Shortcut-Stacked Sentence Encoders for Multi-Domain Inference

Yixin Nie & Mohit Bansal

UNC
DEPARTMENT OF
COMPUTER SCIENCE

Task and Motivation

Premise	Label	Hypothesis	Genre
The Old One always comforted Ca'daan, except today.	*neutral*	Ca'daan knew the Old One very well.	*Fiction*
Your gift is appreciated by each and every student who will benefit from your generosity.	*neutral*	Hundreds of students will benefit from your generosity.	*Letters*
yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual or	*contradiction*	August is a black out month for vacations in the company.	*Telephone Speech*
At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	*entailment*	People formed a line at the end of Pennsylvania Avenue.	*9/11 Report*
A black race car starts up in front of a crowd of people.	*contradiction*	A man is driving down a lonely road.	*SNLI*

Only encoding-based models are eligible for the RepEval 2017 Shared Task.

[https://repeval2017.github.io/shared/], [https://nlp.stanford.edu/projects/snli/]

Motivation of Encoding-based Models

Encoding-based Model: models that transform sentences into **fixed-length** vector representations and reason using **only** those representations without cross-attention between two sentences

Motivation of Encoding-based Models

A **portable** neural model to transform the source sentence into some **sentence-level** meaning representation

- A plug and play module
- Sentence-level knowledge unit

Existing Encoding-based Model Results

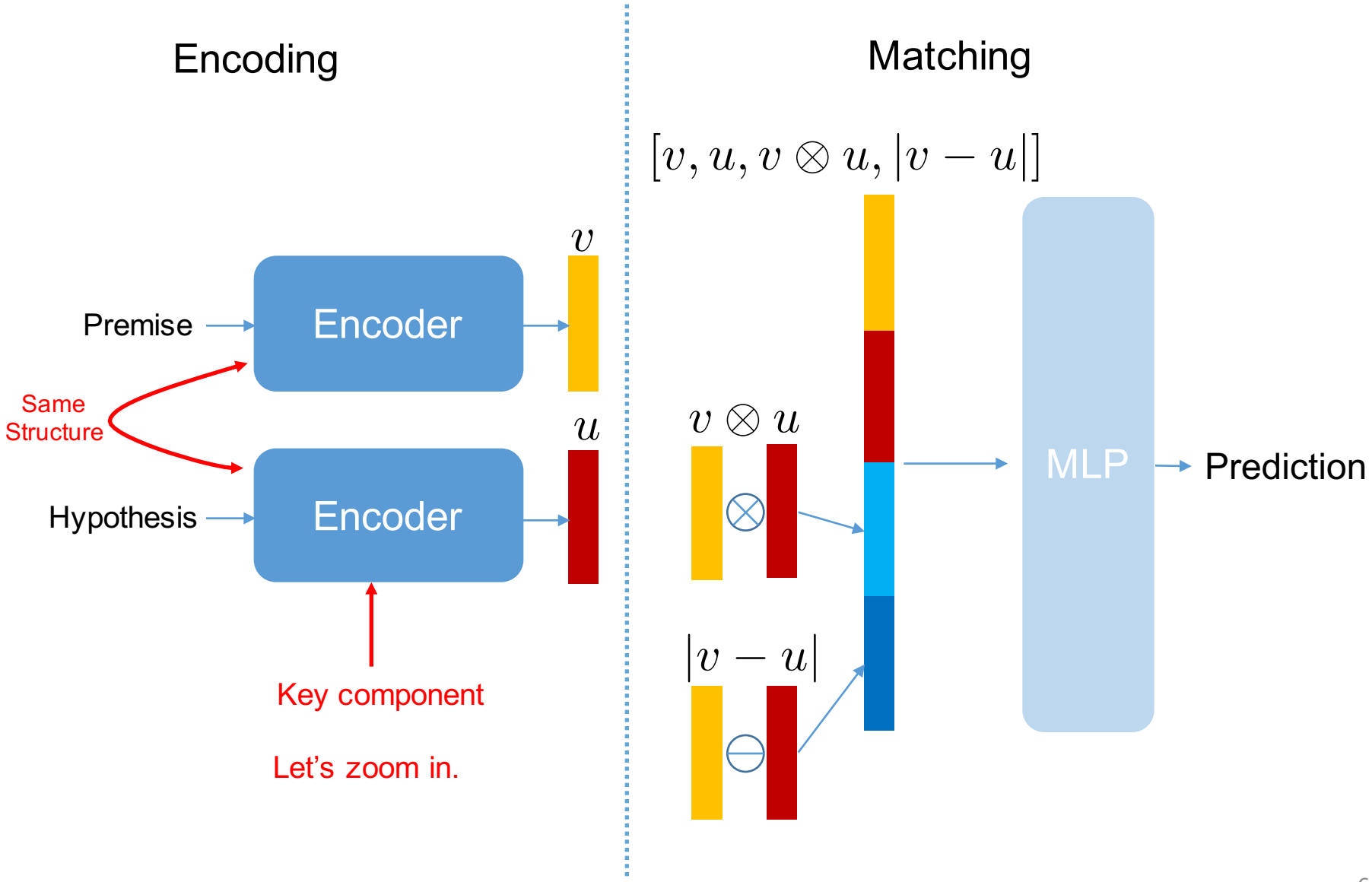
300D NSE encoders (Munkhdalai & Yu 2016)
84.6% on SNLI

