Encoding Logic Rules in Sentiment Classification



Kalpesh Krishna*
UMass Amherst

Preethi Jyothi
IIT Bombay

Mohit Iyyer
UMass Amherst

** Work done at IIT Bombay*

Sentiment Classification

Sentiment Classification

Classify a sentence as **positive** or **negative**

Sentiment Classification

Classify a sentence as **positive** or **negative**

*this movie has a **great*** story

Sentiment Classification

Classify a sentence as **positive** or **negative**

this movie has a ***great*** story

Sentiment = **Positive**

Not Always Easy!

this movie has a ***great*** story

Solution :- Lexicons, Bag of Words

Not Always Easy!

Easy!

this movie has a ***great*** story

Solution :- Lexicons, Bag of Words

Not Always Easy!

Easy!

this movie has a ***great*** story

Solution :- Lexicons, Bag of Words

Contrastive - *this movie is funny, but* ***horribly directed***

Negation - *this is* ***not*** *a movie worth waiting for*

Not Always Easy!

Easy!

this movie has a ***great*** story

Solution :- Lexicons, Bag of Words

Much Harder!

Contrastive - *this movie is funny, but* ***horribly directed***

Negation - *this is* ***not*** *a movie worth waiting for*

Logic Rules

this movie is funny, *but* ***horribly directed***

A-*but*-***B***

Logic Rules

this movie is funny, *but* ***horribly directed***

A-*but*-***B***

sentiment(*A*-*but*-*B*) = sentiment(*B*)

Neural Nets + Logic Rules?

Method	Previous Work	Our Contributions

Neural Nets + Logic Rules?

Method	Previous Work	Our Contributions
Explicit		

Neural Nets + Logic Rules?

Method	Previous Work	Our Contributions
Explicit	Hu et al. (ACL 2016)	

Neural Nets + Logic Rules?

Method	Previous Work	Our Contributions
Explicit	Hu et al. (ACL 2016)	**Contribution #1** replication study finds **incorrect** conclusions

Neural Nets + Logic Rules?

Method	Previous Work	Our Contributions
Explicit	Hu et al. (ACL 2016)	**Contribution #1** replication study finds **incorrect** conclusions
Implicit?		

Neural Nets + Logic Rules?

Method	Previous Work	Our Contributions
Explicit	Hu et al. (ACL 2016)	**Contribution #1** replication study finds **incorrect** conclusions
Implicit?	Peters et al. (NAACL 2018)	

Neural Nets + Logic Rules?

Method	Previous Work	Our Contributions
Explicit	Hu et al. (ACL 2016)	**Contribution #1** replication study finds **incorrect** conclusions
Implicit?	Peters et al. (NAACL 2018)	**Contribution #2** ELMo embeddings learn logic rules **without** explicit supervision

Neural Nets + Logic Rules?

Method	Previous Work	Our Contributions
Explicit	Hu et al. (ACL 2016)	**Contribution #1** replication study finds **incorrect** conclusions
Implicit?	Peters et al. (NAACL 2018)	**Contribution #2** ELMo embeddings learn logic rules **without** explicit supervision

Contribution #3

Robustness of explicit / implicit methods to varying annotator agreement in A-but-B sentences

Digression: Reproducibility

Digression: Reproducibility

Small benchmark datasets (SST, MR, CR)

Significant variation in performance every run
(due to random initialization / GPU parallelization)

Digression: Reproducibility

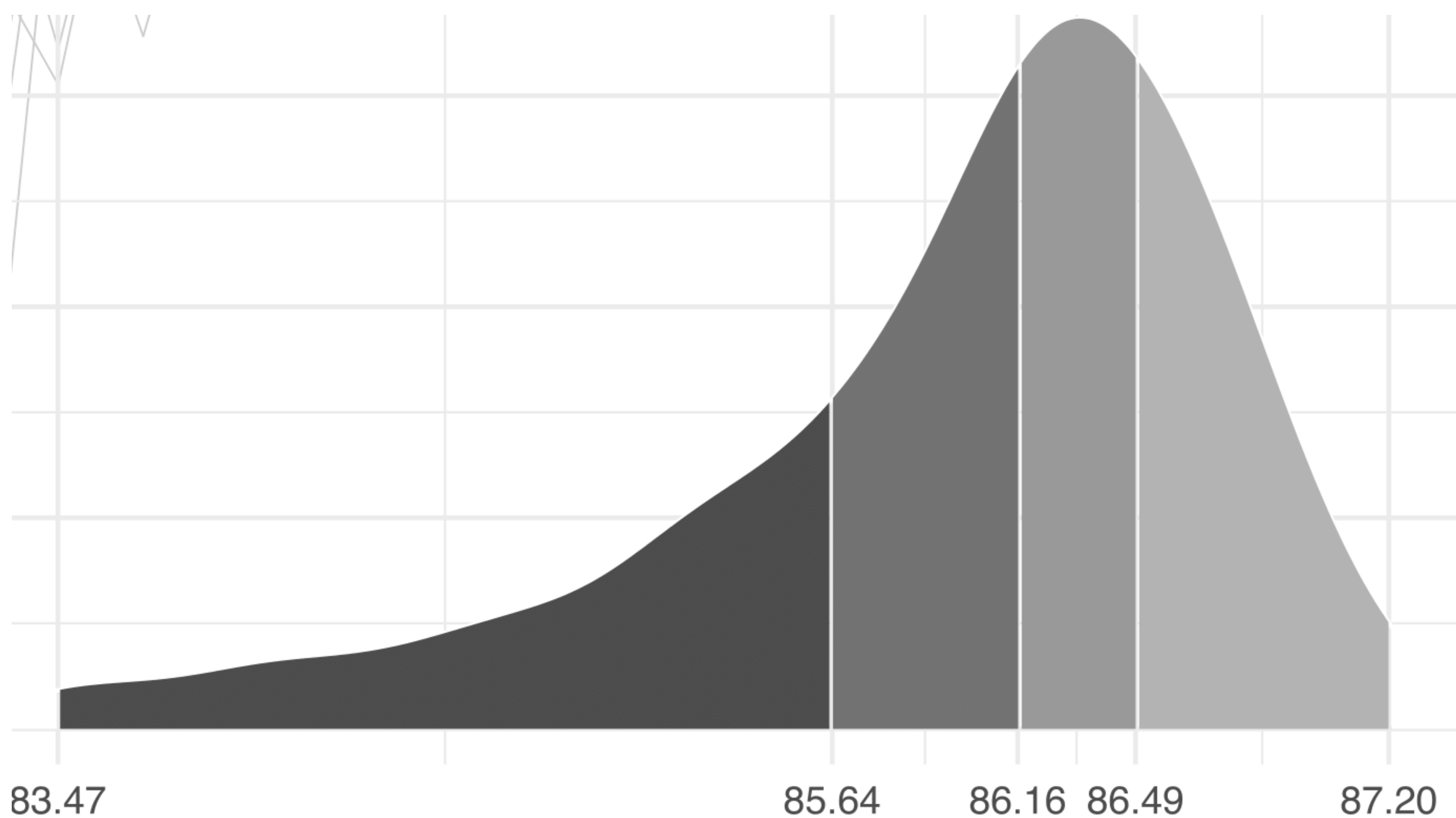
Small benchmark datasets (SST, MR, CR)

