What If We Encoded Words as Matrices and Used Matrix Multiplication as Composition Function?

***Lukas Galke**, Florian Mai, Ansgar Scherp*

Kiel University (DE), Idiap Research Institute (CH), University of Essex (UK)

Extended abstract of: F Mai, L Galke, A Scherp: *CBOW Is Not All You Need: Combining CBOW with the Compositional Matrix Space Model*, ICLR 2019.

Motivation

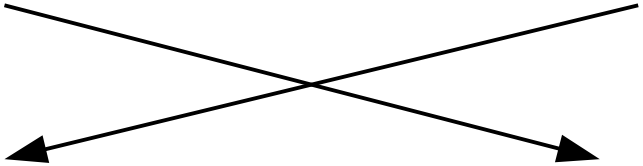
- Word embeddings (Collobert & Weston, ICML 2008; Mikolov et al., NeurIPS 2013)
  - Learn once from unlabelled text, re-use often
- Drawback of aggregated word embeddings
  - Compositions are not *order-aware*

Motivation

The movie was not great, it was rather awful.

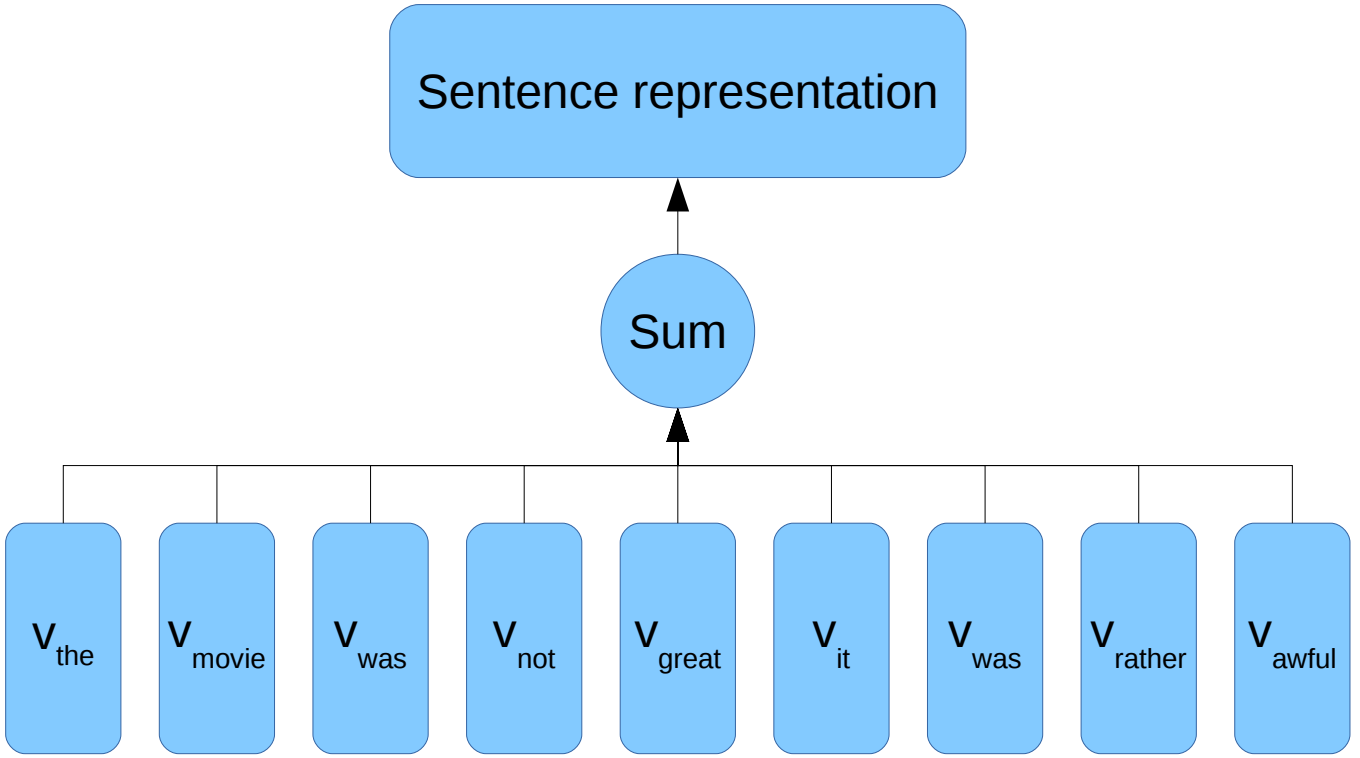
Motivation

The movie was not great, it was rather awful.



