Multimodal CL and NLP: Combining Language and Vision for (Computational) Semantics

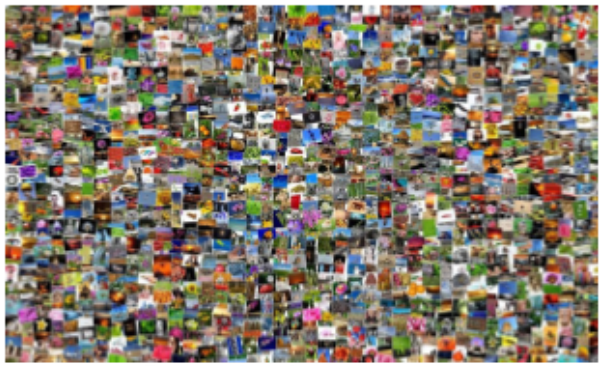
Carina Silberer

IMS, University of Stuttgart

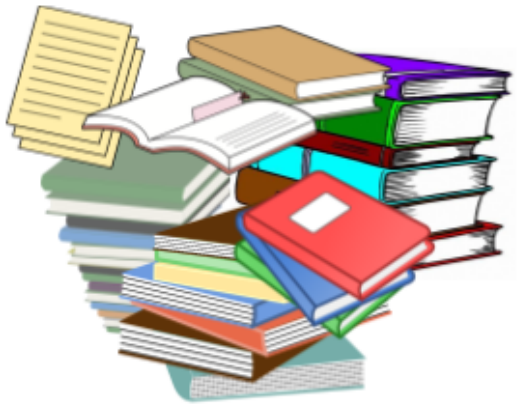
CL Fall School 2024, Passau
16 - 20 September 2024

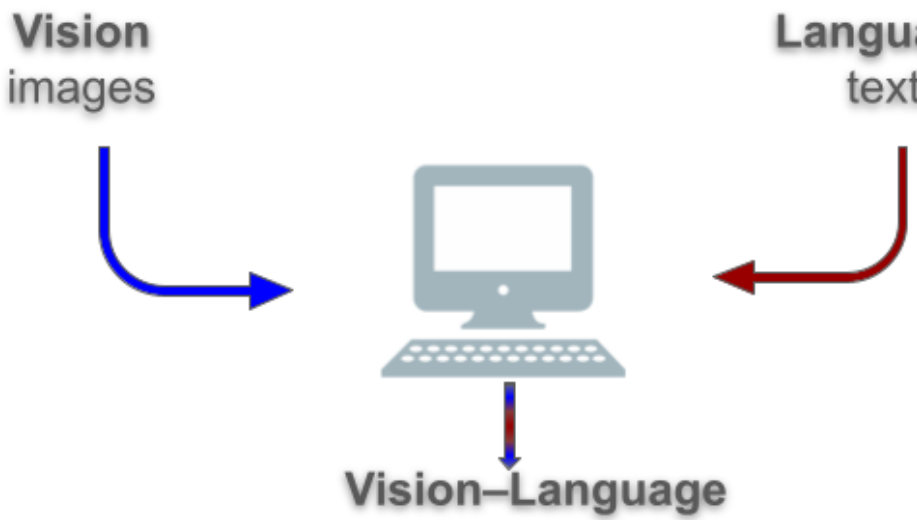
Outline

Multimodal (VL) Representations



https://pixabay.com/photos/photo-collection-pictures-photos-382018





Grounded Semantic Models: Motivation

Conceptual Grounding:
Language is connected to perceptual and sensorimotor interaction with environment [Barsalou, 2008]

Cognitive Perspective

Computational/Practical Perspective: Grounding to Vision

Grounded Semantic Models: Motivation

Conceptual Grounding:
Language is connected to perceptual and sensorimotor interaction with environment [Barsalou, 2008]

Cognitive Perspective

- Purely language-based models that are isolated from situational context are overly artificial
- Conceptual and sensorimotor representations are connected and interacting
- Early Models:[Andrews et al., 2009, Johns and Jones, 2012, ...]

Grounded Semantic Models: Motivation

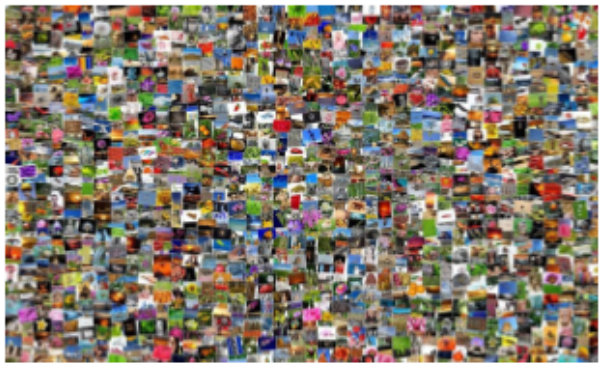
Conceptual Grounding:
Language is connected to perceptual and sensorimotor interaction with environment [Barsalou, 2008]

Cognitive Perspective

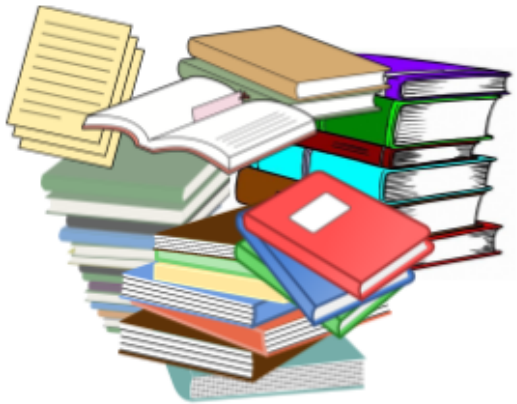
Computational/Practical Perspective: Grounding to Vision

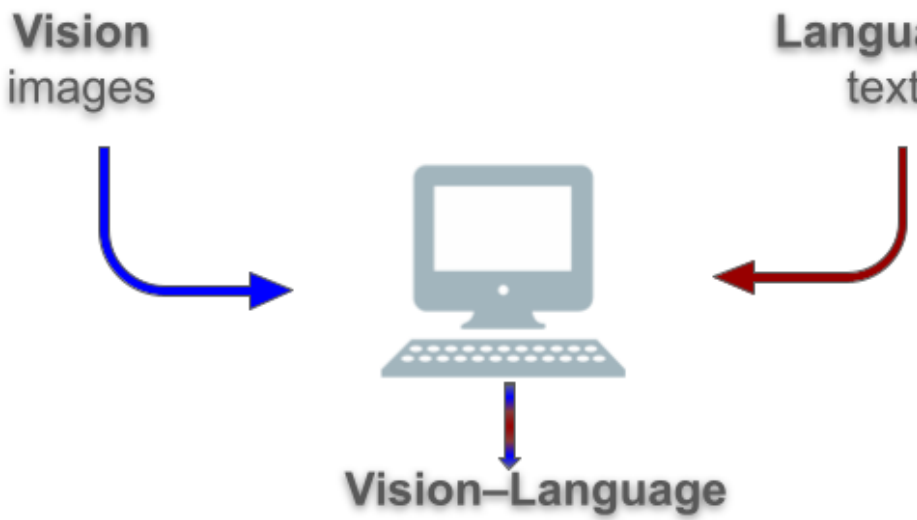
- Perceptually grounded models contain richer, more complete semantic knowledge
- Reduces the collapse of multiple senses in one embedding
- Reduces referential ambiguity
- Early Models: [Bruni et al., 2012, Silberer and Lapata, 2012, Kiela and Bottou, 2014, Lazaridou et al., 2015, Silberer et al., 2017, ...]