Significant variation in performance every run (due to random initialization / GPU parallelization)

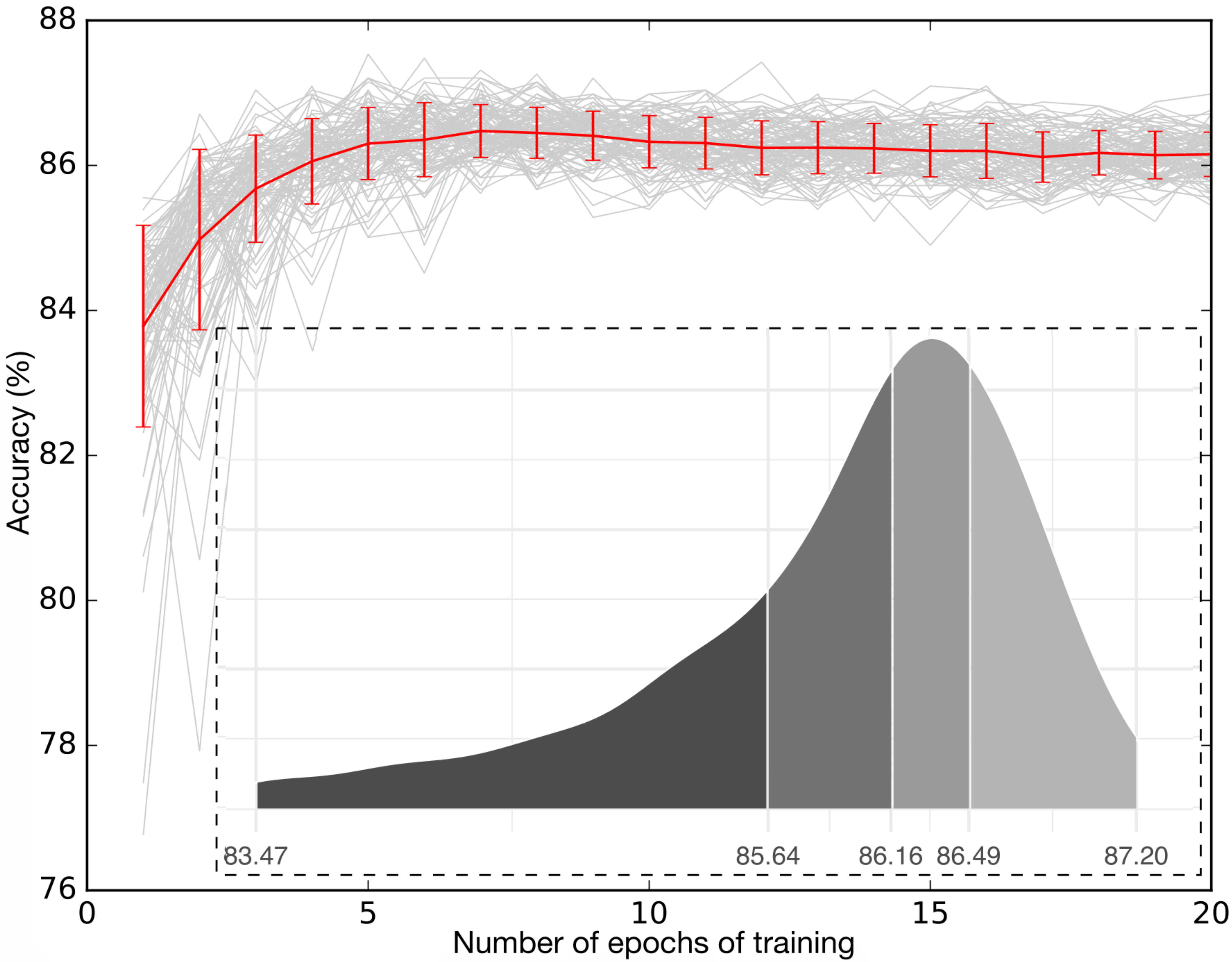
Solution :- Average performance over a large number of random seeds (Reimers and Gurevych 2017)

Large Variation (100 seeds)

Large Variation (100 seeds)



Large Variation (100 seeds)



Outline

Method	Previous Work	Our Contributions
Explicit	Hu et al. (ACL 2016)	**Contribution #1** replication study finds **incorrect** conclusions
Implicit?	Peters et al. (NAACL 2018)	**Contribution #2** ELMo embeddings learn logic rules **without** explicit supervision

Contribution #3

Robustness of explicit / implicit methods to varying annotator agreement in A-but-B sentences

Model in Hu et al. 2016

Expectation-**M**aximization style algorithm

E: Projection (Ganchev et al. 2010)

M: Distillation (Hinton et al. 2014)

E: **Projection** (Ganchev et al. 2010)

E: Projection (Ganchev et al. 2010)

*this movie is funny, but **horribly directed***

E: Projection (Ganchev et al. 2010)

this movie is funny, but ***horribly directed***

$$p_\theta(y|x)$$

negative = 0.34

positive = 0.66

E: Projection (Ganchev et al. 2010)

this movie is funny, but ***horribly directed***

$$p_\theta(y|x)$$

project

negative = 0.34

positive = 0.66

E: Projection (Ganchev et al. 2010)

this movie is funny, but ***horribly directed***

$$p_\theta(y|x) \xrightarrow{\text{project}} q_\theta(y|x)$$

$p_\theta(y\|x)$	$q_\theta(y\|x)$
negative = 0.34	**negative = 0.77**
positive = 0.66	positive = 0.23

E: **Projection** (Ganchev et al. 2010)

*this movie is funny, <u>but</u> **horribly directed***

$$p_\theta(y|x) \xrightarrow{\text{project}} q_\theta(y|x)$$

negative = 0.34
positive = 0.66

negative = 0.77
positive = 0.23

projection is a **convex** optimization problem

E: **Projection** (Ganchev et al. 2010)

this movie is funny, <u>but</u> ***horribly directed***

$$p_\theta(y|x) \xrightarrow{\text{project}} q_\theta(y|x)$$

negative = 0.34
positive = 0.66

negative = 0.77
positive = 0.23

projection is a **convex** optimization problem

new distribution consistent with **logic rules**

M: Distillation (Hinton et al. 2014)

$$L = \lambda H(p_{\mathrm{truth}}, p_\theta) + \boxed{(1-\lambda) H(q_\theta, p_\theta)}$$

train model with projected distribution as **soft-label**

Hu et al. 2016 algorithm

E: Projection

M: Distillation

```
forall minibatch (x,y) {
  p = forward(x)
  q = project(p)
  theta += grad-update(p, q, y)
}
```