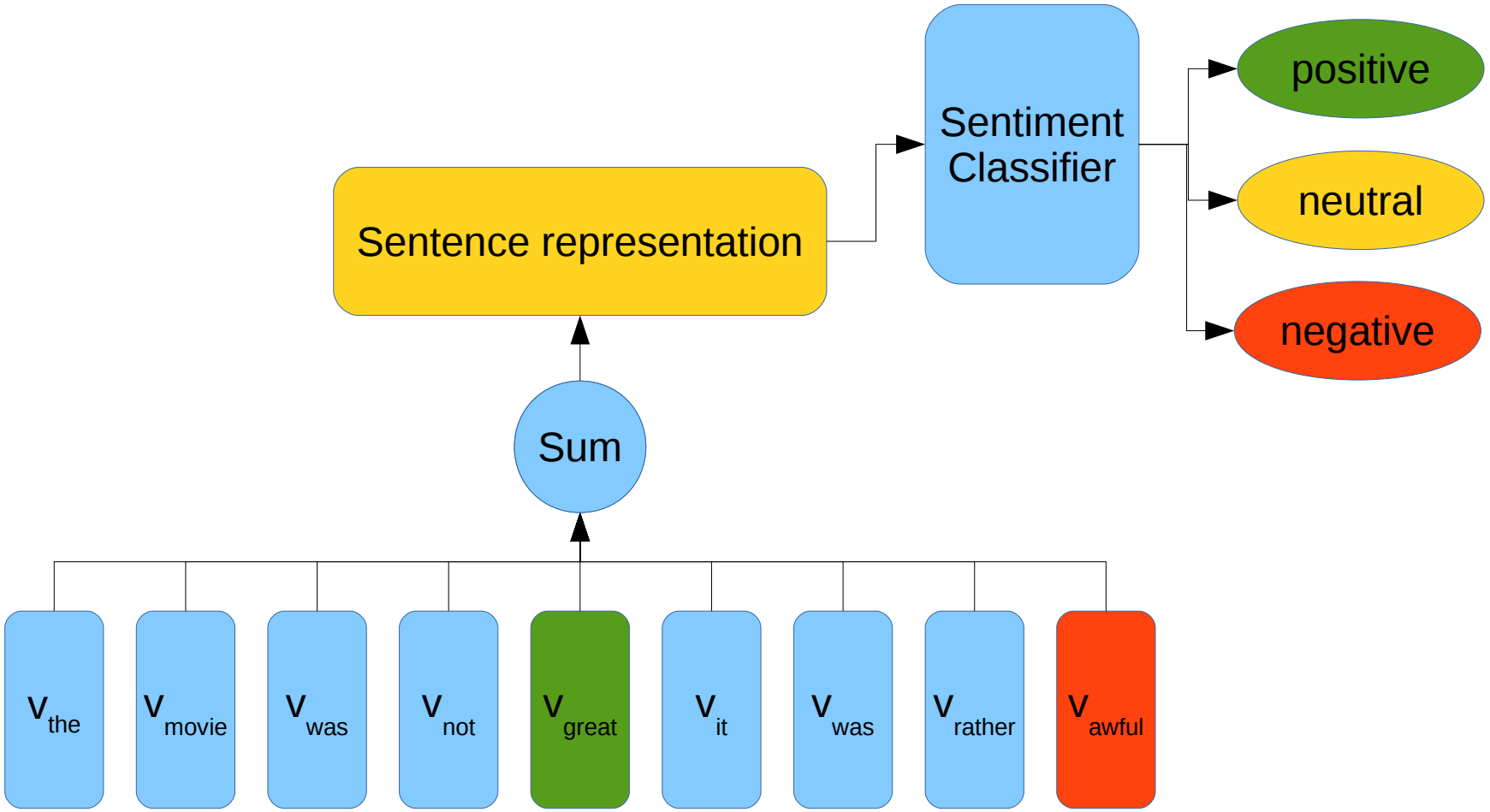
The movie was not awful, it was rather great.

Continuous Bag of Words (CBOW)



The movie was not great, it was rather awful.

