See this post for more info.

4/16/20

this is happening in the last two years
or so this is an interesting block by
researcher

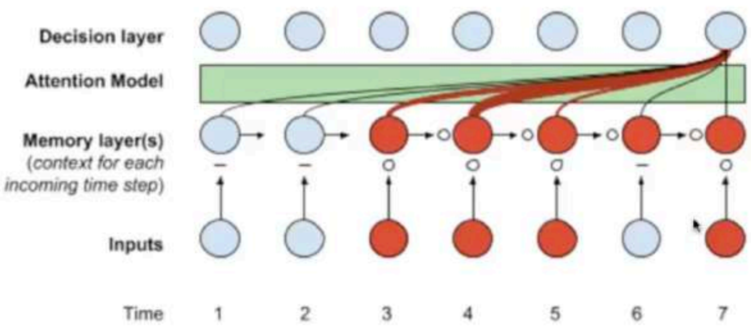
80

04/2018

Components

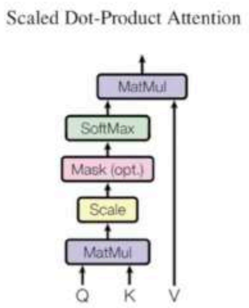
- Encoder
- Decoder
- Position Encoding
- Attention
  - V: value
  - K: key or index
  - Q: query, quering V via K
  - E.g., V: [150, 175, 45]; K: [weight (lb), height (cm), age]; Q: weight → 150 lbs

Attention Mechanism

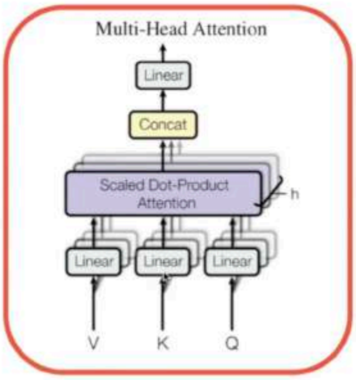


Attention Inputs V, K, Q

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



V, K, and Q

- V: [150, 175, 45]; K: [weight (lb), height (cm), age]
- Q: weight → 150 lbs

- V: [I enjoy Tokyo foliage]; K: [subject, verb, noun, noun]
- Q: subject → I

Complexity and Information Path Length

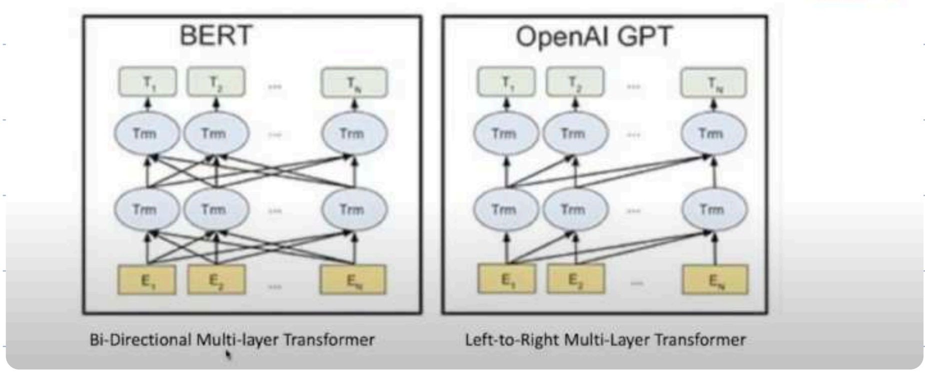
Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention.

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

BERT & GPT

(Self-supervised Pre-trained Model)

BERT vs. GPT



Self-Supervised (Semi-Supervised) Learning

- Computer Vision: compression and decompression
  - Error function: SNR
- NLP: predict missing words
  - Next word?
  - Masked words?

Why Uni- and Bi-directional?

Self-supervised learning needs a language model

- By the definition of LM, one looks from left to right, and hence GPT
- GPT can generate language from left to right (but BERT cannot)

BERT uses masked language model

- Predict masked words (15%)
- Predict if a sentence is *isNext* sentence

LM Example

- The man went to a supermarket. He bought a gallon of milk.