BiLSTM Encoder (Williams et al., 2017)
67.5%/67.1% on MultiNLI (Matched/Mismatched)

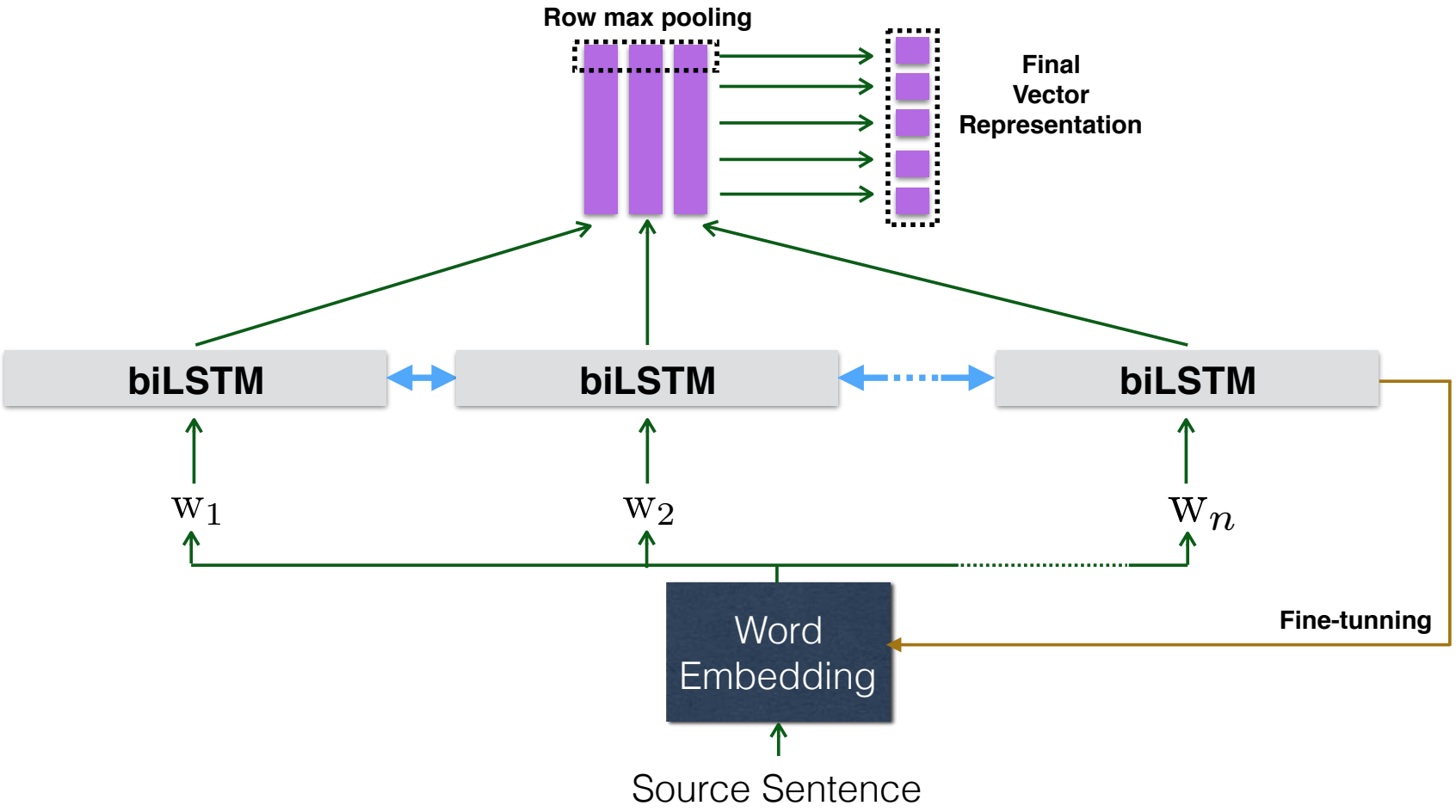
There is still much scope for improvement.