Classification with CBOW

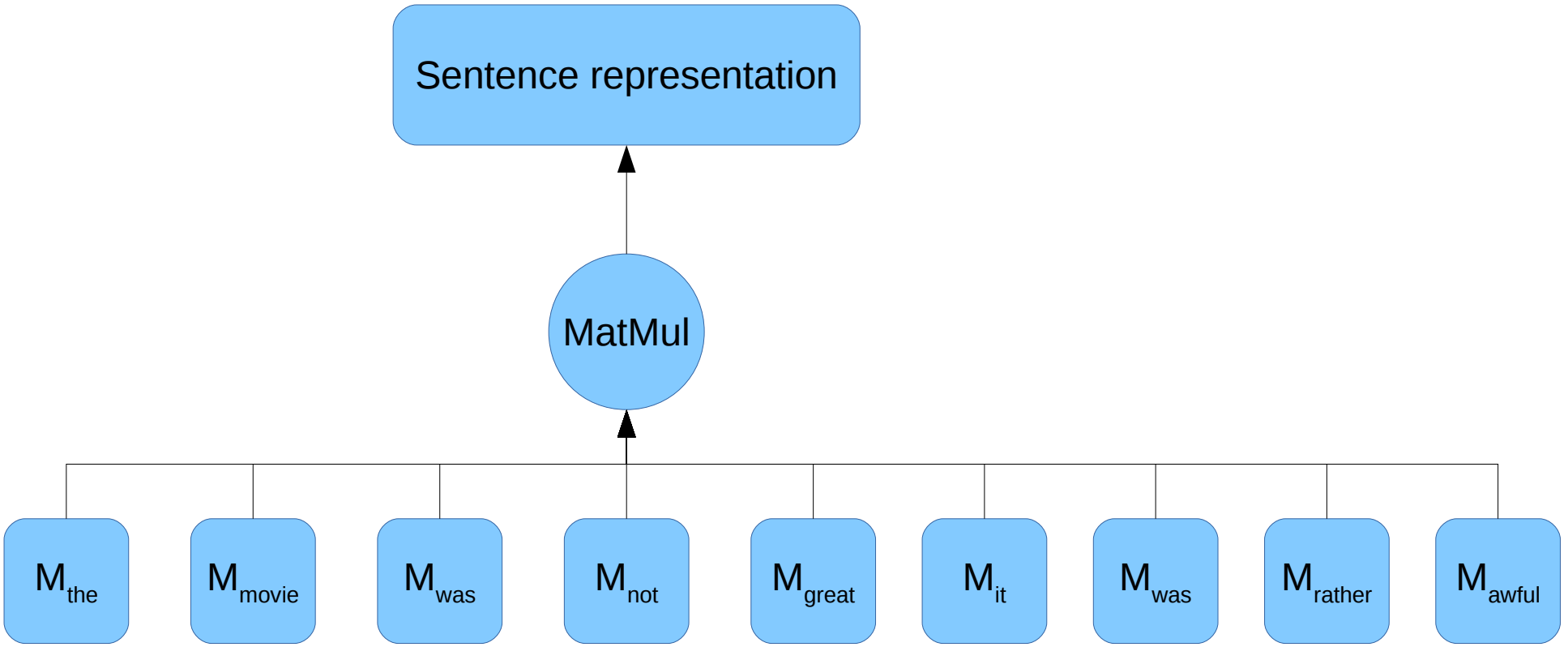


The movie was not great, it was rather awful.

Key Idea

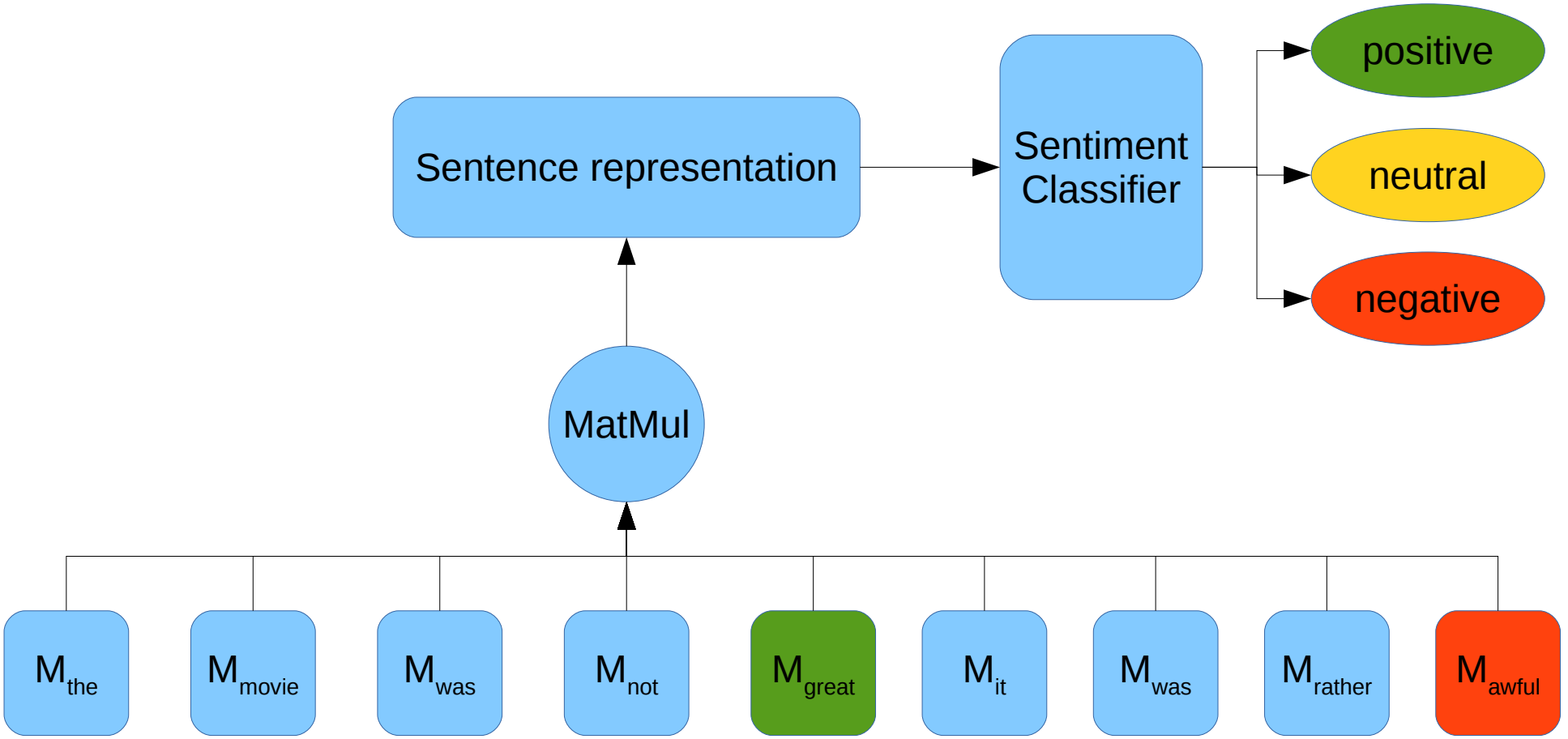
Use matrix multiplication
as composition function.