GPT

- The man
- The man went
- The man went to…

next word

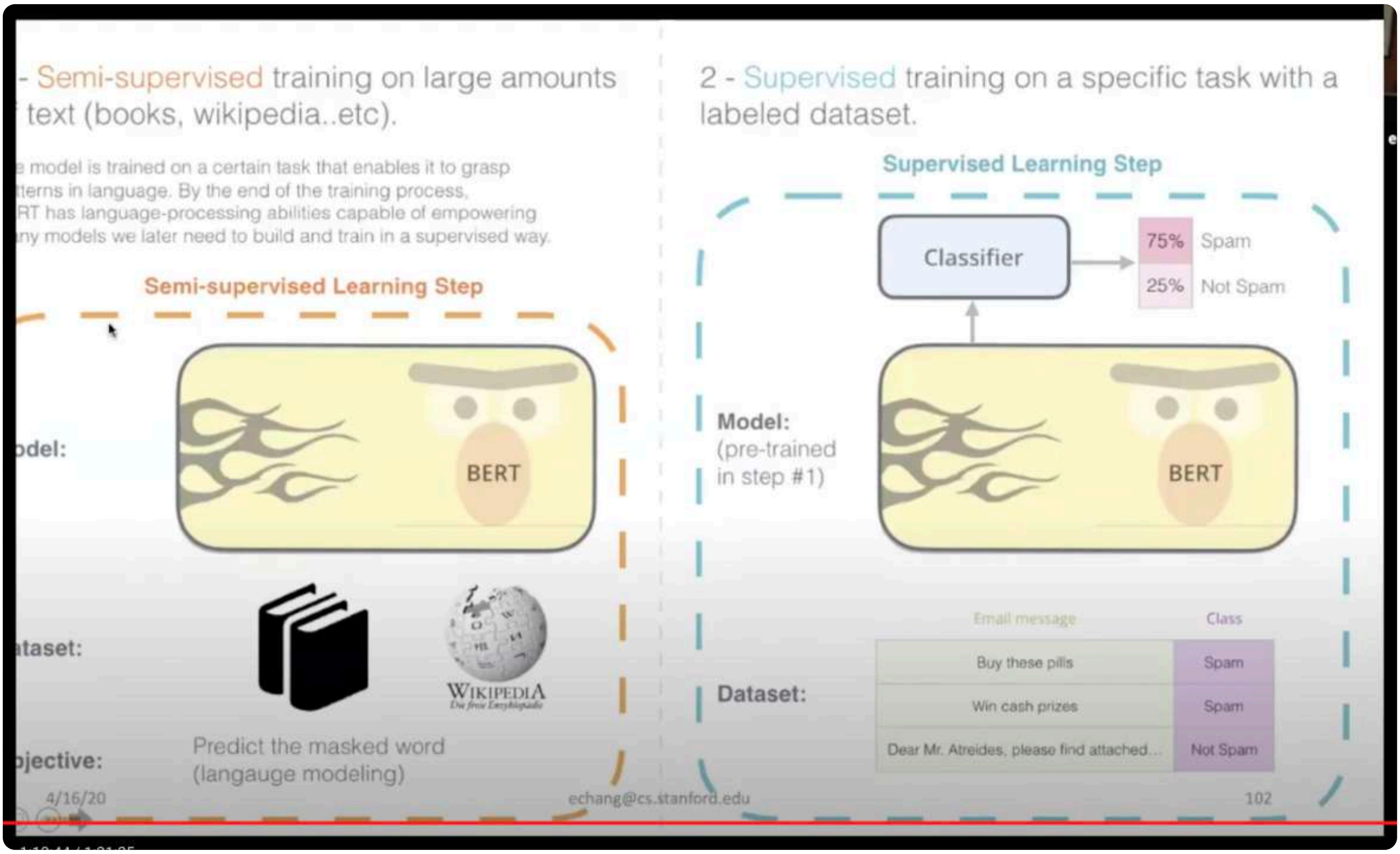
Masked LM Example

- The man went to a supermarket. He bought a gallon of milk.