[https://repeval2017.github.io/shared/], [https://nlp.stanford.edu/projects/snli/]

Typical Architecture of Encoding-based Model



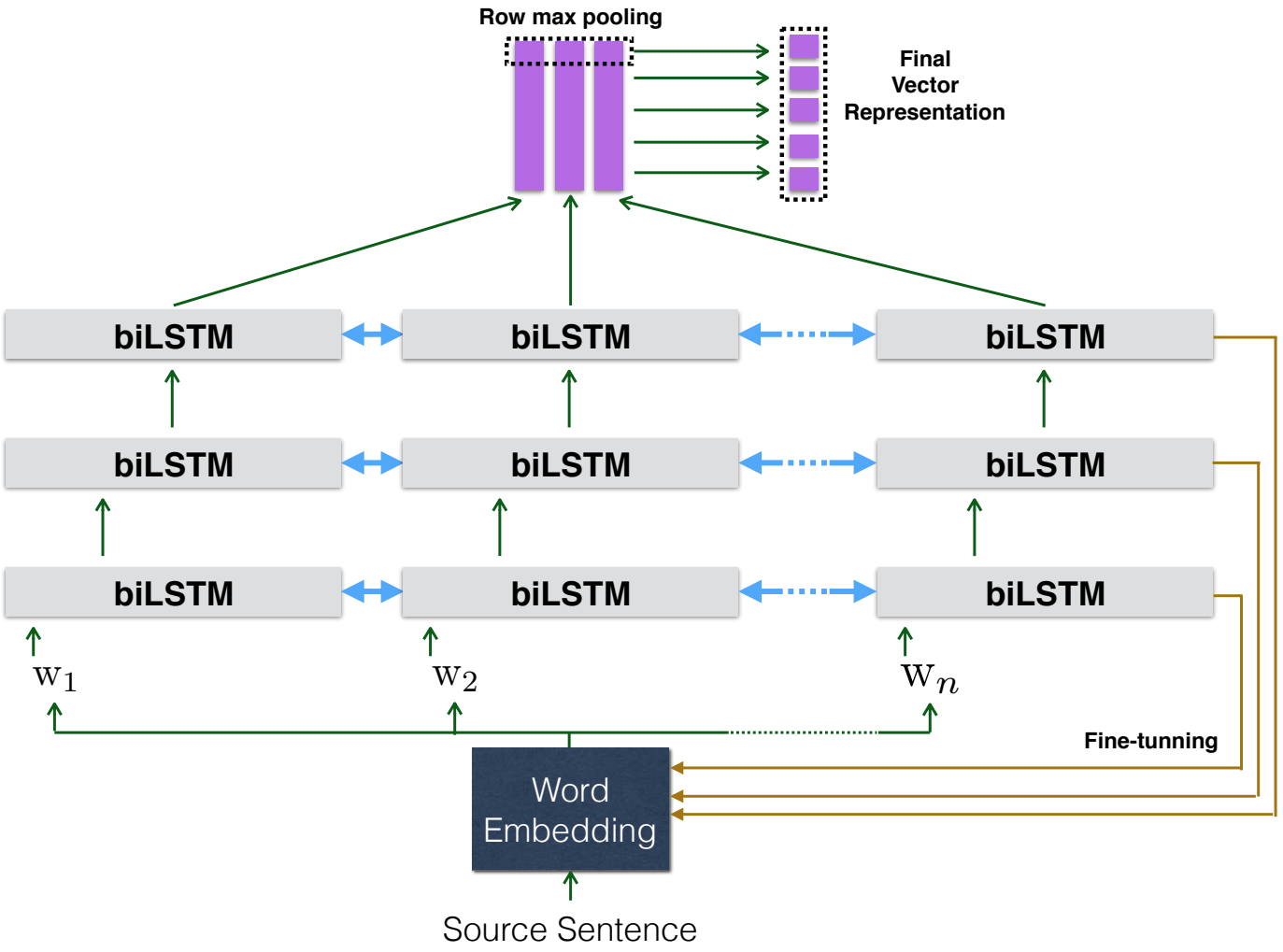
Encoder (Starting Point)



One Layer biLSTM with Max-pooling

[Conneauetal., 2016]