Multimodal (VL) Representations



https://pixabay.com/photos/photo-collection-pictures-photos-382018

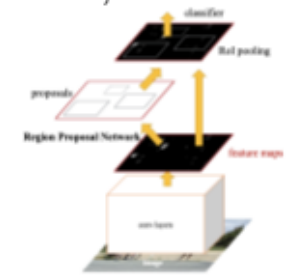




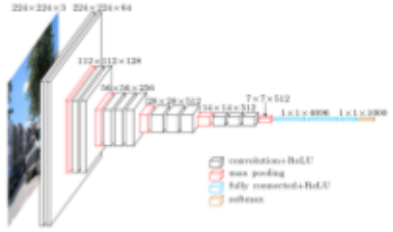
Multimodal (VL) Representations

CNNs

VGG, ResNet



Faster R-CNN

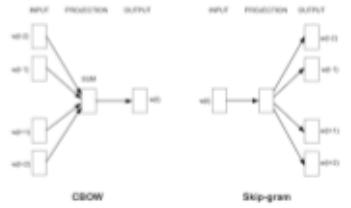


global models

co-occurrence matrix

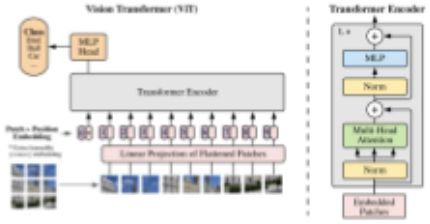
	drink	brewed	beans	exam	teacher
coffee	6891	1023	3028	34	0
tea	7732	643	138	20	0
school	2	0	0	5821	983
study	7	0	0	7112	892

word2vec



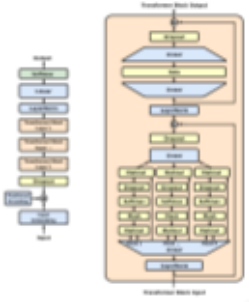
Transformers

ViT

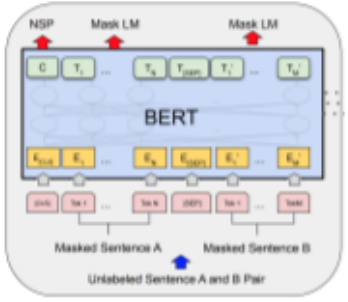


Vision
images

GPT



BERT



Language
text

Visual encoder

Textual encoder / LLM

Outline

Linguistic Representations

How to define the meaning of words?

Example: WordNet

Question: What is the meaning of a _____?

The noun _____ has 4 senses

1. underground pod of the _____vine
2. widely cultivated American plant cultivated in tropical and warm regions; showy yellow flowers on stalks that bend over to the soil so that seed pods ripen underground
3. _____a young child who is small for his age
4. pod of the _____vine containing usually 2 nuts or seeds; ‘groundnut’ and ‘monkey nut’ are British terms

Example: WordNet

Question: What is the meaning of a _____?

Hypernyms:

- Sense 1

```
=> pod, seedpod
    => fruit
        => reproductive structure
            => plant organ
                => plant part, plant structure
                    => natural object
                        => whole, unit
                            => object, physical object
                                => physical entity
                                    => entity
```

- Sense 2

```
=> legume, leguminous plant
    => herb, herbaceous plant
        => vascular plant, tracheophyte
            => plant, flora, plant life
                => organism, being
                    => living thing, animate thing
                        => whole, unit
                            => object, physical object
                                => physical entity
                                    => entity
    => climber
        => vine
            => vascular plant, tracheophyte
                => plant, flora, plant life
                => organism, being
```

Example: WordNet

Question: What is the meaning of a _____?

Hypernyms:

- Senses 3+4

```
=> child, kid, youngster, minor, shaver, nipper, small fry, tid
    => juvenile, juvenile person
        => person, individual, someone, somebody, mortal, soul
            => organism, being
                => living thing, animate thing
                    => whole, unit
                        => object, physical object
                            => physical entity
                                => entity
            => causal agent, cause, causal agency
                => physical entity
                    => entity
```

```
 earthnut, goober, goober pea, groundnut, monkey nut
=> edible nut
    => nut
        => seed
            => fruit
                => reproductive structure
                    => plant organ
                        => plant part, plant structure
                            => natural object
                                => whole, unit
                                    => object, physical object
                                        => physical entity
                                            => entity
```

Example: Images

Question: What is the meaning of a _____?

sources: `https://peanut-institute.com`, `wikipedia.org`

Early Textual Representations

Semantic models / Vector space models

Question: What does _____ mean?

Context sentences

- which is how _____ achieves fermentation
- _____ starter culture contains Bacillus subtilus
- Bacillus subtilis forms spores that are found in the _____
- _____ is frequently eaten in Japan over rice
- You can also eat _____ with condiments such as soy sauce
- Soybeans for _____ vary in size from super large to tiny ones

Question: What does _____ mean?

Context sentences

- which is how _____ achieves **fermentation**
- _____ **starter** culture contains Bacillus subtilus
- Bacillus subtilis forms spores that are found in the _____
- _____ is frequently **eaten** in Japan over **rice**
- You can also eat _____ with condiments such as **soy sauce**
- **Soybeans** for _____ vary in size from super large to tiny ones

Question: What does _____ mean?

Context sentences

- which is how _____ achieves **fermentation**
- _____ **starter** culture contains Bacillus subtilus
- Bacillus subtilis forms spores that are found in the _____
- _____ is frequently **eaten** in Japan over **rice**
- You can also eat _____ with condiments such as **soy sauce**
- **Soybeans** for _____ vary in size from super large to tiny ones

Other sentences without _____

- in **soybean fermentation** for the production of tempeh
- the tempeh **starter** contains spores of Rhizopus oligosporus
- Mapo tofu is usually **eaten** with **rice**
- Steamed tofu with a savory **soy sauce** garlic dressing
- **Soybeans** for tofu differ from those grown for livestock

Question: What does natto mean?

Sentences with natto

- which is how natto achieves **fermentation**
- natto **starter** culture contains Bacillus subtilus
- Bacillus subtilis forms spores that are found in the natto
- Natto is frequently **eaten** in Japan over **rice**
- You can also eat natto with condiments such as **soy sauce**
- **Soybeans** for natto vary in size from super large to tiny ones

Other sentences ...

- **Soybeans** for tofu differ from those grown for livestock
- in **soybean fermentation** for the production of tempeh
- the tempeh **starter** contains spores of Rhizopus oligosporus
- Mapo tofu is usually **eaten** with **rice**
- Steamed tofu with a savory **soy sauce** garlic dressing

Question: What does natto mean?

Sentences with natto

- which is how natto achieves **fermentation**
- natto **starter** culture contains Bacillus subtilus
- Bacillus subtilis forms spores that are found in the natto
- Natto is frequently **eaten** in Japan over **rice**
- You can also eat natto with condiments such as **soy sauce**
- **Soybeans** for natto vary in size from super large to tiny ones

Natto: a food made of (fermented) soybeans, like tempeh or tofu

Natto

source: serious-eats.com

Distributional Semantics

Defining word meaning by its use in language

- Ludwig Wittgenstein; Zellig Harris (1954)
- Words are defined by their environment
- If two words share the same environments, they are synonyms
- Word that have similar contexts tend to be similar in meaning

Distributional Semantics:

Vector Space Models and Embeddings

- The standard way to represent meaning in NLP
- Fine-grained model

Why using vectors as representations?

- Consider QA: "African restaurants nearby?"
  Feature "African" is a numeric word vector

$\Rightarrow$ We can compute its similarity to other words
"There is an Eritrean and South African restaurant nearby"

$\Rightarrow$ We can compute its similarity to other words
"There is an Eritrean and South African restaurant nearby"

- Consider Sentiment Analysis: "The book was hilarious"

$\Rightarrow$ allows for generalisation!
"The book was good."

Count-based Methods

Semantic Space Models: Term-term matrices

	drink	brewed	beans	exam	teacher
coffee	6891	1023	3028	34	0
tea	7732	643	138	20	0
school	2	0	0	5821	983
study	7	0	0	7112	892

Semantic Space Models: Term-term matrices

	drink	brewed	beans	exam	teacher
coffee	6891	1023	3028	34	0
tea	7732	643	138	20	0
school	2	0	0	5821	983
study	7	0	0	7112	892

The prices of arabica **coffee** *beans* are climbing to record highs . . .

Coffee can be *brewed* in several different ways . . .

Arabica **coffee** *beans* grow best when the temperature . . .