Continual Matrix Multiplication (CMOW)



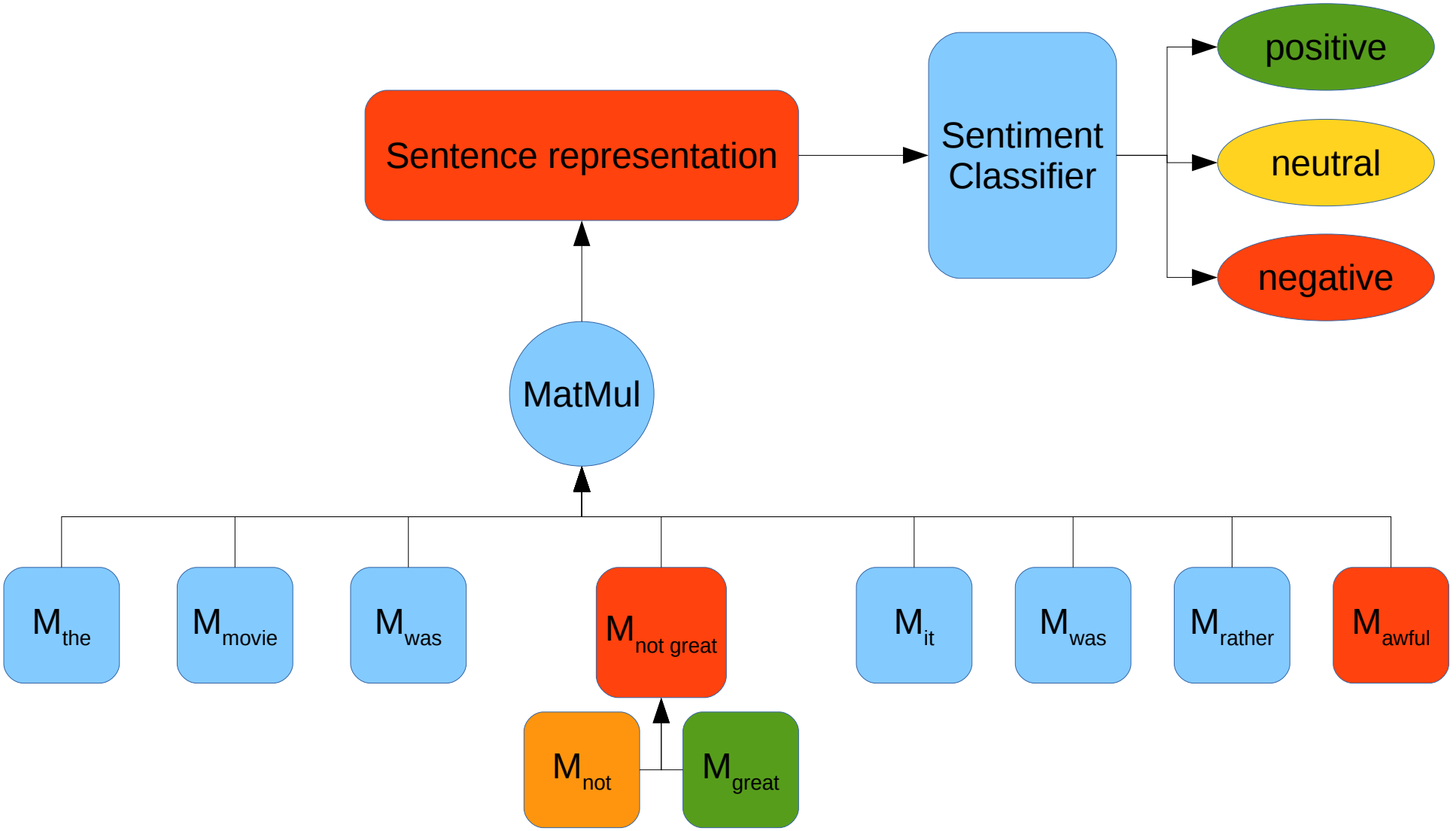
The movie was not great, it was rather awful.

Classification with CMOW



The movie was not great, it was rather awful.