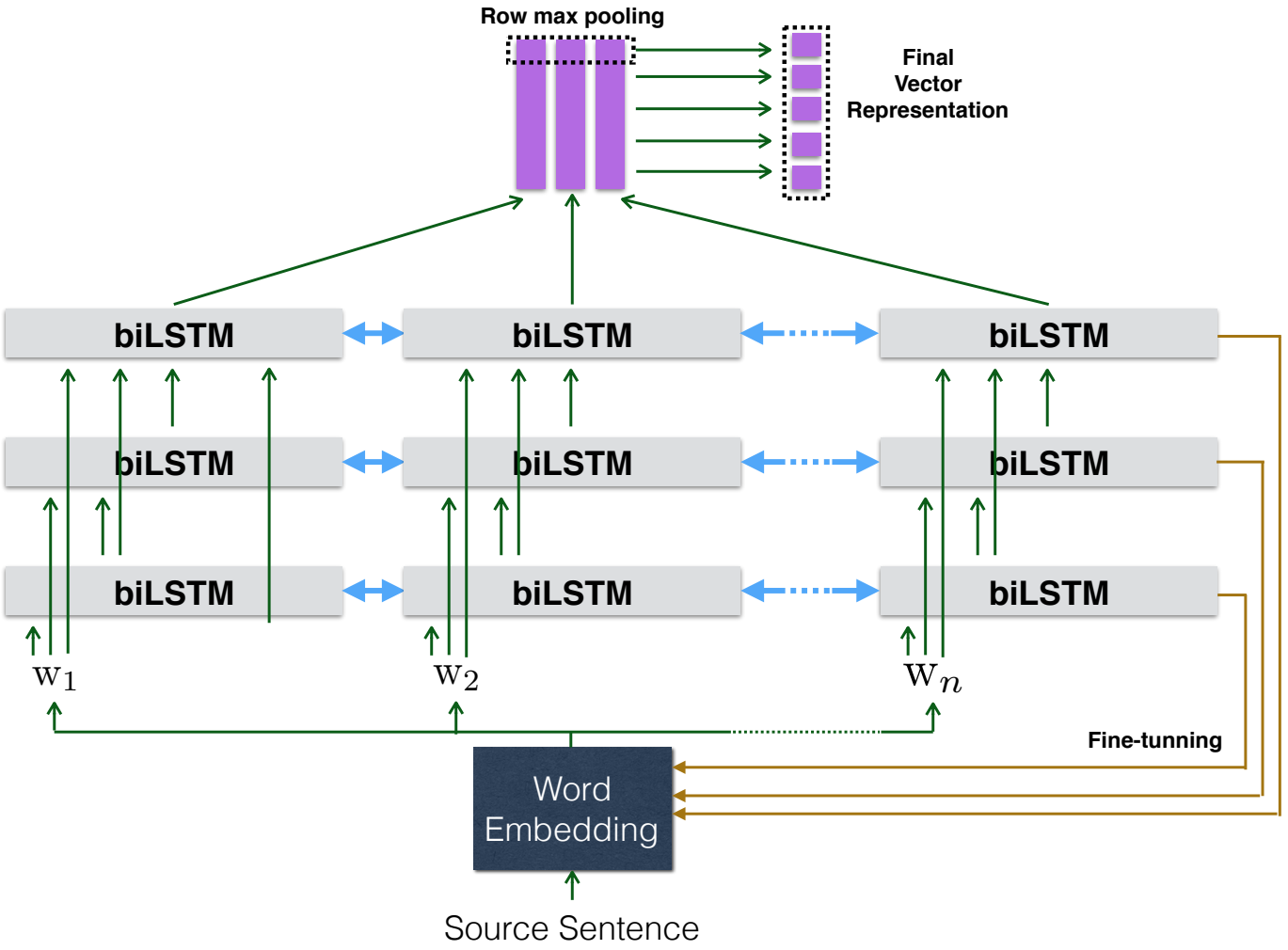
Encoder (Stacking bi-LSTM)



By stacking layers of biLSTM the model was able to learn some high-level semantic features that are useful for natural language inference task.

[Simonyan et al., 2016]

Encoder (Shortcut-connection)



Shortcut-connections help sparse gradient from max-pooling to flow into lower layers.

[Hashimoto et al., 2016]

Shared Task Competition Results

Team Name	Authors	Matched	Mismatched	Model Details
alpha (ensemble)	Chen et al.	**74.9%**	**74.9%**	STACK, CHAR, ATTN., POOL, PRODDIFF
YixinNie-UNC-NLP	Nie and Bansal	74.5%	73.5%	STACK, POOL, PRODDIFF, SNLI
alpha	Chen et al.	73.5%	73.6%	STACK, CHAR, ATTN, POOL, PRODDIFF
Rivercorners (ensemble)	Balazs et al.	72.2%	72.8%	ATTN, POOL, PRODDIFF, SNLI
Rivercorners	Balazs et al.	72.1%	72.1%	ATTN, POOL, PRODDIFF, SNLI
LCT-MALTA	Vu et al.	70.7%	70.8%	CHAR, ENHEMB, PRODDIFF, POOL
TALP-UPC	Yang et al.	67.9%	68.2%	CHAR, ATTN, SNLI
BiLSTM baseline	Williams et al.	67.0%	67.6%	POOL, PRODDIFF, SNLI

RepEval 2017 shared task competition results

[Nangia et al., 2017]

Ablation Analysis

Layers and Dimensions		Accuracy	
#layers	bilstm-dim	Matched	Mismatched
1	512	72.5	72.9
2	512 + 512	73.4	73.6
1	1024	72.9	72.9
2	512 + 1024	73.7	74.2
1	2048	73.0	73.5
2	512 + 2048	73.7	74.2
2	1024 + 2048	73.8	74.4
2	2048 + 2048	74.0	74.6
3	512 + 1024 + 2048	**74.2**	**74.7**

Results for models with different of biLSTM layers and their hidden state dimensions

Natural language inference tasks do require some high-level features that could be learned after applying multiple bi-RNN layers in sequence

Ablation Analysis

	Matched	Mismatched
without any shortcut connection	72.6	73.4
only word shortcut connection	74.2	74.6
full shortcut connection	**74.2**	**74.7**

Results with and without shortcut connections.