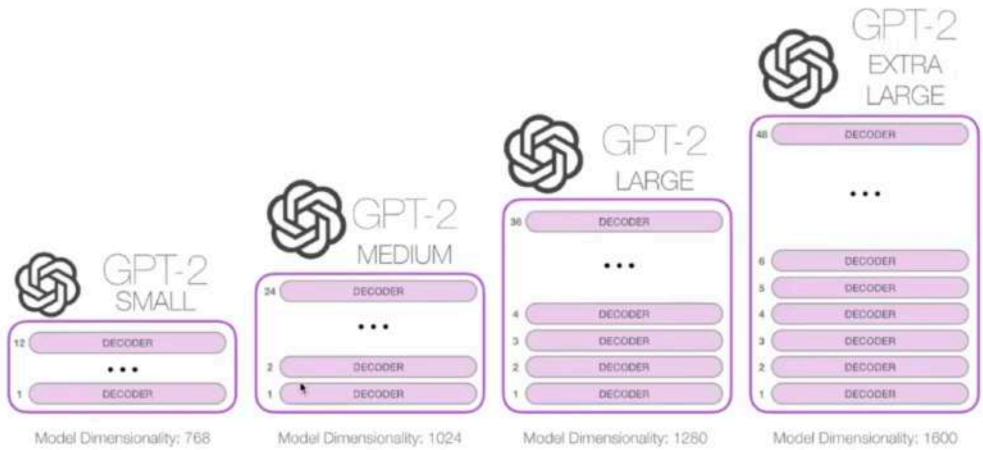
BERT

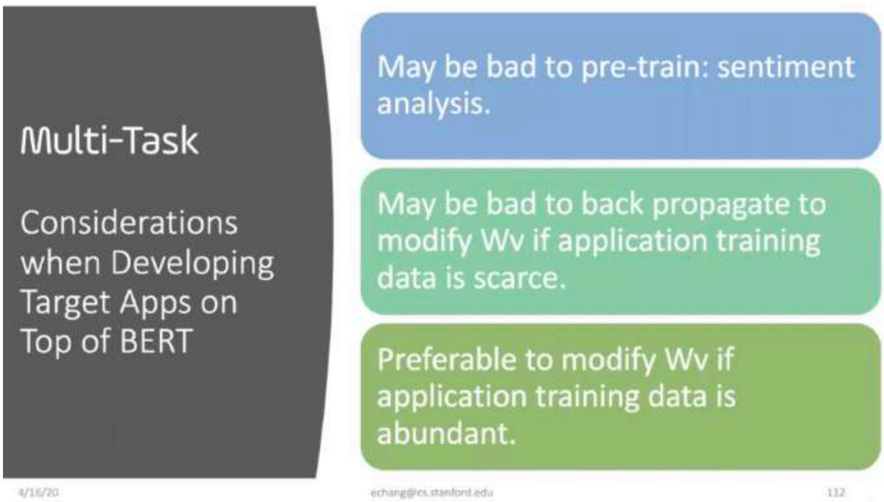
- Input: [cls] The man went to a ~~supermarket~~ [sep] He ~~bought~~ a gallon of milk. [sep]
- Label: isNext True