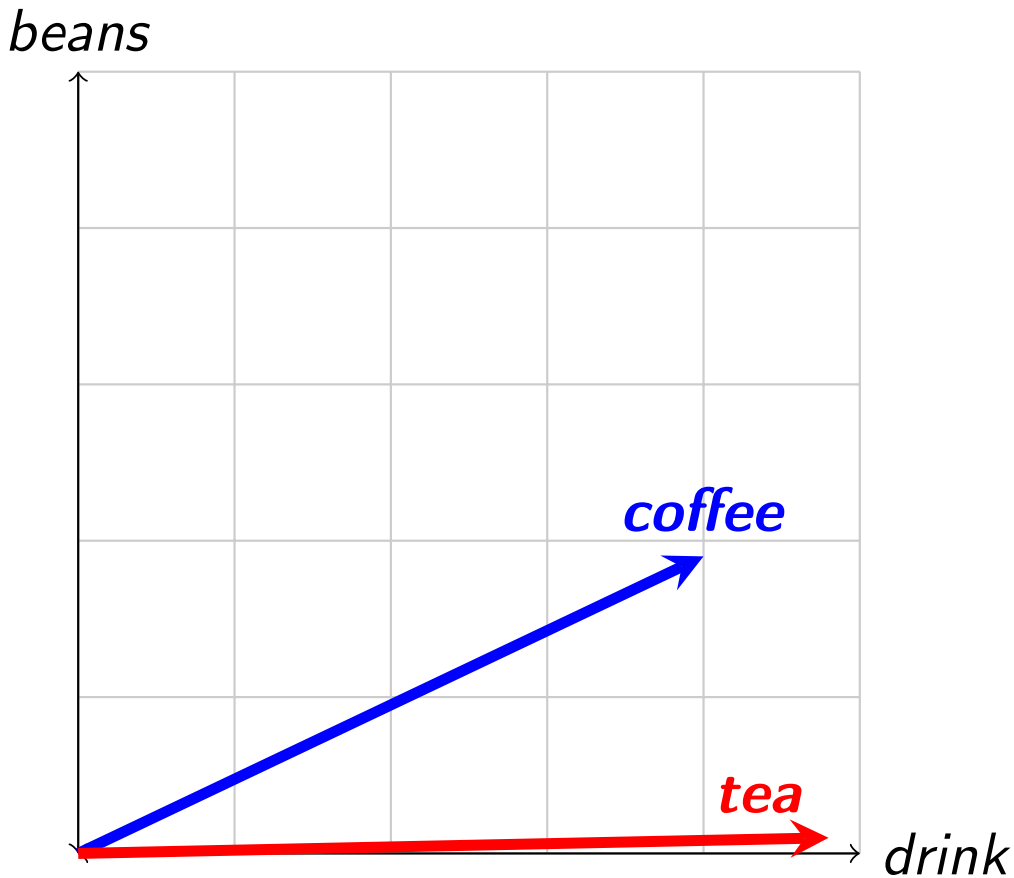
Once *brewed*, the desirable **coffee** flavors . . .

. . . how to **study** for an *exam* at a German university . . .

. . . arrange to take the *exam* at your **school** . . .

Semantic Space Models: Term-term matrices

Visualising word vectors



Vector representations/embeddings: Computing Word Similarity

Cosine Similarity[1]

The cosine of the angle between the two vectors **u**,**v**:

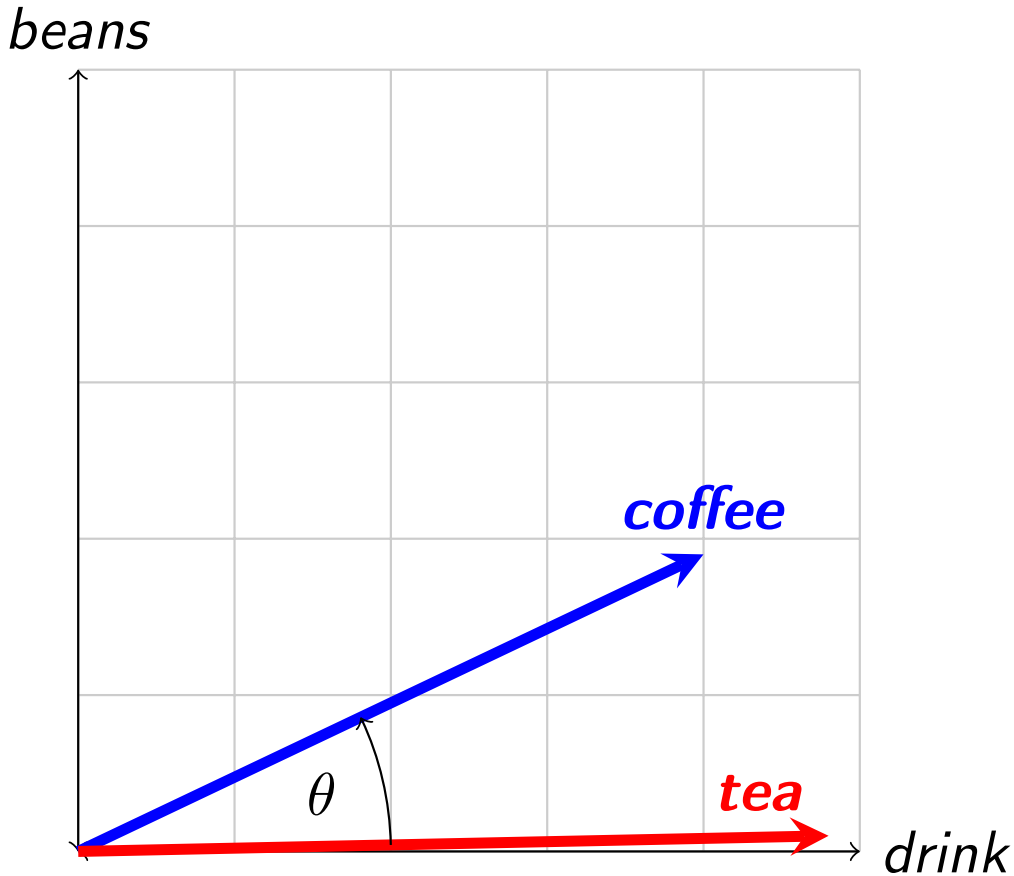
$$\text{cosine}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}| \cdot |\mathbf{v}|} = \frac{\sum_{i=1}^{N} u_i v_i}{\sqrt{\sum_{i=1}^{N} u_i} \cdot \sqrt{\sum_{i=1}^{N} v_i}}$$

- The normalised dot product between two vectors
- Normalised by the *lengths* of the vectors
- The cosine similarity is invariant to vector length

[1] The symbols **u** and $\vec{u}$ are used interchangeably to denote a vector.

Semantic Space Models: Term-term matrices

Visualising word vectors



Prediction-based Methods

Prediction-based models

Advantages

- Yield short, dense vectors (embeddings)
- Short: easier to use as features in machine learning (fewer weights to tune than long vectors)
- Dense: may generalise better than explicit counts
- They capture various patterns, like similarity and analogy
- In practice, they work better than count-based methods

Popular Models

\+ New words can be added easily, by resuming training

- word2vec (skipgram & CBOW)
  [Mikolov et al., 2013a, Mikolov et al., 2013b]
- GloVe [Pennington et al., 2014]

Prediction-based models

Popular Models

+ New words can be added easily, by resuming training
+ Very fast to train
  - word2vec (skipgram & CBOW) [Mikolov et al., 2013a]
    - Skip-gram: Given target word, predict its context words
    - CBOW: Given context words, predict target word (continuous bag-of-words)
    - Frequent words are subsampled
    - scales with corpus size
  - GloVe [Pennington et al., 2014]
    - Hybrid model: starts with co-occurrence-counts, followed by prediction-based learning
    + fast training (co-occurrence with one sweep through corpus)
    + good on small corpora, scalable to huge corpora

Prediction-based models: The word2vec framework

[Mikolov et al., 2013a, Mikolov et al., 2013b]