Main performance gain from shortcut property comes from shortcut-connection for word-embedding

Ablation Analysis

# of MLPs	Activation	Matched	Mismatched
1	tanh	73.7	74.1
2	tanh	73.5	73.6
1	relu	74.1	74.7
2	relu	**74.2**	**74.7**

Results for different MLP classifiers

Rectified linear unit is better than hyperbolic tangent function in this task

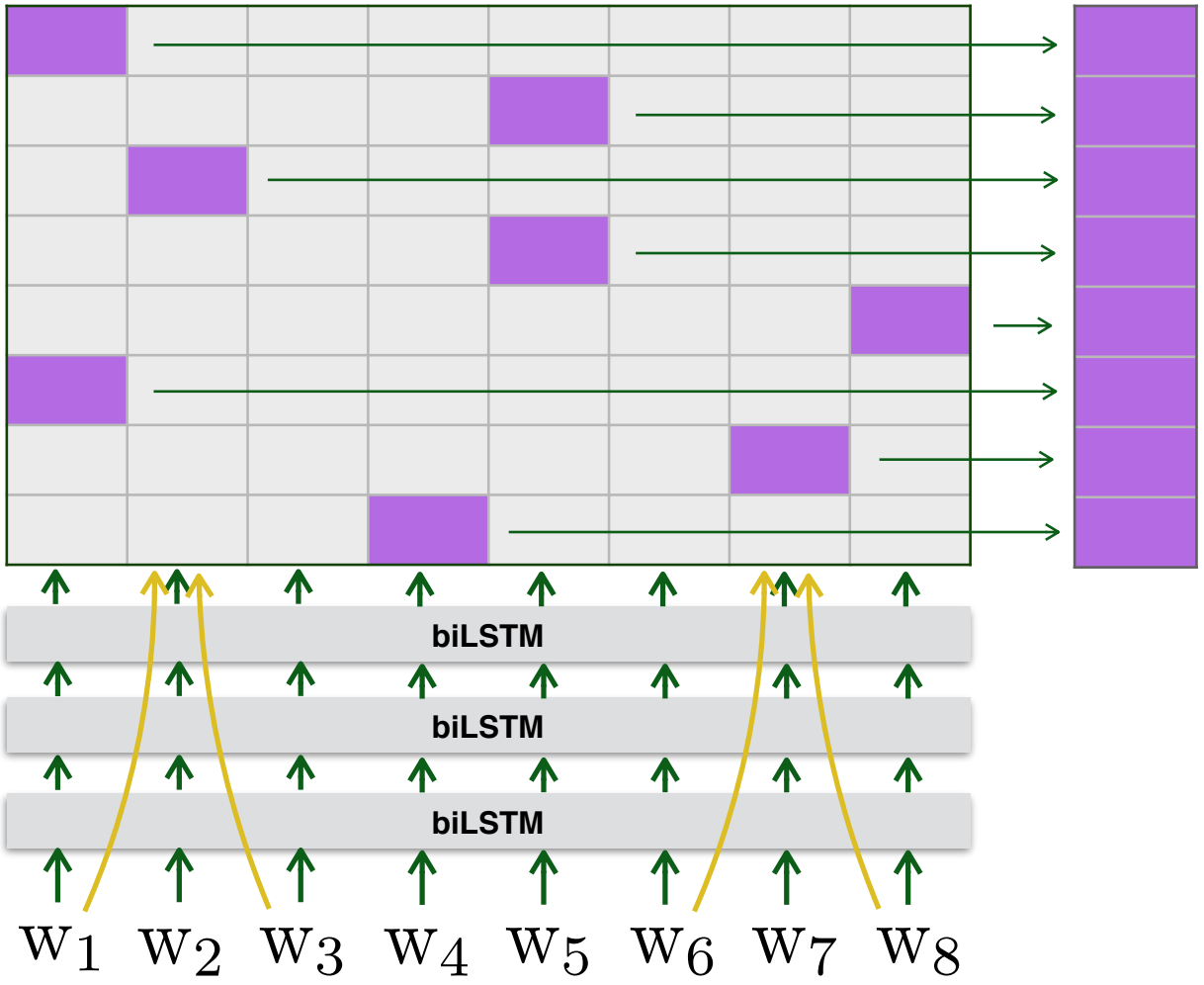
Results on SNLI and MultiNLI

Model	Accuracy		
	SNLI	Multi-NLI Matched	Multi-NLI Mismatched
CBOW (Williams et al., 2017)	80.6	65.2	64.6
biLSTM Encoder (Williams et al., 2017)	81.5	67.5	67.1
300D Tree-CNN Encoder (Mou et al., 2015)	82.1	–	–
300D SPINN-PI Encoder (Bowman et al., 2016)	83.2	–	–
300D NSE Encoder (Munkhdalai and Yu, 2016)	84.6	–	–
biLSTM-Max Encoder (Conneau et al., 2017)	84.5	–	–
Our biLSTM-Max Encoder	85.2	71.7	71.2
Our Shortcut-Stacked Encoder	**86.1**	**74.6**	**73.6**