Conclusions in Hu et al. 2016

1) Distilled model **better** than single projection

2) Distilled neural network has **significant gain** on SST2 as it **learns A-but-B rule**

Our Conclusions

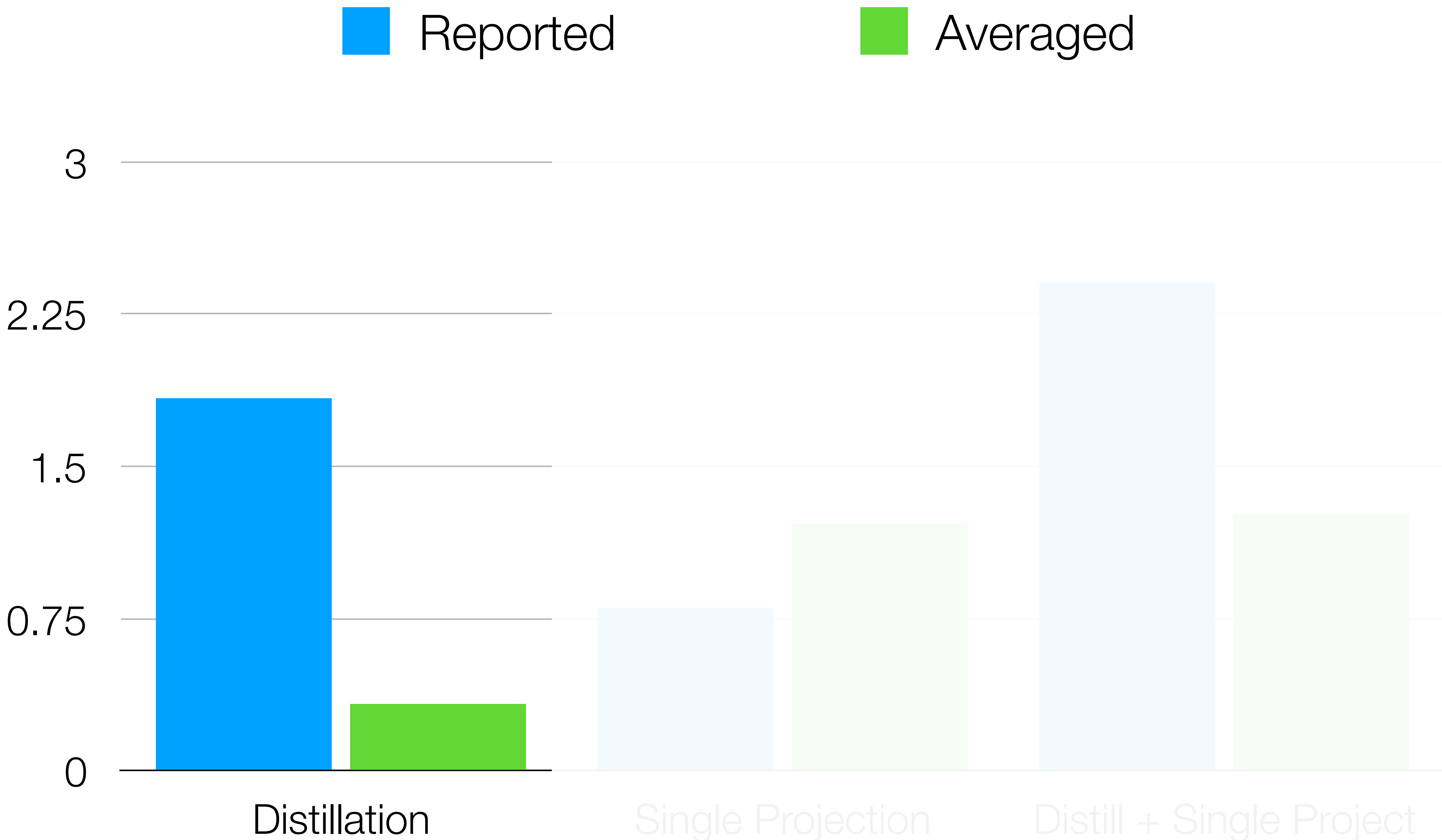
~~1) Distilled model **better** than single projection~~

~~2) Distilled neural network has **significant gain** on SST2 as it **learns A-but-B rule**~~

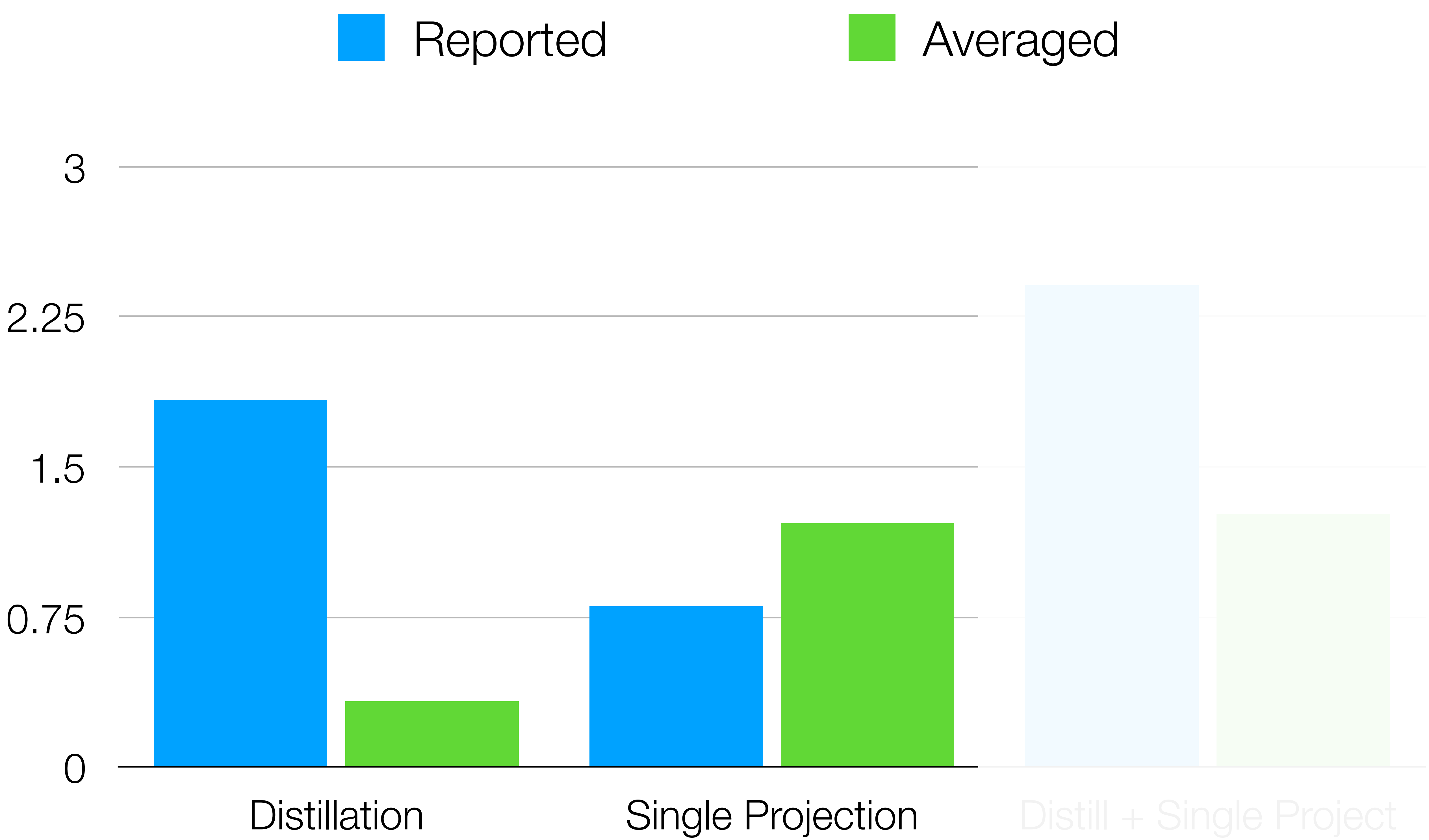
1) A **single projection** is a good way to explicitly encode logic rules