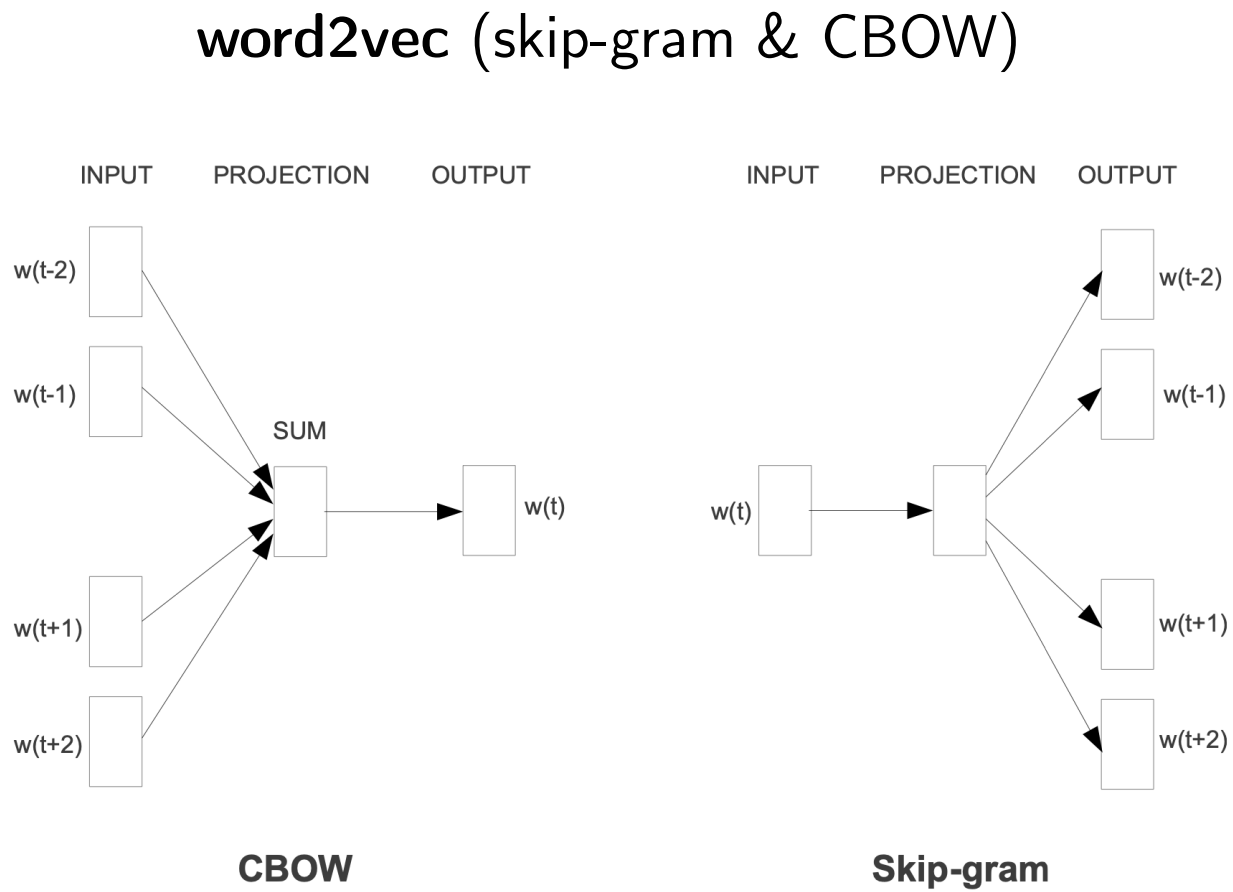


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

Prediction-based models: The word2vec framework

[Mikolov et al., 2013a, Mikolov et al., 2013b]

- Don't count how often "coffee" occurs near *beans*

Prediction-based models: The word2vec framework

[Mikolov et al., 2013a, Mikolov et al., 2013b]

- Don't count how often "coffee" occurs near *beans*
- Instead, train a classifier on a binary prediction task:
  - *Is beans likely to occur near "coffee"*?
- Goal: Learn classifier weights $\Rightarrow$ the embeddings
- Idea: Self-supervised learning
  - Correct answers: Words that occur near "coffee"
  - No human labels needed

Linguistic/Semantic Representations

Static Representations

- Count-based or prediction-based models
- One vector for each word **type**
- Common baseline models
- Methods: Counts, SVD; word2vec, GloVe, Fasttext, ...

Current Linguistic Representations

Transformer-based Methods /
Large Language Models (LLMs)

Contextualised Embeddings

What is the meaning of a word in a particular context?

"They caress their colt every evening."
"They clean their colt every evening."
"They put their colt into the safe every evening."
"bert talks nonsense"
"apple releases a new mac"
"the kid loves the apple and hates the pear"
"they visited the big apple dozens of times"

How can we incorporate linguistic context into word embeddings?

$\Leftrightarrow$ How can we represent individual word tokens?

Transformers: Basic Components

How can we encode a word as a vector that captures its meaning in a particular context?

Question: What and how much of is the contribution of each context item to the meaning of “colt” and “apple”, resp.?

Transformers: Basic Components

How can we encode a word as a vector that captures its meaning in a particular context?

Question: What and how much of is the contribution of each context item to the meaning of "colt" and "apple", resp.?

- "They caress their colt every evening."
- "The kid loves the apple and hates the pear."

Transformers: Basic Components

How can we encode a word as a vector that captures its meaning in a particular context?

Question: What and how much of is the contribution of each context item to the meaning of "colt" and "apple", resp.?

- "They caress their colt every evening."
- "The kid loves the apple and hates the pear."

$\Rightarrow$ A words's meaning (representation) in a particular context depends on its own embedding and the embeddings of the context items *cf. compositionality*

Neural Network Architectures for Linguistic Sequences

Transformers [Vaswani et al., 2017]

- no recurrent connections
- fully connected networks
- efficient: parallel matrix computations

Fine-tuning stage for classification tasks [Howard and Ruder, 2018]

The following slides are based on material from Chris Potts (Stanford cs22u, 2021); J. Alammar's blog; Ch.9, Jurafsky & Martin, SLP3; [Vaswani et al., 2017]; [?]

Transformers: Basic Components

Self-attention: Basic Principle [Vaswani et al., 2017]

- *Attention weights*
  - Obtain proportional amount of attention to pay to each context word for the target word
    ← How relevant is each context word to the target word?

Transformers: Basic Components

Self-attention: Basic Principle [Vaswani et al., 2017]

- *Attention weights*
  - Obtain proportional amount of attention to pay to each context word for the target word
    ← How relevant is each context word to the target word?
  - Based on similarity between each target and context word input embedding
- Obtain contextualised embedding: target embedding combined with weighted mean of context embeddings

Transformers: Overall Architecture [Vaswani et al., 2017]

Training a transformer as a language model

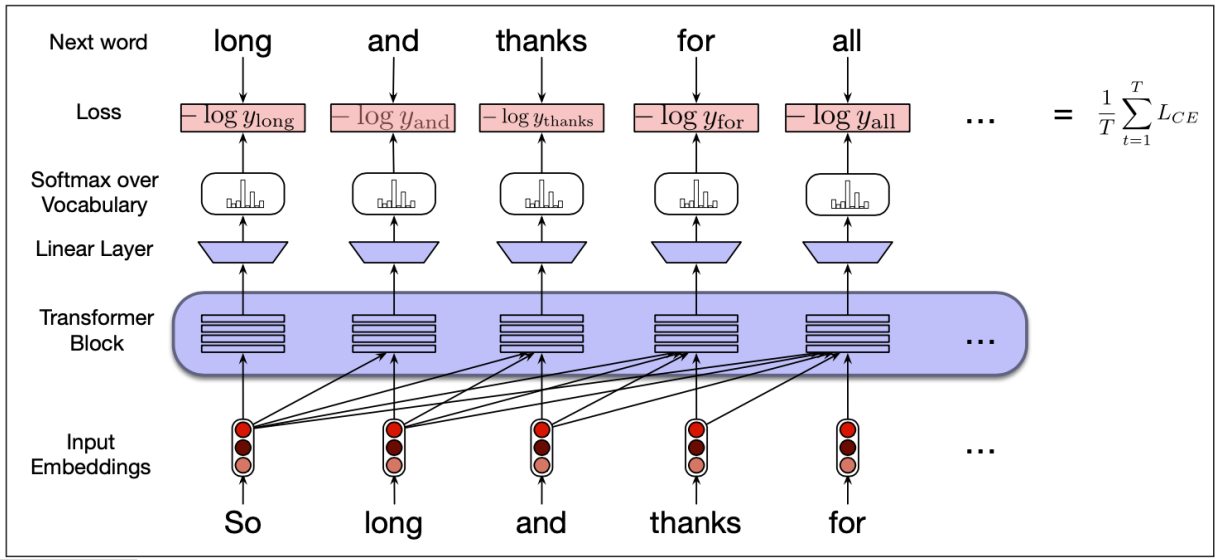


Figure 10.7 Training a transformer as a language model.