Test Results on SNLI and Multi-NLI datasets

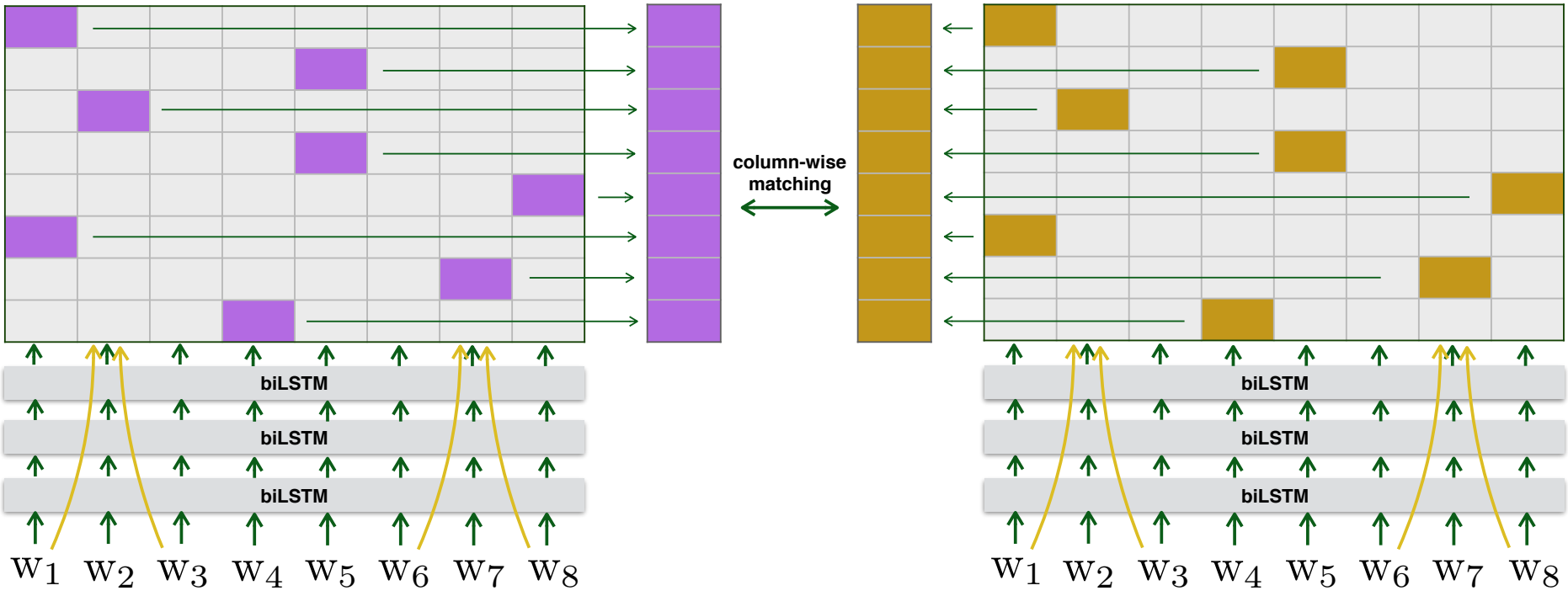
Our encoding-based model achieves new state-of-the-art on SNLI

Thoughts about Max-pooling



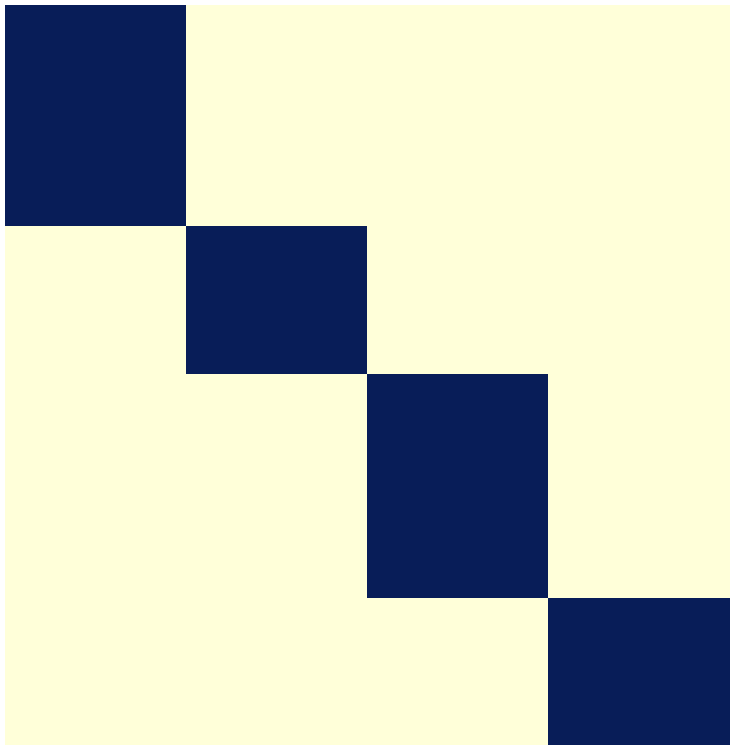
Each column in the final vector representation corresponds to **each word** in the source sentence and its **surroundings/context**