2) Distilled neural nets **aren't learning logic rules**

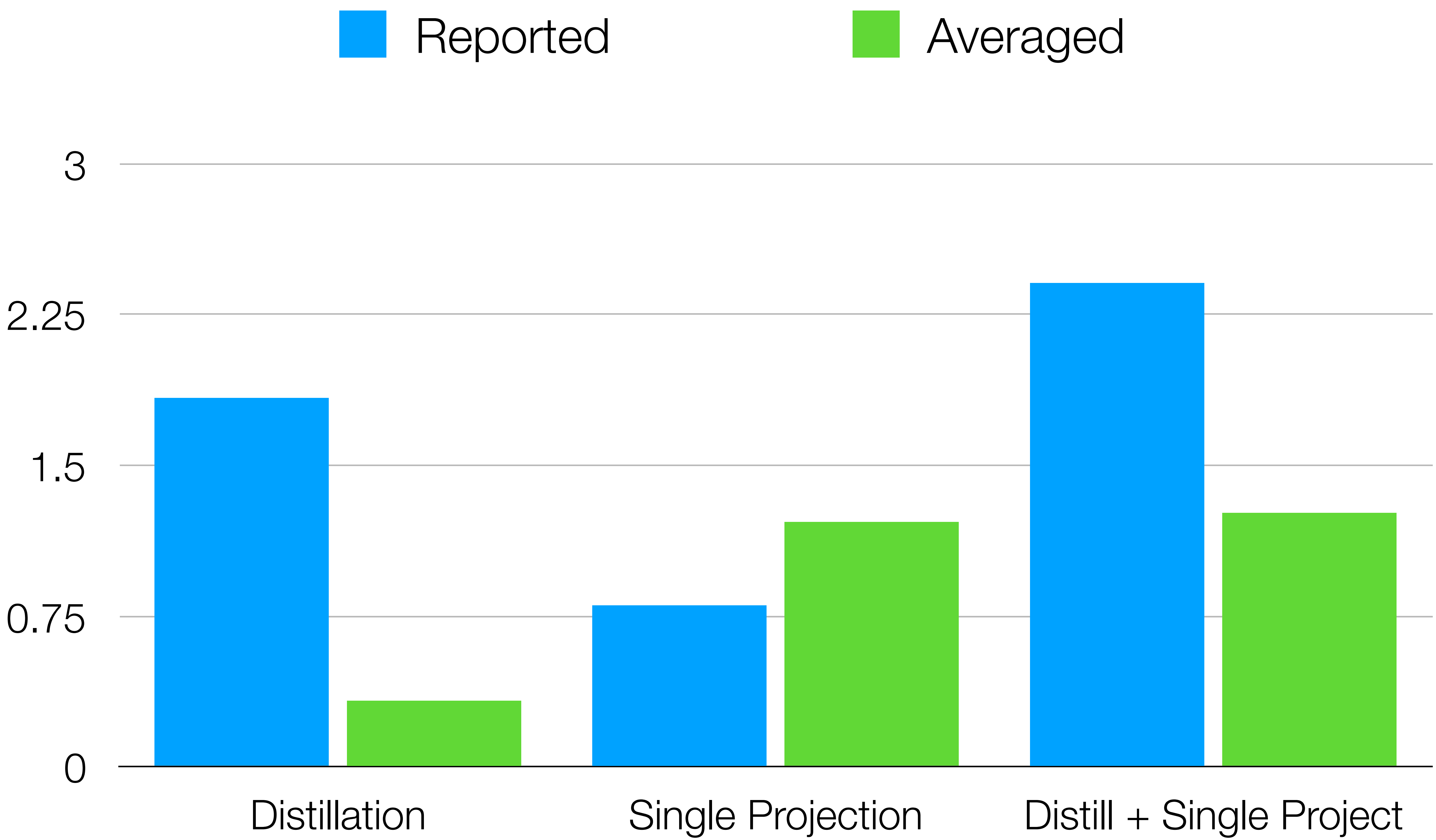
Distillation is ineffective



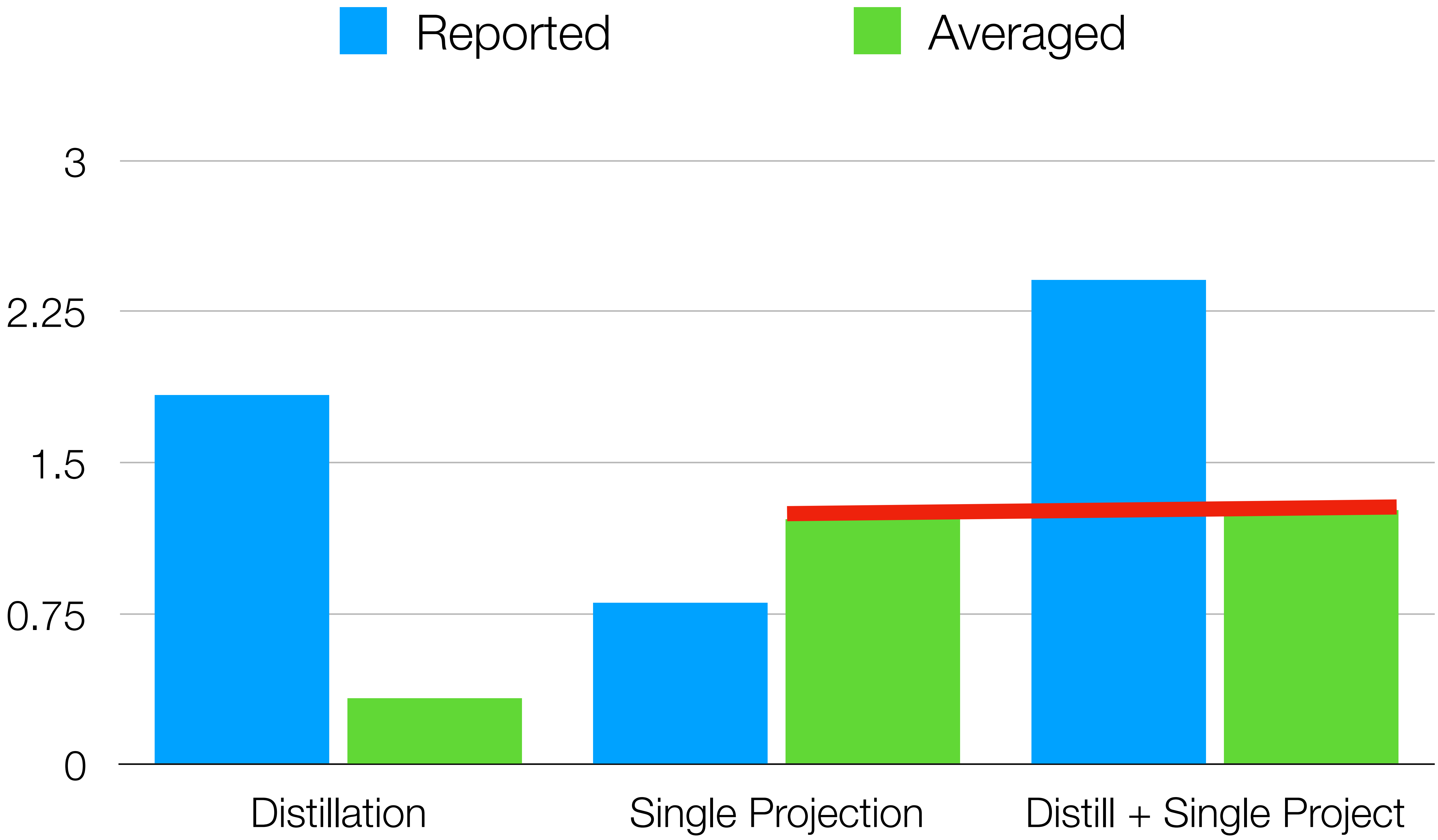
Single Projection good!



Single Projection sufficient!



Single Projection sufficient!



Consistent Trend on A-but-B

Reported	N / A
Averaged Gain %	Distillation = 1.9% **Single Projection = 9.3%** Distillation + Projection = 8.9%

Again, a **single** projection at test time is sufficient!

Our Conclusions

~~1) Distilled model **better** than single projection~~

~~2) Distilled neural network has **significant gain** on SST2 as it **learns A-but-B rule**~~

1) A **single projection** is a good way to explicitly encode logic rules

2) Distilled neural nets **aren't learning logic rules**

Outline

Method	Previous Work	Our Contributions
Explicit	Hu et al. (ACL 2016)	**Contribution #1** replication study finds **incorrect** conclusions
Implicit?	Peters et al. (NAACL 2018)	**Contribution #2** ELMo embeddings learn logic rules **without** explicit supervision

Contribution #3

Robustness of explicit / implicit methods to varying annotator agreement in A-but-B sentences

ELMo Representations

Embeddings from **L**anguage **Mo**dels

ELMo Representations

Embeddings from **L**anguage **Mo**dels

large language model trained on the 1 Billion Words dataset

ELMo Representations

Embeddings from **L**anguage **Mo**dels

large language model trained on the 1 Billion Words dataset

learnt representations used for downstream task

ELMo Representations

Embeddings from **L**anguage **Mo**dels

large language model trained on the 1 Billion Words dataset

learnt representations used for downstream task

Unlike word2vec, these embeddings are **contextual**

ELMo Results (100 seeds)

ELMo Results (100 seeds)

Model	SST2	A-but-B	A-but-B + negation
CNN (Baseline)	86.0	78.7	80.1
CNN + ELMo	88.9	86.5	87.2
Gain %	**2.9**	**7.8**	**7.1**

ELMo Results (100 seeds)

Model	SST2	A-but-B	A-but-B + negation
CNN (Baseline)	86.0	78.7	80.1
CNN + ELMo	88.9	86.5	87.2
Gain %	**2.9**	**7.8**	**7.1**

Significant improvement, *even after* averaging!

Is ELMo Learning Logic Rules?

Model	SST2	A-but-B	A-but-B + negation
CNN (Baseline)	86.0	78.7	80.1
CNN + ELMo	88.9	86.5	87.2
Gain %	**2.9**	**7.8**	**7.1**

60% of the improvement is on A-but-B sentences and negations

Is ELMo Learning Logic Rules?

Model	SST2	A-but-B	A-but-B + negation
CNN (Baseline)	86.0	78.7	80.1
CNN + ELMo	88.9	86.5	87.2
Gain %	**2.9**	**7.8**	**7.1**

60% of the improvement is on
A-but-B sentences and negations

*(Only **24.5%** of corpus is A-but-B / negations)*

ELMo + Explicit (Projection)

Model	SST2	A-but-B
ELMo	88.9	86.5
ELMo + project	89.0	87.2
Gain %	0.1	0.7

ELMo + Explicit (Projection)

Model	SST2	A-but-B
ELMo	88.9	86.5
ELMo + project	89.0	87.2
Gain %	0.1	0.7

Test-time projection is **ineffective** for ELMo

ELMo + Explicit (Projection)

Model	SST2	A-but-B
ELMo	88.9	86.5
ELMo + project	89.0	87.2
Gain %	0.1	0.7

Test-time projection is **ineffective** for ELMo

Distance between ELMo distribution and projected distribution is **0.13** (vs **0.26** distillation, **0.27** baseline)