Classification with CMOW

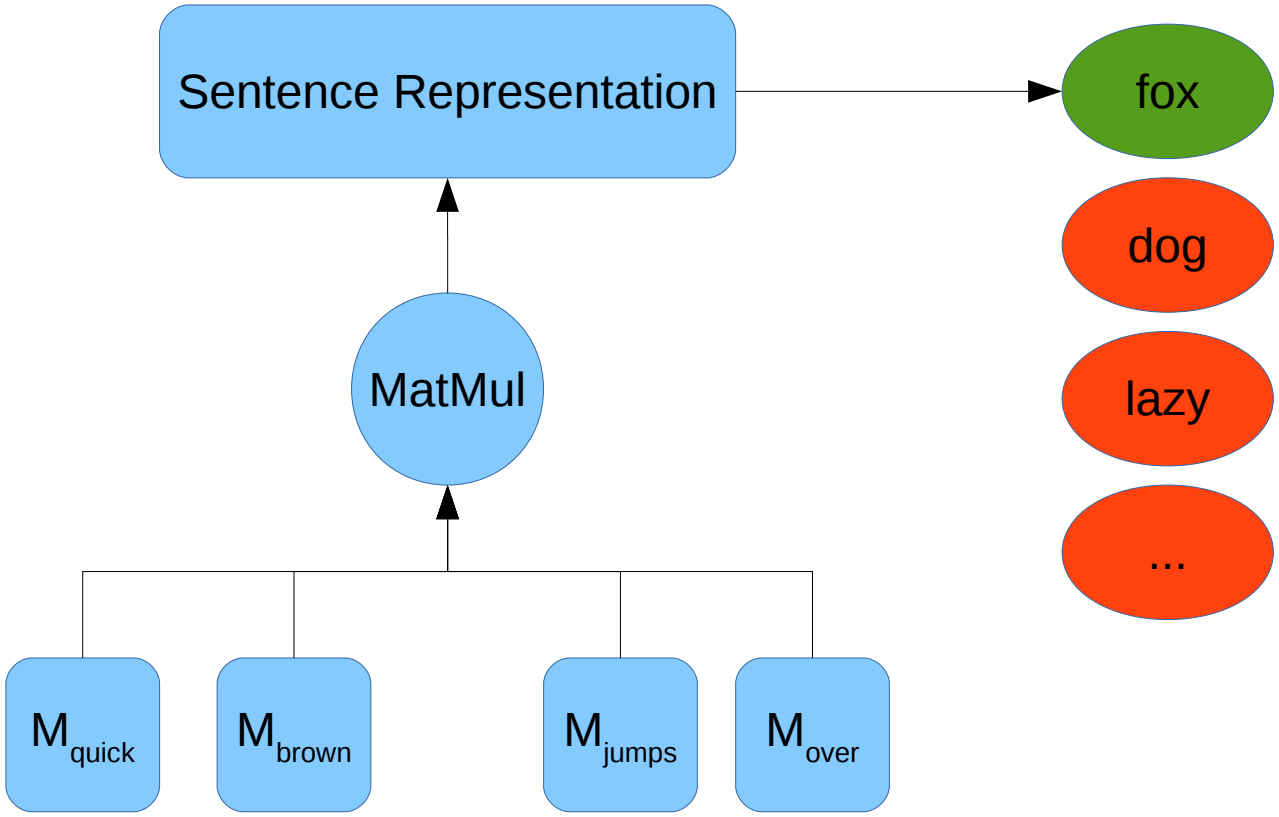


The movie was not great, it was rather awful.

Unsupervised Training

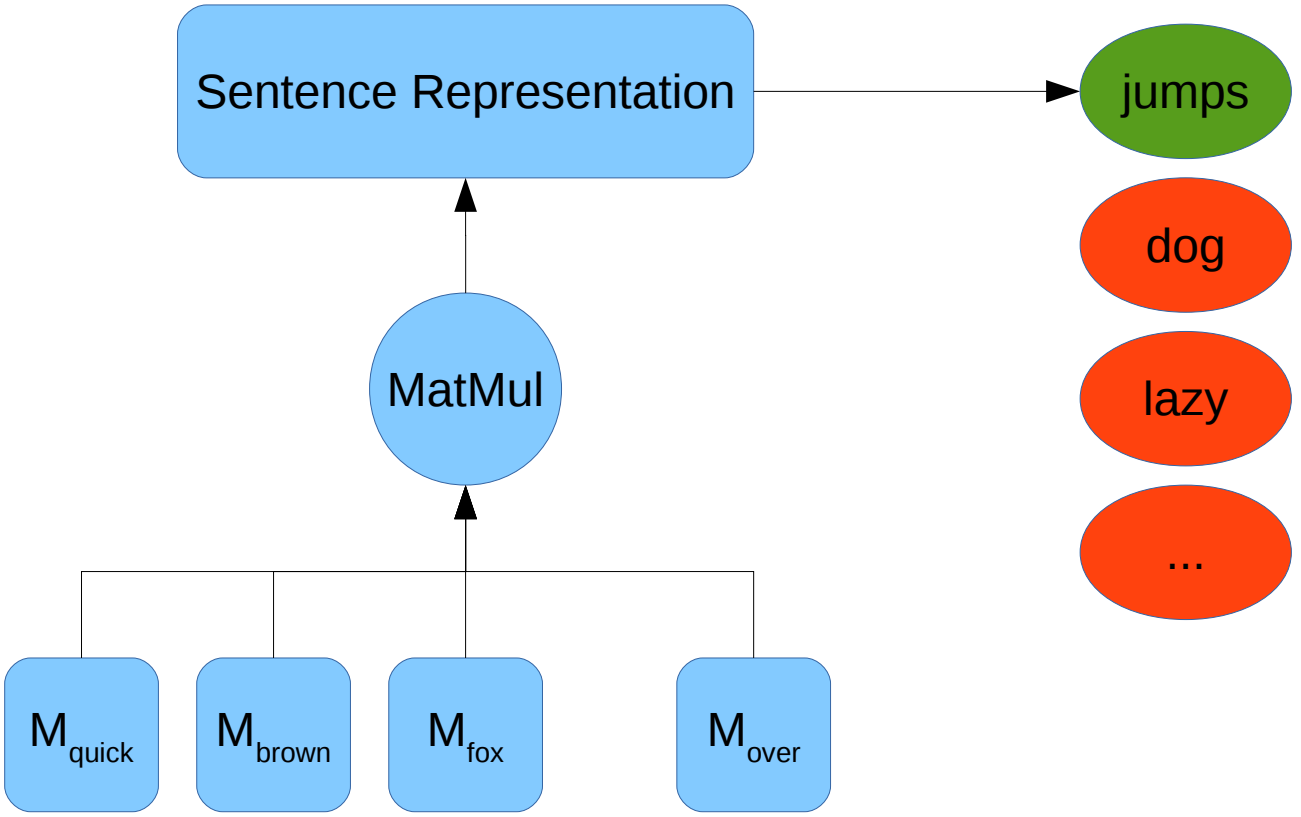
- Training as in Word2Vec (Mikolov et al, NeurIPS 2013)
- Matrix multiplication as composition function
- Initialization close to identity matrix
  - Such that values don’t vanish or explode
- Randomize target word
  - To alleviate bias regarding center words