Thoughts about Max-pooling

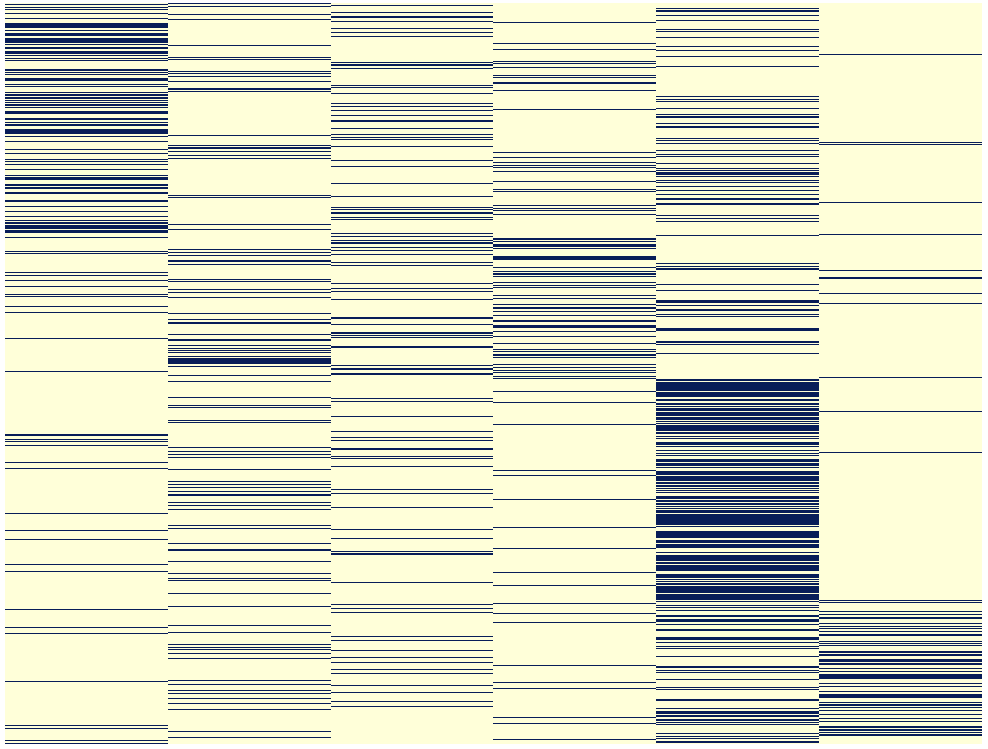


Column-wise matching between final vector representation of the two sentence corresponds to **word matching** between two sentence → similar to attention between two sentences

Thoughts about Max-pooling



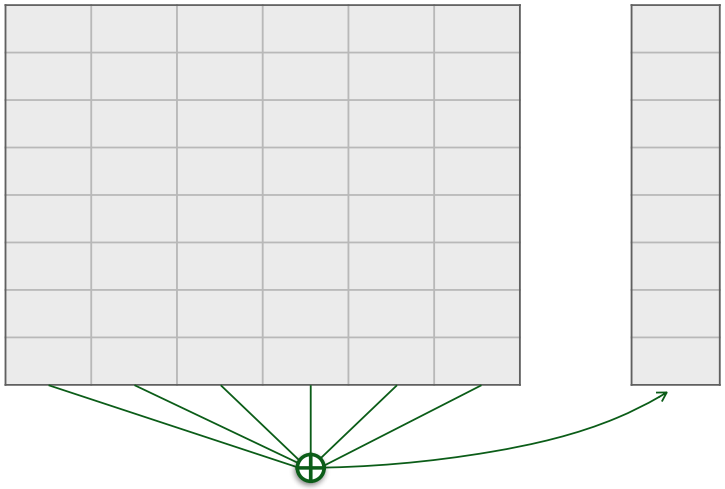
I like research .



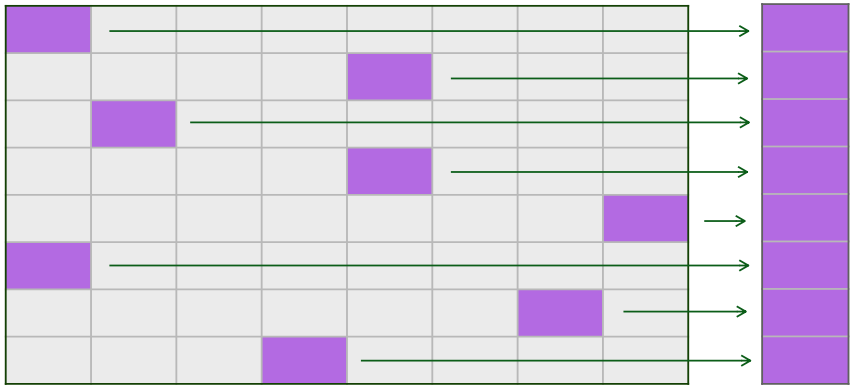
I do not like research .