[Jurafsky and Martin, 2023]

here: Output == Input

Transformers: Using the Embeddings

Fine-tuning on Task

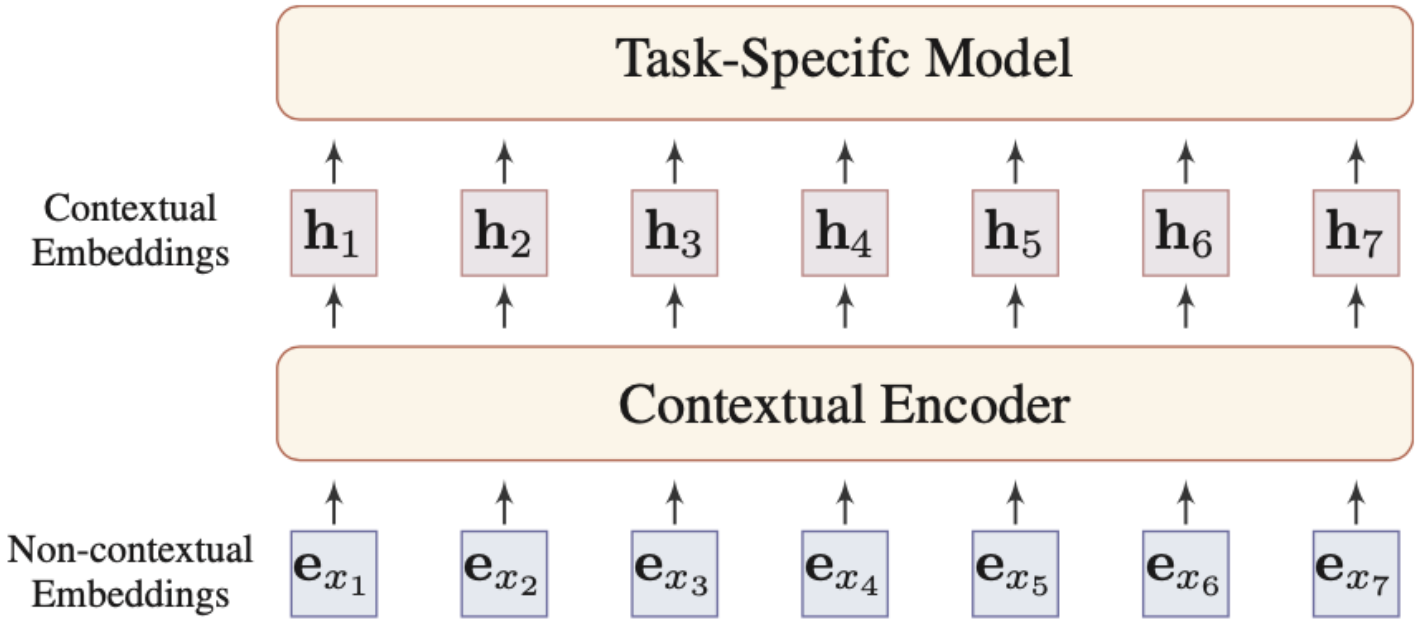


Figure 1: Generic Neural Architecture for NLP

Existing Models: Architecture Types

Decoders vs. Encoders

- **Decoders**: Language Models – can only attend to left context

  "They caress their ? ... ..."

- **Encoders**: *Masked* Language Models – attend to left and right context aka *cloze task*

  "They caress their ? every evening"

GPT, GPT-2, GPT-3: Generative Pre-Training

[Arora et al., 2020, Radford et al., 2019, Brown et al., 2020]

Transformer-based Language Models

- GPT: Language Model – token can only attend to left context
- Can be used for *autoregressive* text generation

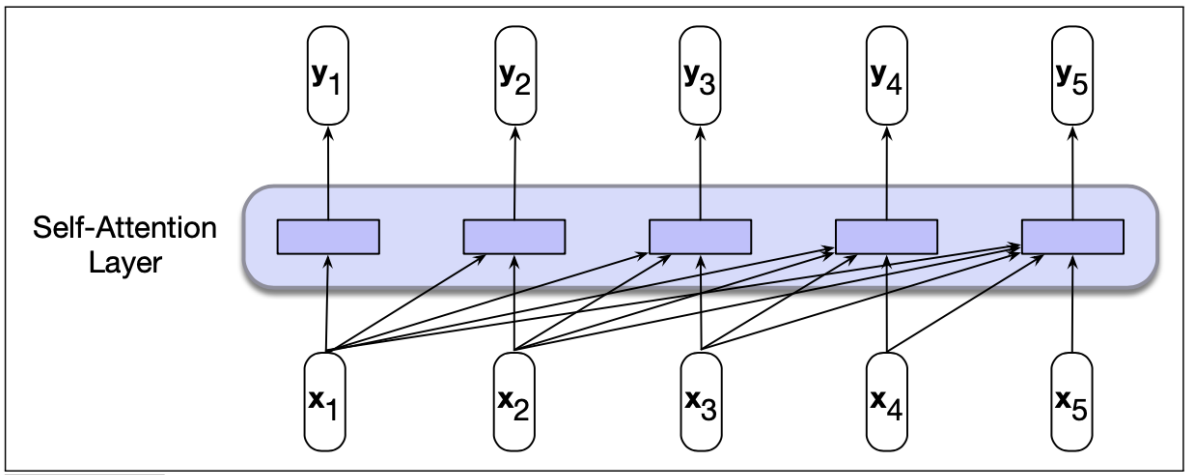


Figure 11.1 A causal, backward looking, transformer model like Chapter 10. Each output is computed independently of the others using only information seen earlier in the context.

BERT: Unsupervised Pre-training [Devlin et al., 2019]

Bidirectional Encoder Representations from Transformers

$\Rightarrow$ Trained with Masked Language Modelling
(aka cloze task) cf. CBOW [Mikolov et al., 2013a]