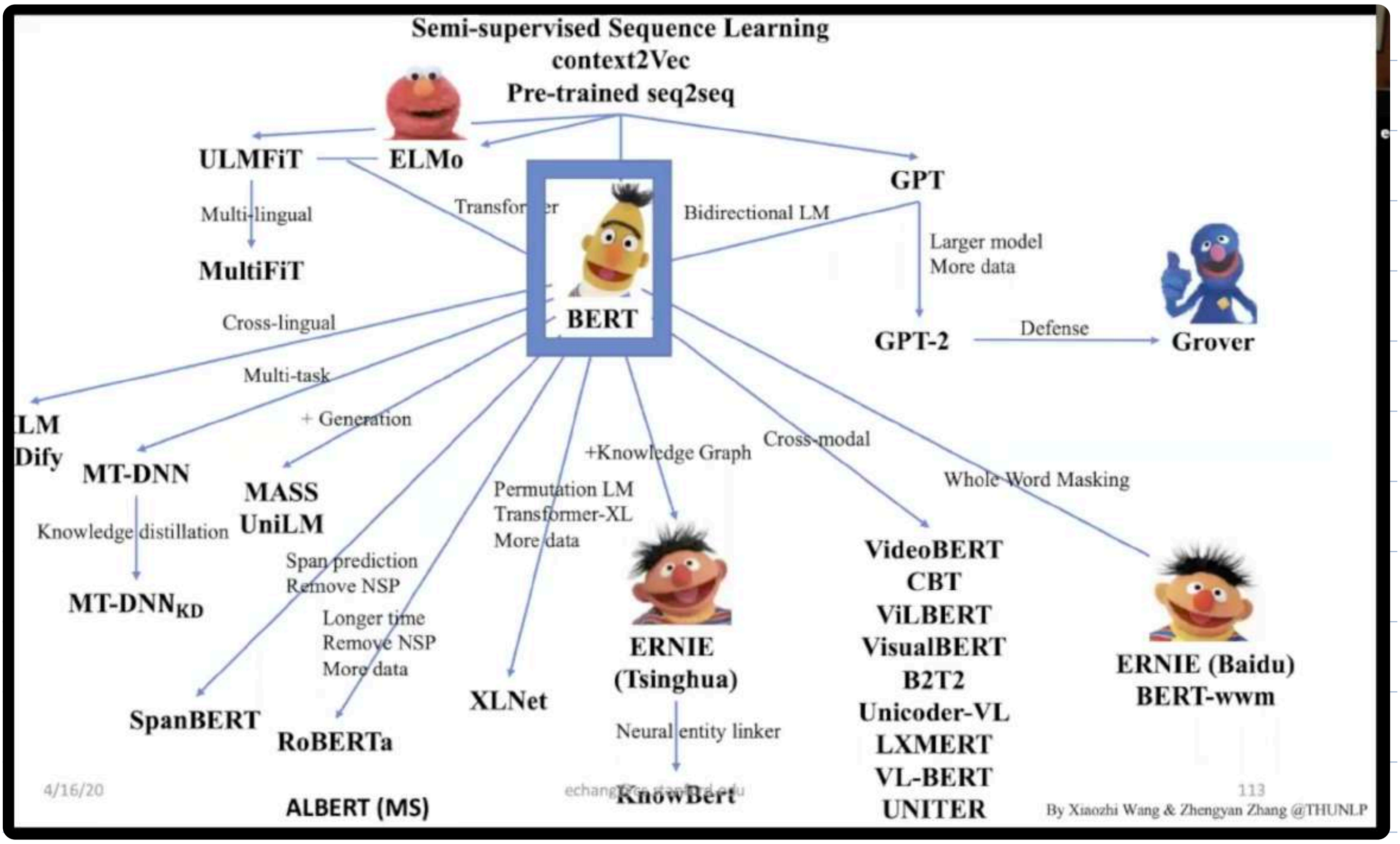
- Input: [cls] The man went to a ~~supermarket~~ [sep]. Mary walks his ~~dogs~~ every day.[sep]
- Label: isNext False