Max-pooling vs. Attention

Soft-attention



Max-pooling



$$e_i = f(w_i, ...)$$
$$a = softmax(e)$$
$$v = \sum a_i h_i$$

Selectively combining information from each item of the source into a compact representation.

We are trying better/advanced max-pooling methods currently.

Vector Rep (1-NN Genre Accuracy)

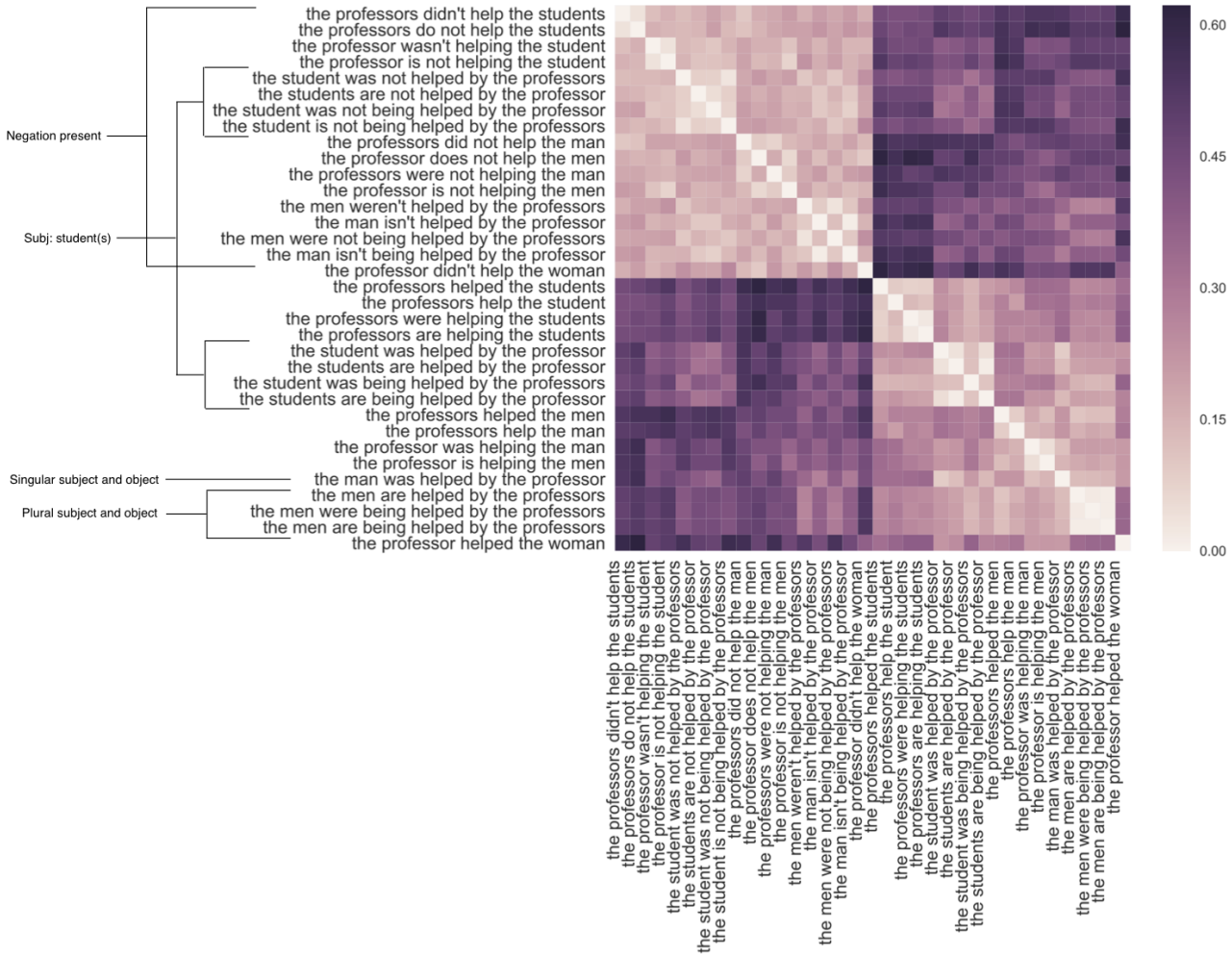
Authors	1-NN Genre Accuracy
Chen et al.	67.3%
Nie and Bansal	74.0%
Balazs et al.	69.2%
Vu et al.	67.0%
Yang et al.	54.7%

Table shows the percentage of times the first nearest neighbor belongs to the same genre as the sample sentence.

- Learned representations are not genre-agnostic
- Potential ability to handle genre classification task

[Nangia et al., 2017]

Vector Rep (Heatmap)



A heatmap showing the cosine similarity between sentence vectors.

Sentences tend to be more similar to one another if they have more structural features in common.

[Nangia et al., 2017]

Thanks

Yixin Nie
yixin1@cs.unc.edu
www.cs.unc.edu/~yixin1

Mohit Bansal
mbansal@cs.unc.edu
www.cs.unc.edu/~mbansal

UNC
DEPARTMENT OF
COMPUTER SCIENCE