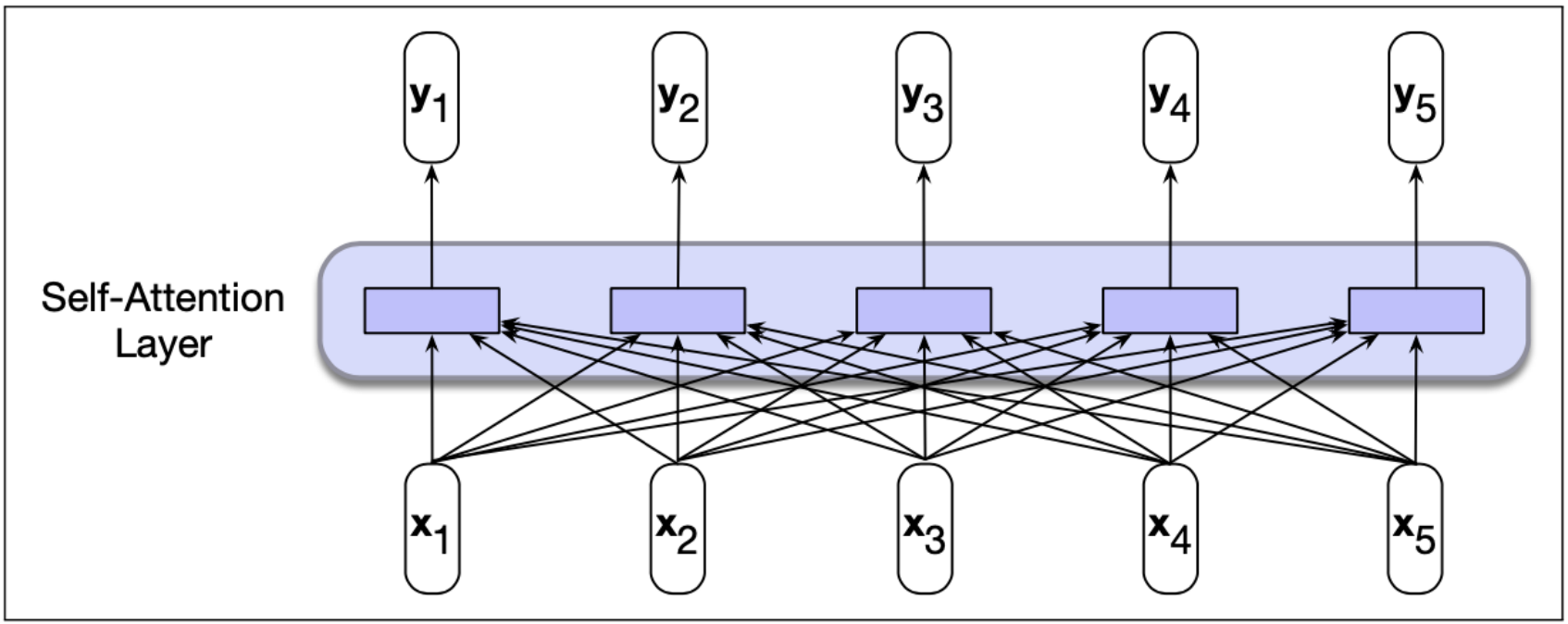
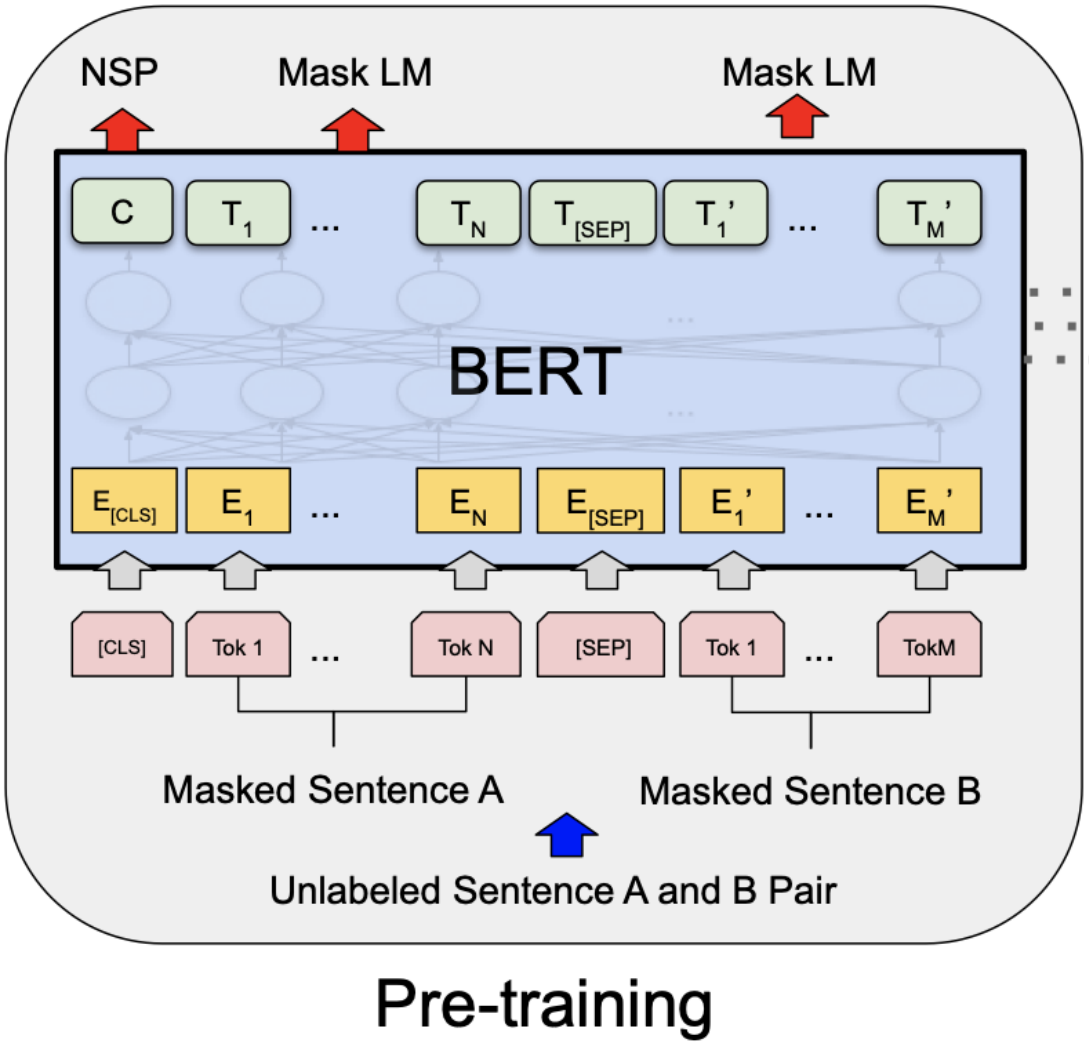


Figure 11.2 Information flow in a bidirectional self-attention model. In processing each element of the sequence, the model attends to all inputs, both before and after the current one.

BERT: Unsupervised Pre-training [Devlin et al., 2019]

Bidirectional Encoder Representations from Transformers



- Pre-training tasks: next sentence prediction and masked LM

BERT: Unsupervised Pre-training [Devlin et al., 2019]

Bidirectional Encoder Representations from Transformers

Pre-training task: Masked Language Modeling (MLM)

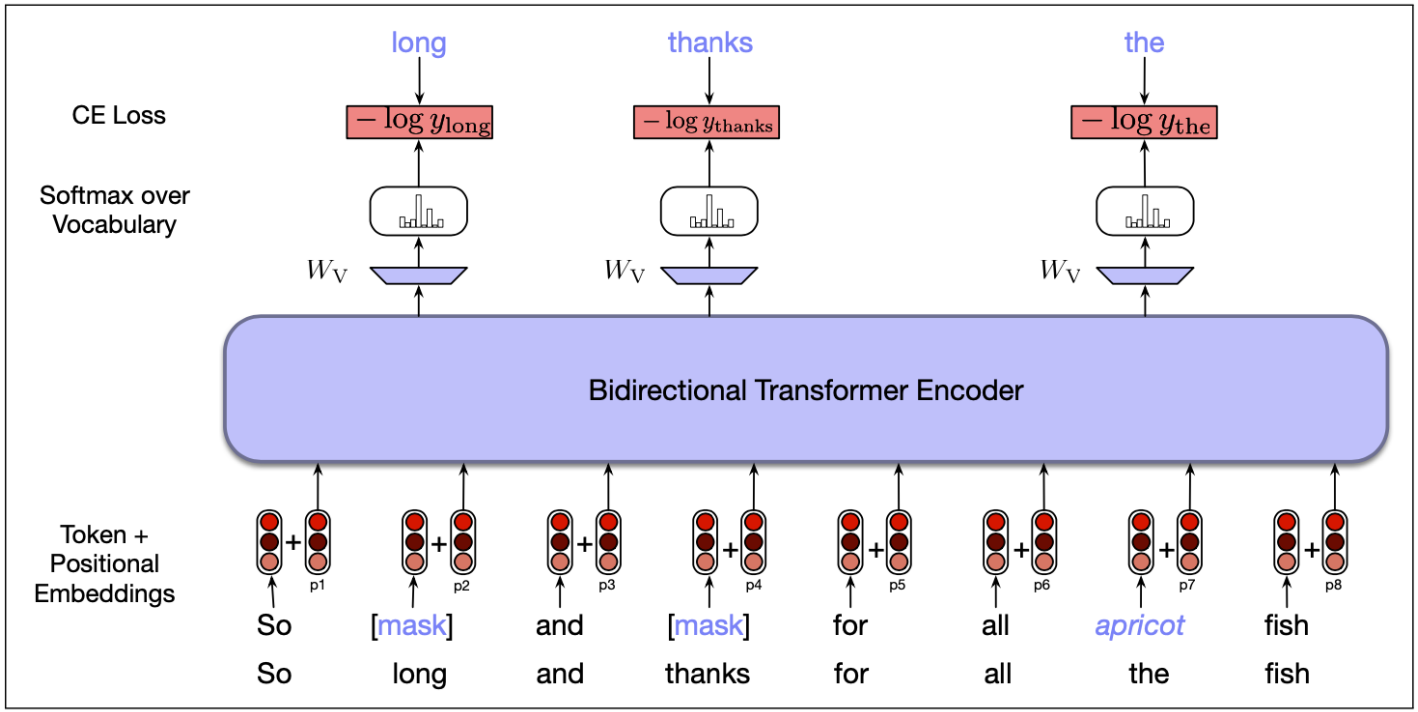


Figure 11.5 Masked language model training. In this example, three of the input tokens are selected, two of which are masked and the third is replaced with an unrelated word. The probabilities assigned by the model to these three items are used as the training loss. (In this and subsequent figures we display the input as words rather than subword tokens; the reader should keep in mind that BERT and similar models actually use subword tokens instead.)