Unsupervised Training



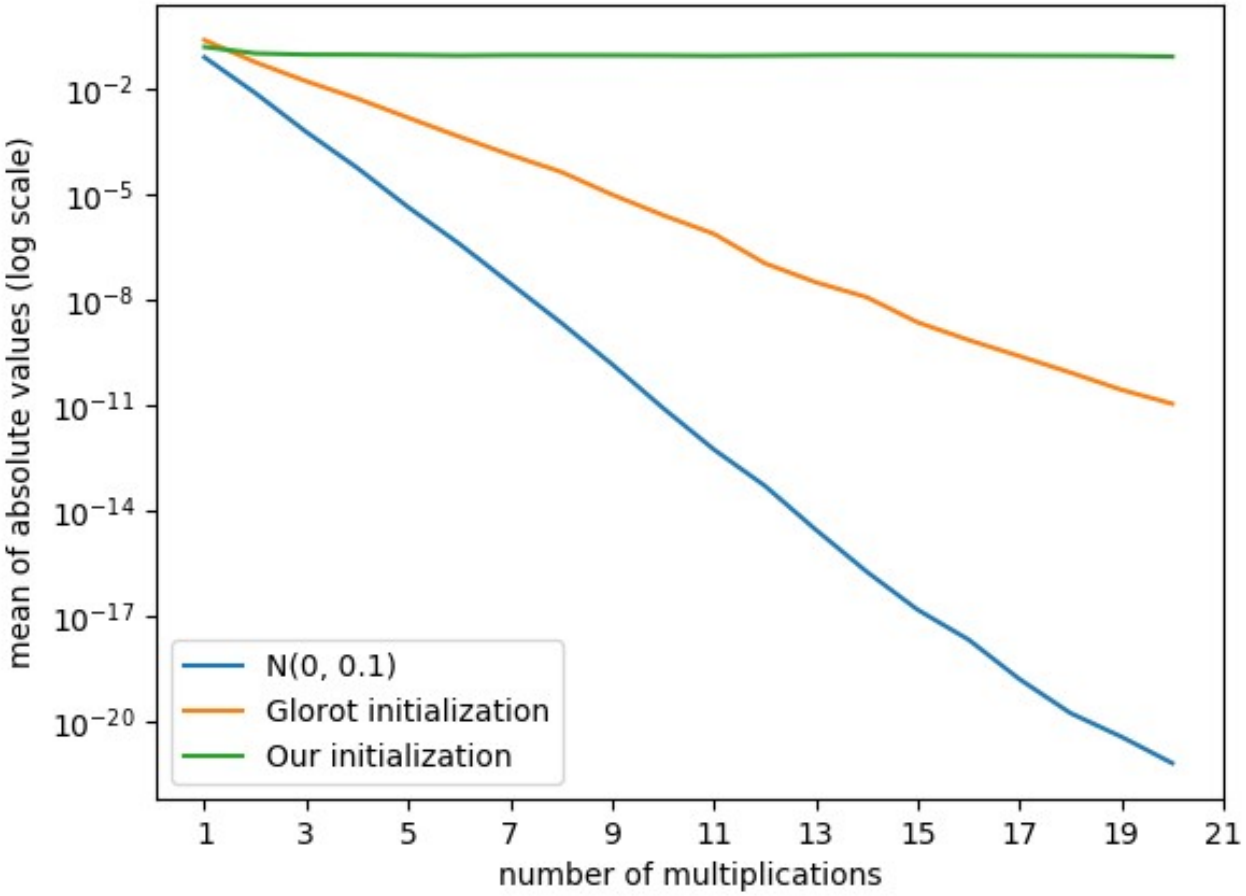
The quick brown ___ jumps over the lazy dog.

Training: Randomize Target Word



The quick brown fox ___ over the lazy dog.

Initialize Close to Identity Matrix



Ours: Initialize word matrices as $I_n + N(0, 0.1)$

Fair Evaluation

- Same training data: 3B tokens
- Same vocabulary size: 30k
- Same number of parameters
- 10 linguistic probing tasks
- 16 downstream tasks

Results: CBOW vs CMOW

Linguistic Probing Tasks

Method	Depth	BShift	SubjNum	Tense	Coord Inv	Length	Obj Num	TopConst	SOMO	WordContent
CBOW	33.0	49.6	79.3	78.4	53.6	74.5	78.6	72.0	49.6	**89.5**
CMOW	**35.1**	**70.8**	**82.0**	**80.2**	**61.8**	**82.8**	**79.7**	**74.2**	**50.7**	72.9

Supervised Downstream Tasks

Method	SUBJ	CR	MR	MPQA	MRPC	TREC	SICK-E	SST2	SST5	STS-B	SICK-R
CBOW	**90.0**	**79.2**	**74.0**	87.1	**71.6**	85.6	**78.9**	**78.5**	**42.1**	**61.0**	**78.1**
CMOW	87.5	73.4	70.6	**87.3**	69.6	**88.0**	77.2	74.7	37.9	56.5	76.2

Unsupervised Downstream Tasks

Method	STS12	STS13	STS14	STS15	STS16
CBOW	**43.5**	**50.0**	**57.7**	**63.2**	**61.0**
CMOW	39.2	31.9	38.7	49.7	52.2

More details on tasks: github.com/facebookresearch/SentEval

Results: CBOW vs CMOW

Linguistic Probing Tasks

Method	Depth	BShift	SubjNum	Tense	Coord Inv	Length	Obj Num	TopConst	SOMO	WordContent
CBOW	33.0	49.6	79.3	78.4	53.6	74.5	78.6	72.0	49.6	**89.5**
CMOW	**35.1**	**70.8**	**82.0**	**80.2**	**61.8**	**82.8**	**79.7**	**74.2**	**50.7**	72.9