Clustering ELMo Vectors

Cosine similarity between every pair of words

Clustering ELMo Vectors

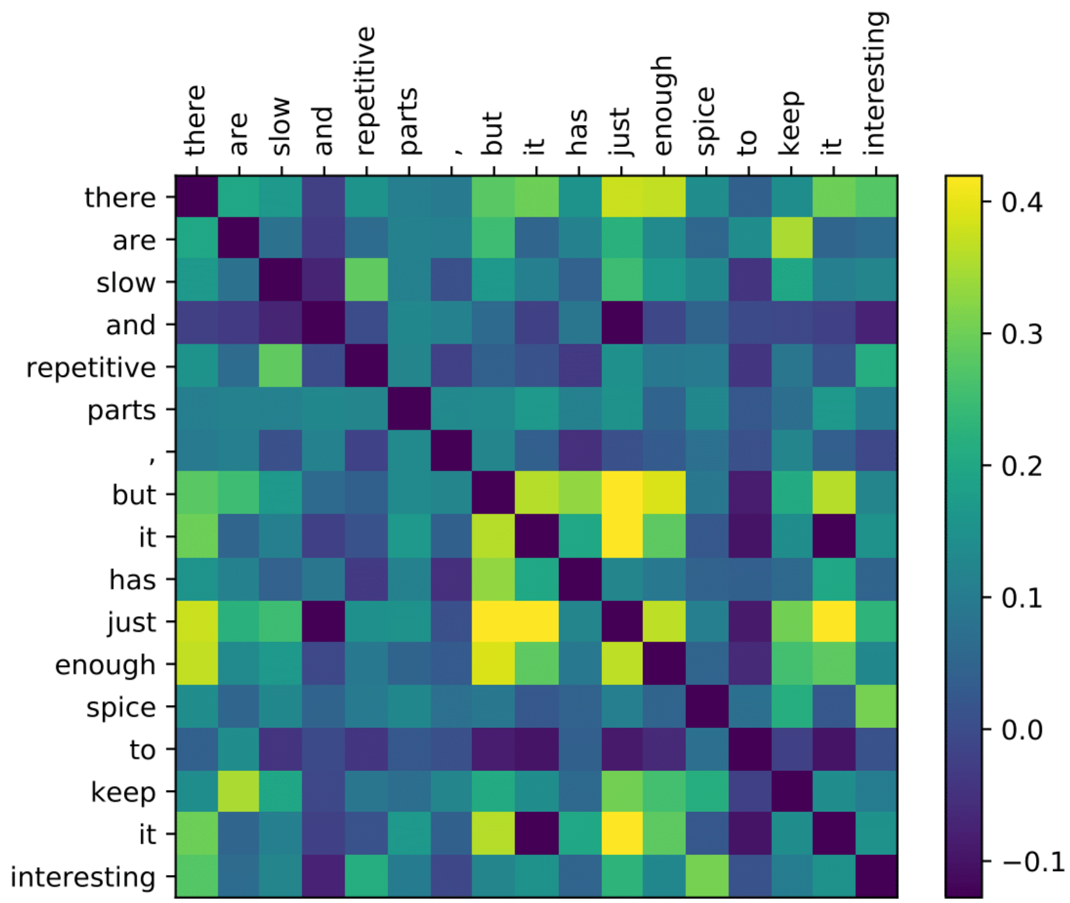
Cosine similarity between every pair of words

Contrastive (A-<u>but</u>-B)

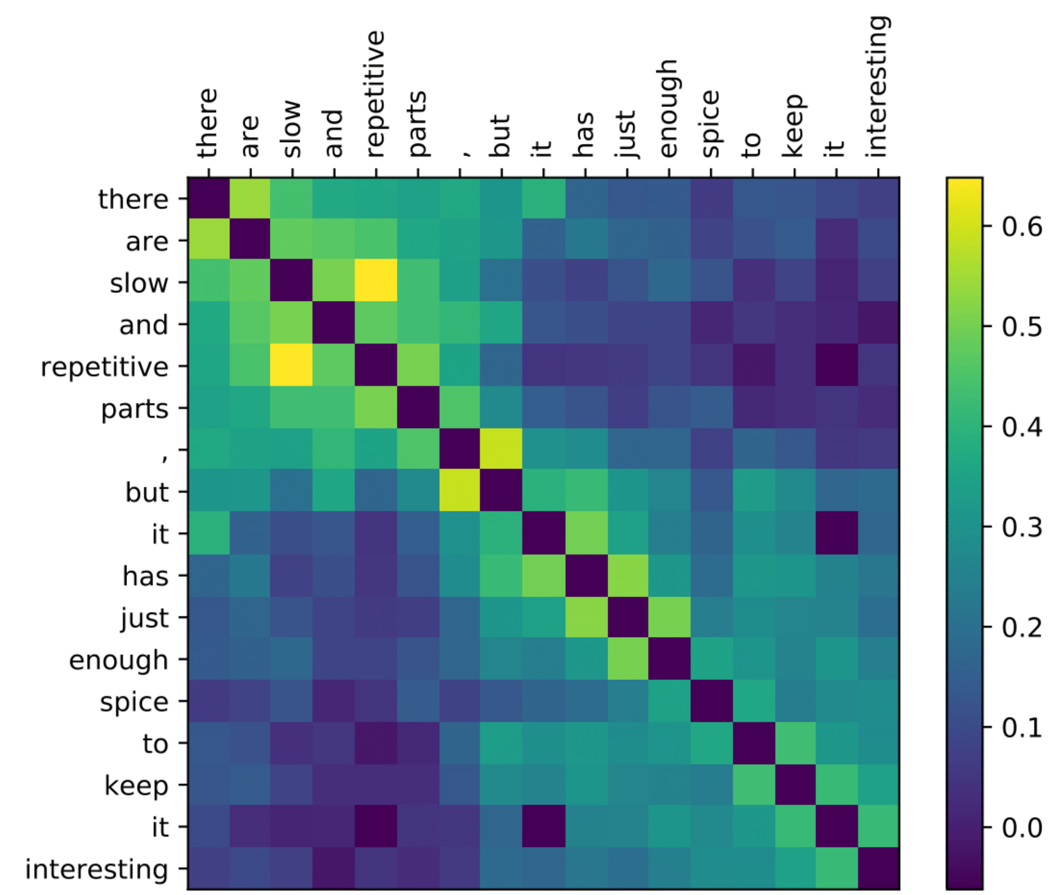
there are slow and repetitive parts, <u>but</u> it has just enough spice to keep it interesting

there are slow and repetitive parts, <u>but</u> it has just enough spice to keep it interesting

there are slow and repetitive parts, but it has just enough spice to keep it interesting