- Semi-supervised training on large amounts
text (books, wikipedia..etc).
e model is trained on a certain task that enables it to grasp
tterns in language. By the end of the training process,
RT has language-processing abilities capable of empowering
ny models we later need to build and train in a supervised way.
Semi-supervised Learning Step
odel:
BERT
ataset:
WIKIPEDIA
bjective:
Predict the masked word
(langauge modeling)
2 - Supervised training on a specific task with a
labeled dataset.
Supervised Learning Step
Classifier
75% Spam
25% Not Spam
Model:
(pre-trained
in step #1)
BERT
Dataset:
Email message
Class
Buy these pills
Spam
Win cash prizes
Spam
Dear Mr. Atreides, please find attached...
Not Spam
4/16/20
echang@cs.stanford.edu
102

fine tuning


GPT-2
SMALL
DECODER
Model Dimensionality: 768
GPT-2
MEDIUM
DECODER
Model Dimensionality: 1024
GPT-2
LARGE
DECODER
Model Dimensionality: 1280
GPT-2
EXTRA
LARGE
DECODER
Model Dimensionality: 1600


Multi-Task
Considerations when Developing Target Apps on Top of BERT
May be bad to pre-train: sentiment analysis.
May be bad to back propagate to modify Wv if application training data is scarce.
Preferable to modify Wv if application training data is abundant.
4/16/20
echang@cs.stanford.edu
112



What's Next in NLP Research?

- From BERT to many BERTs
- Multilingual BERTs (Zero-shot Transfer Learning)
- Multimodal BERTs (Language + Vision)
- Unified-Task Layer: https://decanlp.com/

Date	Paper	Architecture	Visual token	Pre-train datasets	Pre-train tasks	Downstream tasks
/9	VisualBERT	Single cross-modal transformer	Image RoI	COCO caption	• Masked language modeling • Sentence-image matching	• visual question answering (VQA) • visual commonsense reasoning (VCR) • natural language visual reasoning (NLVR^2) • region-to-phrase grounding (Flickr30K)
/6 NIPS19	ViLBERT	One single-modal transformer (language) + One cross-modal transformer (Two stream)	Image RoI	Conceptual Captions	(Simultaneously) • Masked language modeling • Masked region modeling • Sentence-image matching	• visual question answering (VQA) • visual commonsense reasoning (VCR) • referring expression comprehension • caption-based image retrieval
/3 ICCV19	VideoBERT	Single cross-modal transformer	Video frame	Cooking312K	• Masked language modeling • Masked region modeling • Sentence-image matching	• zero-shot action classification • video captioning
/20 EMNLP19	LXMERT	Two single-modal transformer + One cross-modal transformer	Image RoI	COCO caption Visual Genome VQA v2.0 GQA balanced version VG-QA	• Masked language modeling • Masked region modeling • Masked object feature regression • Sentence-image matching • Image QA	• visual question answering (VQA) • GQA • natural language visual reasoning (NLVR^2)
/22	VL-BERT	Single cross-modal transformer	Image RoI	Conceptual Captions BooksCorpus English Wikipedia	• Masked language modeling • Masked region modeling	• visual commonsense reasoning (VCR) • visual question answering (VQA) • referring expression comprehension

/16	Unicoder-VL	Single cross-modal transformer	Image RoI	Conceptual Captions	• Masked language modeling
					• Masked region modeling
					• Sentence-image matching
						• image-text retrieval
						• will do soon
						  • Image captioning
						  • VQA
						  • Visual commonsense reasoning (VCR)

/14 EMNLP19	B2T2	Single cross-modal transformer	Image RoI	Conceptual Captions	• Masked language modeling • Sentence-image matching	• visual commonsense reasoning (VCR)
/13	CBT	Two single-modal transformer + One cross-modal transformer	Video frame	HowTo100M		• action anticipation • video captioning • action segmentation
/25	UNITER	Single cross-modal transformer	Image RoI	COCO Visual Genome Conceptual captions SBU Captions	• Masked language modeling • Masked region modeling • Sentence-image matching	• visual question answering (VQA) • visual commonsense reasoning (VCR) • natural language visual reasoning (NLVR^2) • visual entailment • image-text retrieval • referring expression comprehension

Does BERT really understand language?

- Is BERT a hack …
  - Linzen @ JHU is not convinced with the BERT hack: “we are trying to understand to what extent these models are really understanding language.”
  - Counter evidences provided by Linzen’s HANS and NCKU paper