Supervised Downstream Tasks

Method	SUBJ	CR	MR	MPQA	MRPC	TREC	SICK-E	SST2	SST5	STS-B	SICK-R
CBOW	**90.0**	**79.2**	**74.0**	87.1	**71.6**	85.6	**78.9**	**78.5**	**42.1**	**61.0**	**78.1**
CMOW	87.5	73.4	70.6	**87.3**	69.6	**88.0**	77.2	74.7	37.9	56.5	76.2

Unsupervised Downstream Tasks

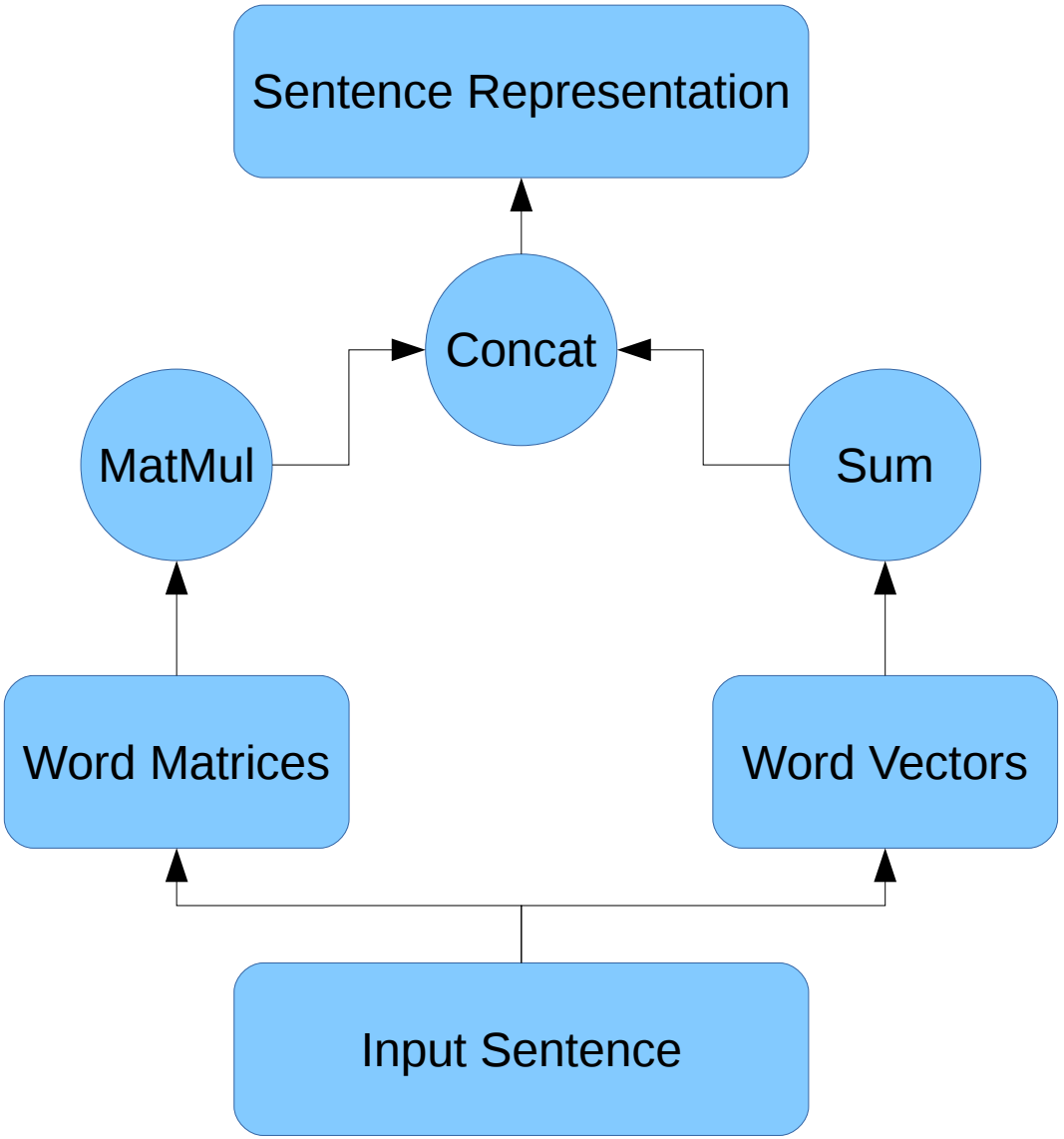
Method	STS12	STS13	STS14	STS15	STS16
CBOW	**43.5**	**50.0**	**57.7**	**63.2**	**61.0**
CMOW	39.2	31.9	38.7	49.7	52.2