BERT: Unsupervised Pre-training [Devlin et al., 2019]

Bidirectional Encoder Representations from Transformers

Pre-training task: Next Sentence Prediction (NSP)

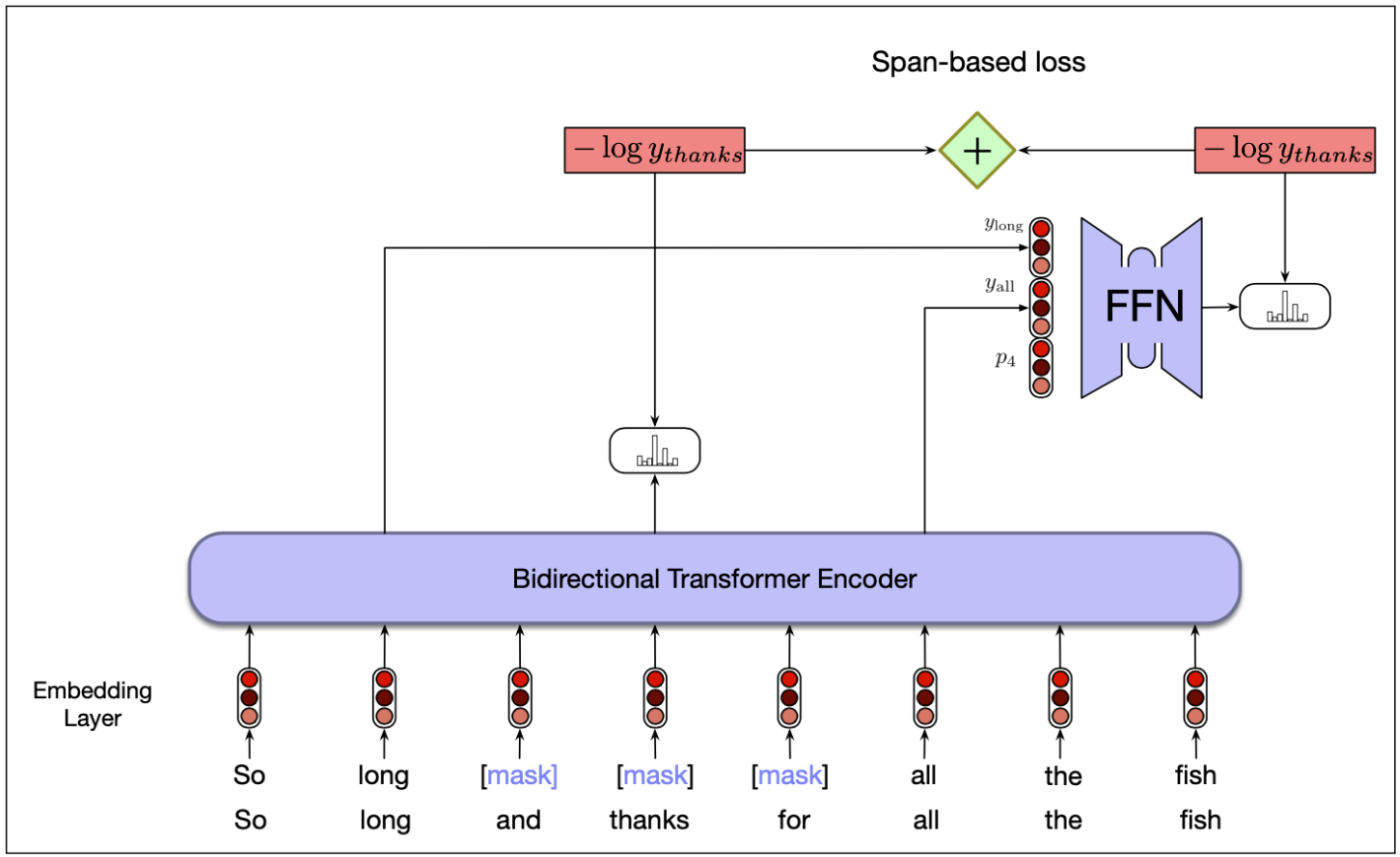


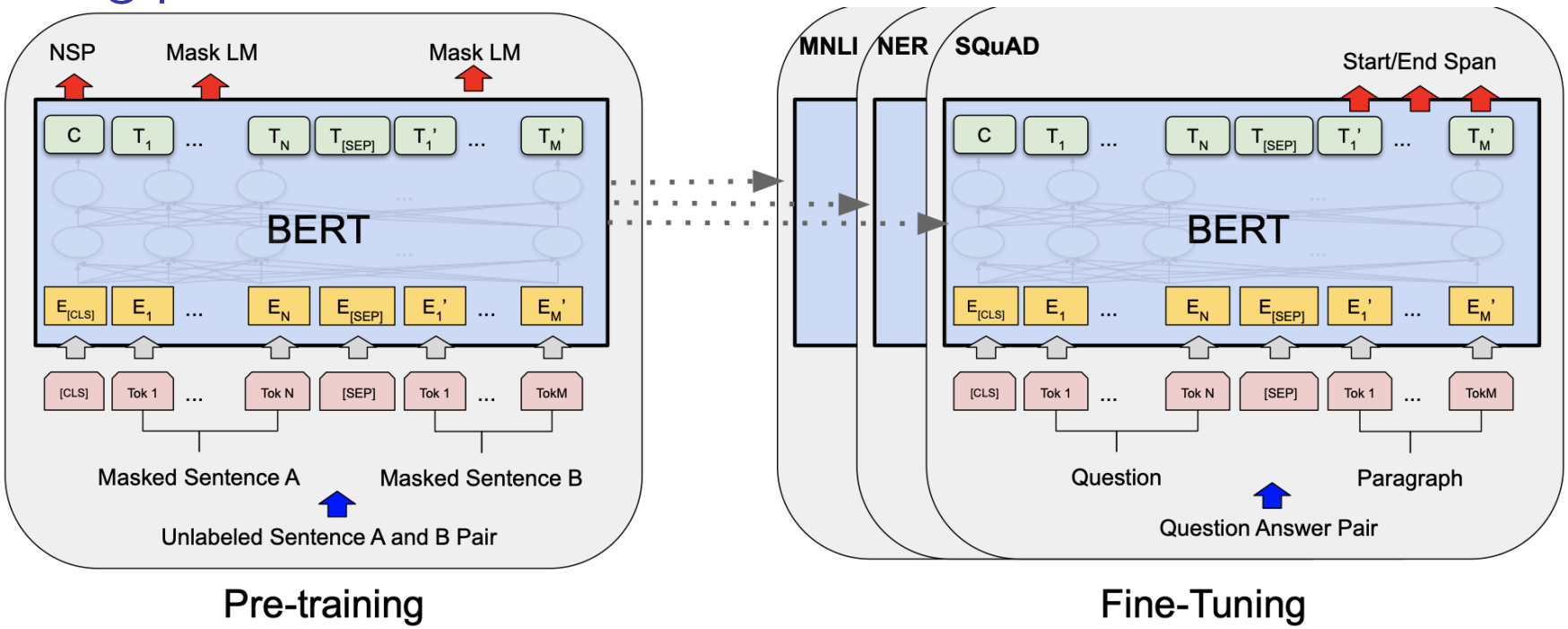
Figure 11.6 Span-based language model training. In this example, a span of length 3 is selected for training and all of the words in the span are masked. The figure illustrates the loss computed for word *thanks*; the loss for the entire span is based on the loss for all three of the words in the span.