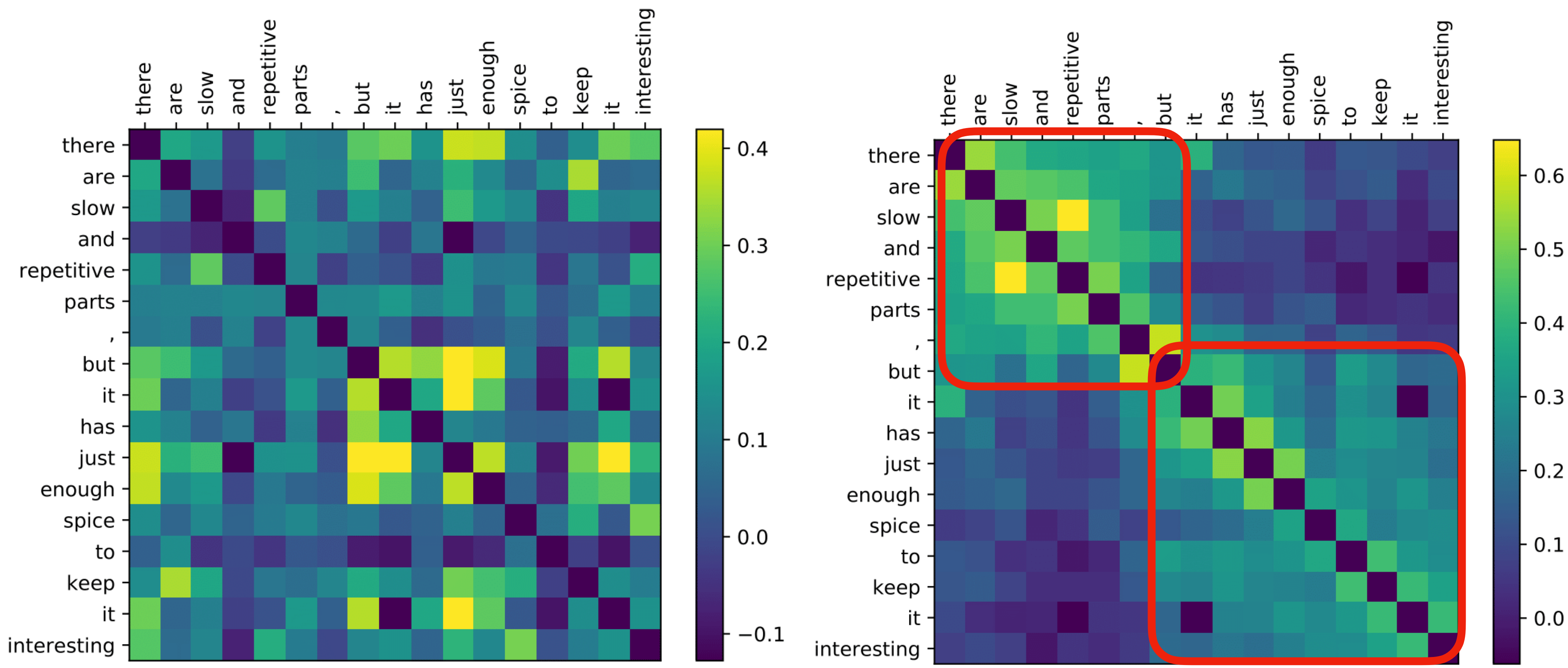


word2vec



ELMo

there are slow and repetitive parts, but it has just enough spice to keep it interesting