More details on tasks: github.com/facebookresearch/SentEval

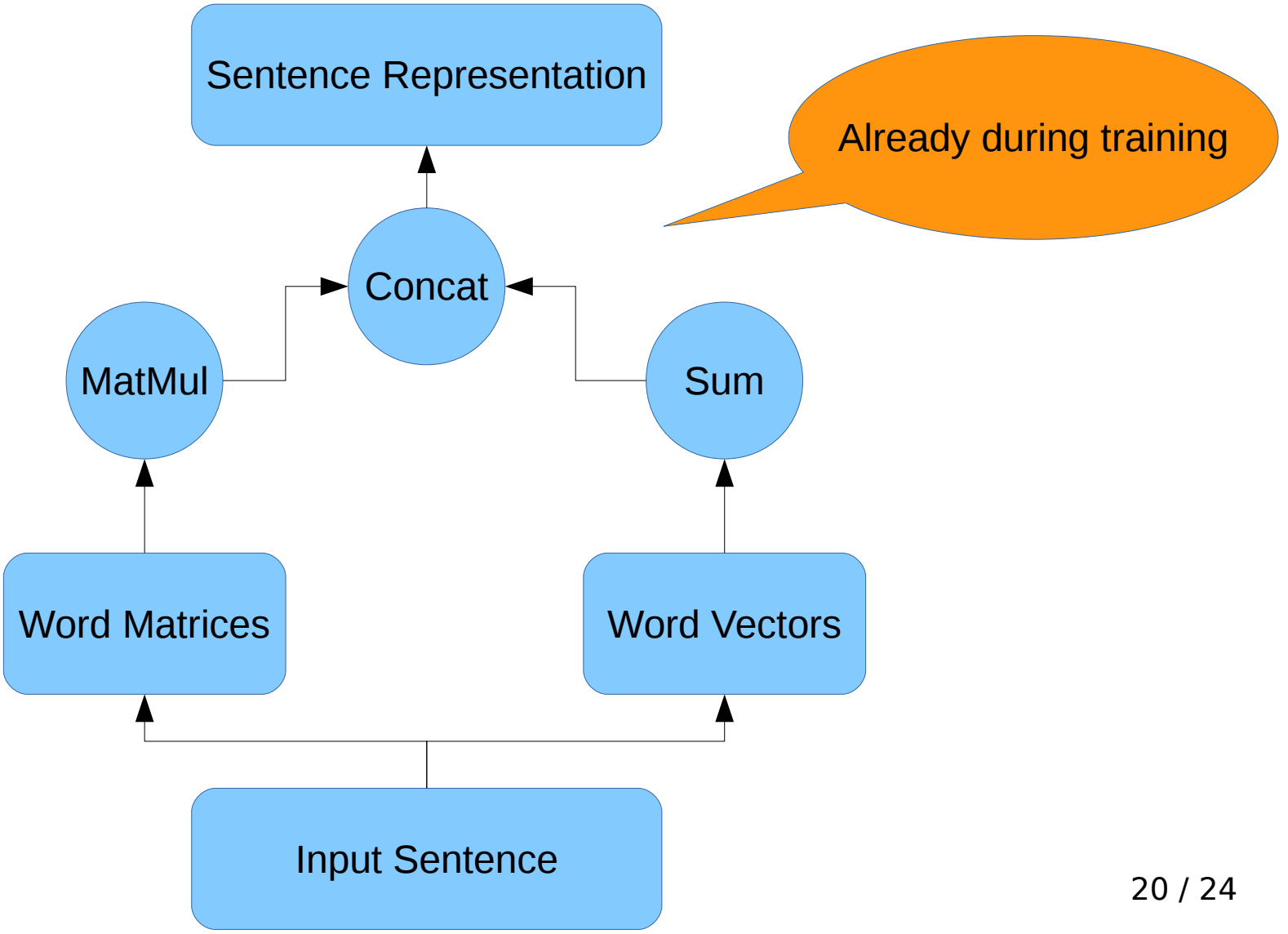
Vector-based and matrix-based models are complementary

- Vector-based models succeed at word content memorization
- Matrix-based models are superior in all other linguistic probing task
- **Idea:** use both vectors and matrices

Hybrid CMOW-CBOW Approach



Hybrid CMOW-CBOW Approach



Results: Hybrid Approach

Linguistic Probing Tasks

Method	Depth	BShift	SubjNum	Tense	Coord Inv	Length	Obj Num	TopConst	SOMO	WordContent
CBOW	33.0	49.6	79.3	78.4	53.6	74.5	78.6	72.0	49.6	**89.5**
CMOW	**35.1**	**70.8**	**82.0**	80.2	**61.8**	82.8	**79.7**	74.2	**50.7**	72.9
Hybrid	35.0	**70.8**	81.7	**81.0**	59.4	**84.4**	79.0	**74.3**	49.3	87.6

Supervised Downstream Tasks

Method	SUBJ	CR	MR	MPQA	MRPC	TREC	SICK-E	SST2	SST5	STS-B	SICK-R
CBOW	90.0	**79.2**	**74.0**	87.1	**71.6**	85.6	78.9	78.5	42.1	61.0	**78.1**
CMOW	87.5	73.4	70.6	**87.3**	69.6	**88.0**	77.2	74.7	37.9	56.5	76.2
Hybrid	**90.2**	78.7	73.7	**87.3**	**72.7**	87.6	**79.4**	**79.6**	**43.3**	**63.4**	77.8

1.2% avg. improvement

Unsupervised Downstream Tasks

Method	STS12	STS13	STS14	STS15	STS16
CBOW	43.5	**50.0**	**57.7**	**63.2**	**61.0**
CMOW	39.2	31.9	38.7	49.7	52.2
Hybrid	**49.6**	46.0	55.1	62.4	62.1

Discussion

- **Alternatives**
  - RNNs are weakly parallelizable, require O(n) sequential steps
  - BERT (Devlin et al. 2018) and other Transformers are quadratic in sequence length
  - 1D-Convolutions are similar, but wordorder might get lost during pooling
- **Our hybrid model**
  - Same #parameters as word embeddings
  - Requires only O(log n) sequential steps
  - Allows dynamic programming

Conclusion

- Matrix- and vector-based methods capture complementary properties
- Hybrid model achieves 1.2% improvement across 16 downstream tasks compared to CBOW
- First unsupervised training scheme for matrix-based embedding models
- Opens up new opportunities for order awareness

Acknowledgment

- This research was supported by the Swiss National Science Foundation under the project Learning Representations of Abstraction for Opinion Summarisation (LAOS), grant number “FNS-30216”.