BERT: Unsupervised Pre-training

[Devlin et al., 2019]

Bidirectional Encoder Representations from Transformers

Using pre-trained BERT for tasks



- Final hidden state of [CLS]: used as the aggregate sequence representation for classification tasks

BERT: Unsupervised Pre-training [Devlin et al., 2019]

Bidirectional Encoder Representations from Transformers

Architecture

- Stack of transformer encoders: block of self-attention layer + feed-forward neural network
- Input: Spans of contiguous text and special tokens
  - First token of every sequence[2]: always a special classification token [CLS]
    final hidden state of [CLS]: used as the aggregate sequence representation for classification tasks
  - Sentences are separated through [SEP] token
  - Learned input embedding for each token, indicating if it belongs to sentence A or B
  - Positional embedding for each token

[2] 1-2 spans of contiguous text

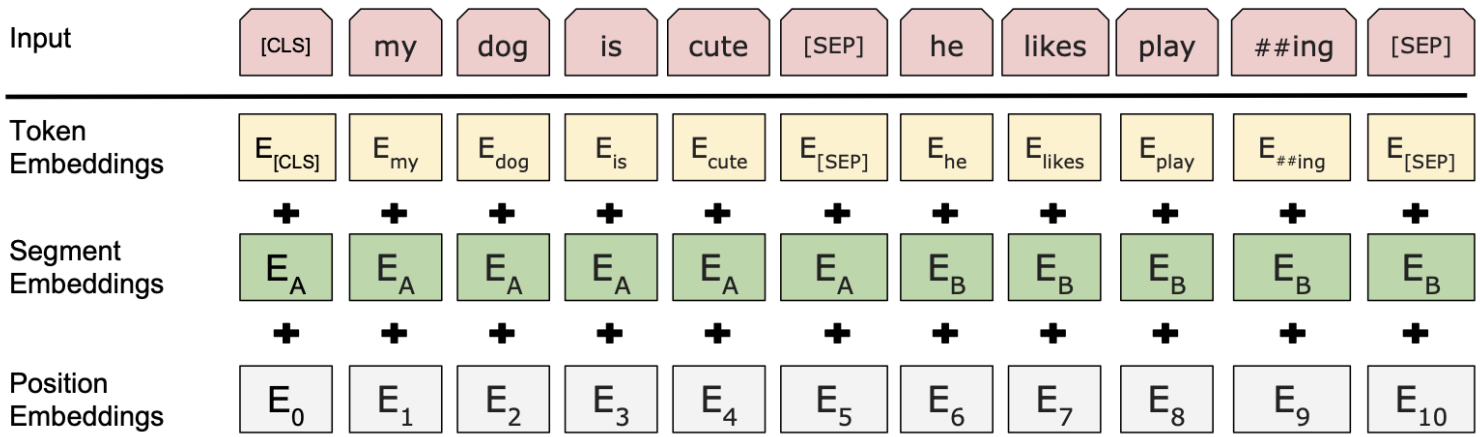
BERT: Unsupervised Pre-training

[Devlin et al., 2019]

Bidirectional Encoder Representations from Transformers

Architecture: Inputs

- Stack of transformer encoders
- Input: 3 embeddings for each token



Large Language Models: LLaMA

https://llama.meta.com

- Auto-regressive foundation LM
- Trained on publicly available data

Linguistic/Semantic Representations

Static Representations

- Count-based or prediction-based models
- One vector for each word **type**
- Common baseline models
- Methods: Counts, SVD; word2vec, GloVe, Fasttext, ...

Linguistic/Semantic Representations

Static Representations

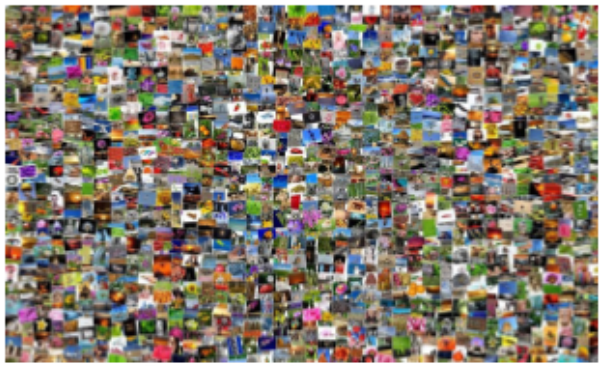
- Count-based or prediction-based models
- One vector for each word **type**
- Common baseline models
- Methods: Counts, SVD; word2vec, GloVe, Fasttext, ...

Contextualised Representations

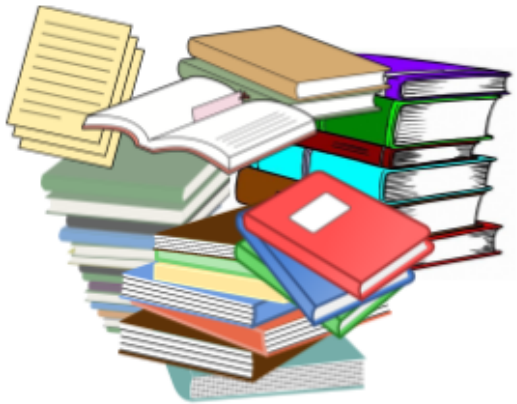
- Transformer-based (special deep NN) [Vaswani et al., 2017]
- One embedding for each **word *token*** in a text
- State-of-the-art
- Methods: BERT, ELMO, GPT, ELECTRA, RoBERTa, ...

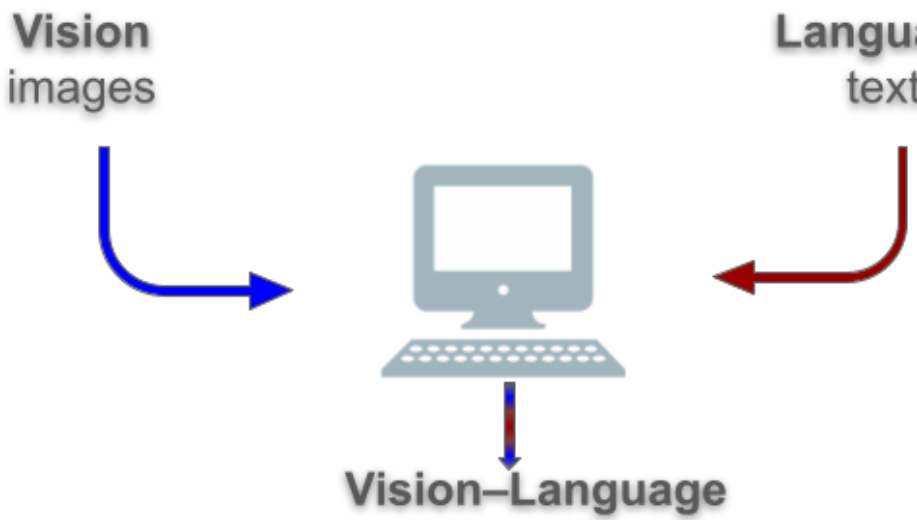
Common to all: Underlying principle is the distributional hypothesis

Multimodal (VL) Representations



https://pixabay.com/photos/photo-collection-pictures-photos-382018





References I

Andrews, M., Vigliocco, G., and Vinson, D. (2009).
Integrating experiential and distributional data to learn semantic representations.
Psychological review, 116(3):463.

Arora, S., May, A., Zhang, J., and Ré, C. (2020).
Contextual embeddings: When are they worth it?
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2663, Online. Association for Computational Linguistics.

Barsalou, L. W. (2008).
Grounded Cognition.
Annual Review of Psychology, 59(1):617–645.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020).
Language models are few-shot learners.
In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901.

Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K. (2012).
Distributional semantics in technicolor.
In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019).
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

References II

Howard, J. and Ruder, S. (2018).

Universal language model fine-tuning for text classification.
In *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 1, pages 328–339. Association for Computational Linguistics.

Johns, B. T. and Jones, M. N. (2012).

Perceptual inference through global lexical similarity.
Topics in Cognitive Science, 4(1):103–120.

Jurafsky, D. and Martin, J. H. (2023).

Speech and Language Processing (3rd ed. draft).

Kiela, D. and Bottou, L. (2014).

Learning image embeddings using convolutional neural networks for improved multi-modal semantics.
In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 36–45. ACL.

Lazaridou, A., Baroni, M., et al. (2015).

Combining language and vision with a multimodal skip-gram model.
In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a).

Efficient estimation of word representations in vector space.
In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

References III

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b).
Distributed representations of words and phrases and their compositionality.
In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Pennington, J., Socher, R., and Manning, C. D. (2014).
Glove: Global vectors for word representation.
In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019).
Language models are unsupervised multitask learners.

Silberer, C., Ferrari, V., and Lapata, M. (2017).
Visually grounded meaning representations.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11):2284–2297.

Silberer, C. and Lapata, M. (2012).
Grounded models of semantic representation.
In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, Jeju Island, Korea. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).
Attention is all you need.
Advances in neural information processing systems, 30.