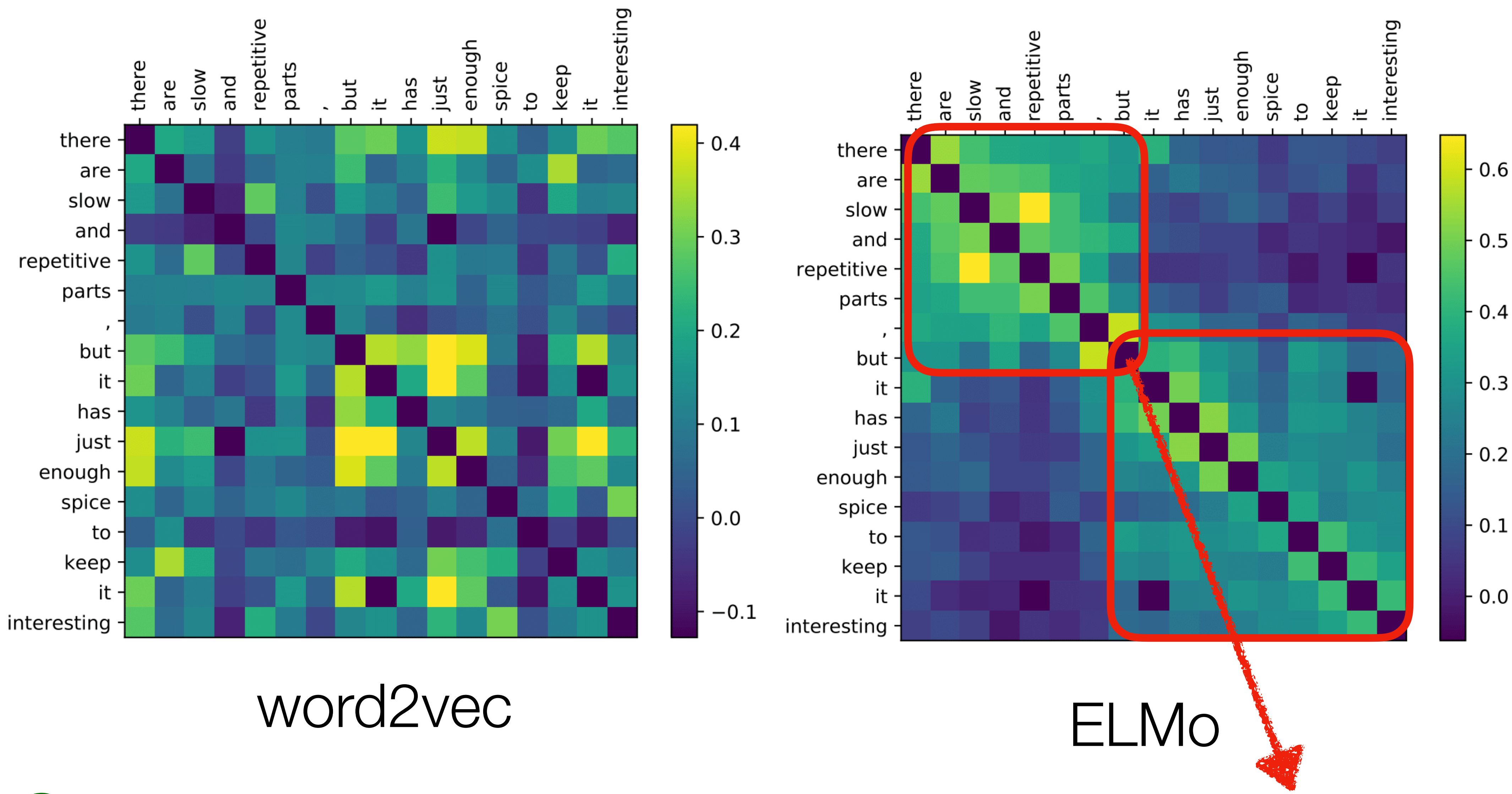


word2vec

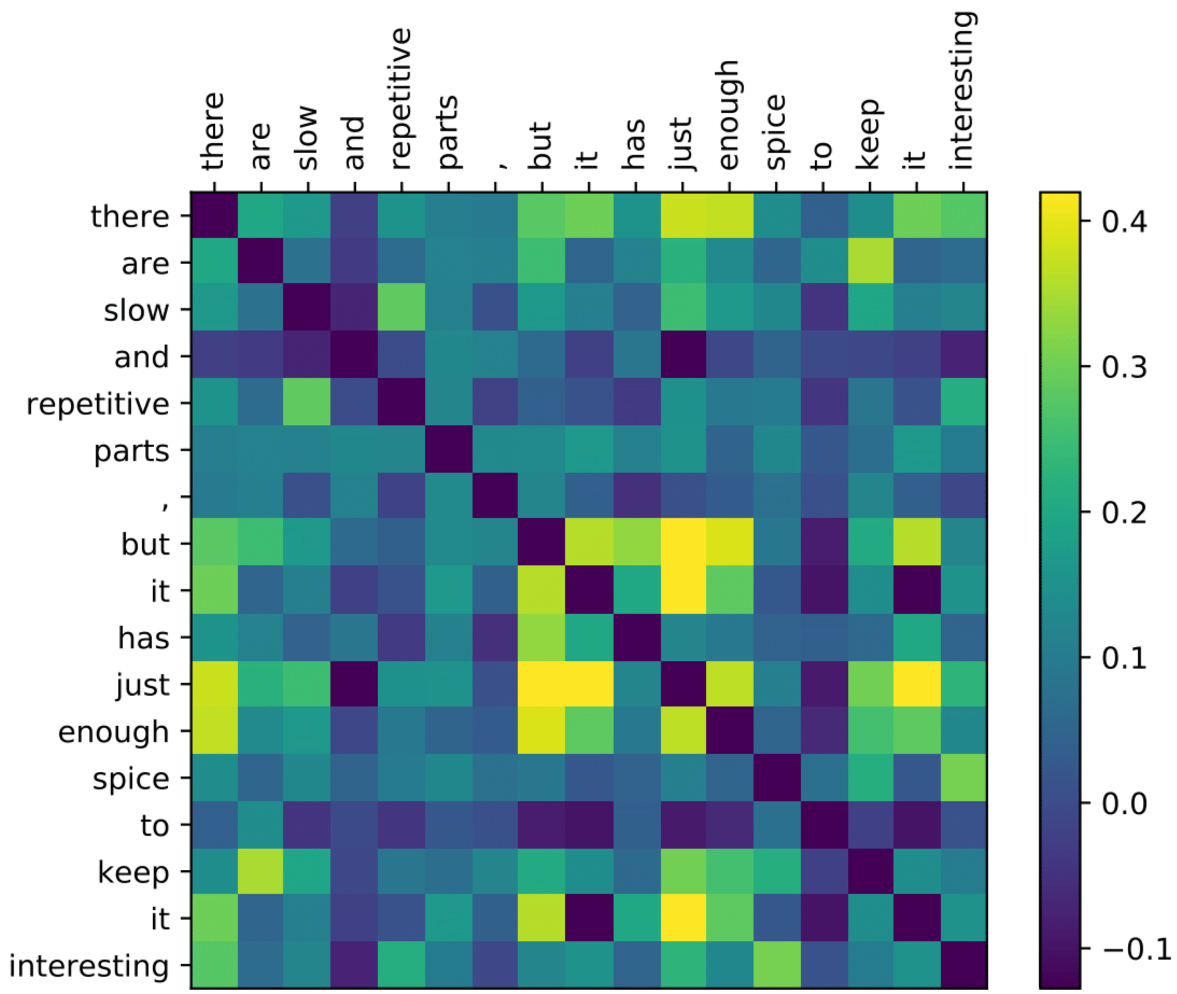
ELMo

Clustering for A part and B part in A-*but*-B sentences for ELMo embeddings

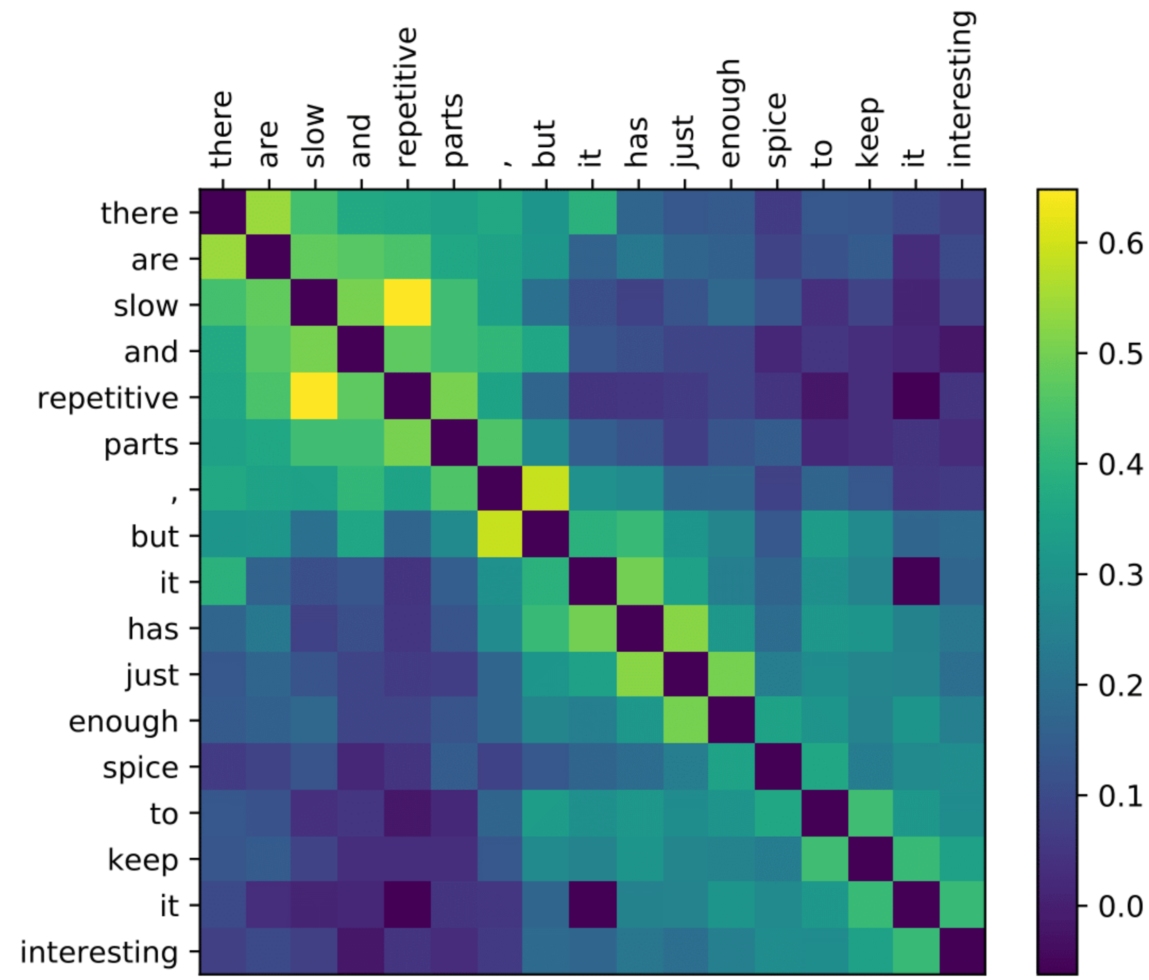
there are slow and repetitive parts, <u>but</u> it has just enough spice to keep it interesting



Clustering for A part and B part in A-*but*-B sentences for ELMo embeddings



word2vec



ELMo

ELMo Representations learn the scope of a contrastive conjunction!

Outline

Method	Previous Work	Our Contributions
Explicit	Hu et al. (ACL 2016)	**Contribution #1** replication study finds **incorrect** conclusions
Implicit?	Peters et al. (NAACL 2018)	**Contribution #2** ELMo embeddings learn logic rules **without** explicit supervision

Contribution #3
Robustness of explicit / implicit methods to varying annotator agreement in A-but-B sentences

Sentiment is Ambiguous!

beautiful film, but those who have read the book will be disappointed

Sentiment is Ambiguous!

beautiful film, but those who have read the book will be disappointed

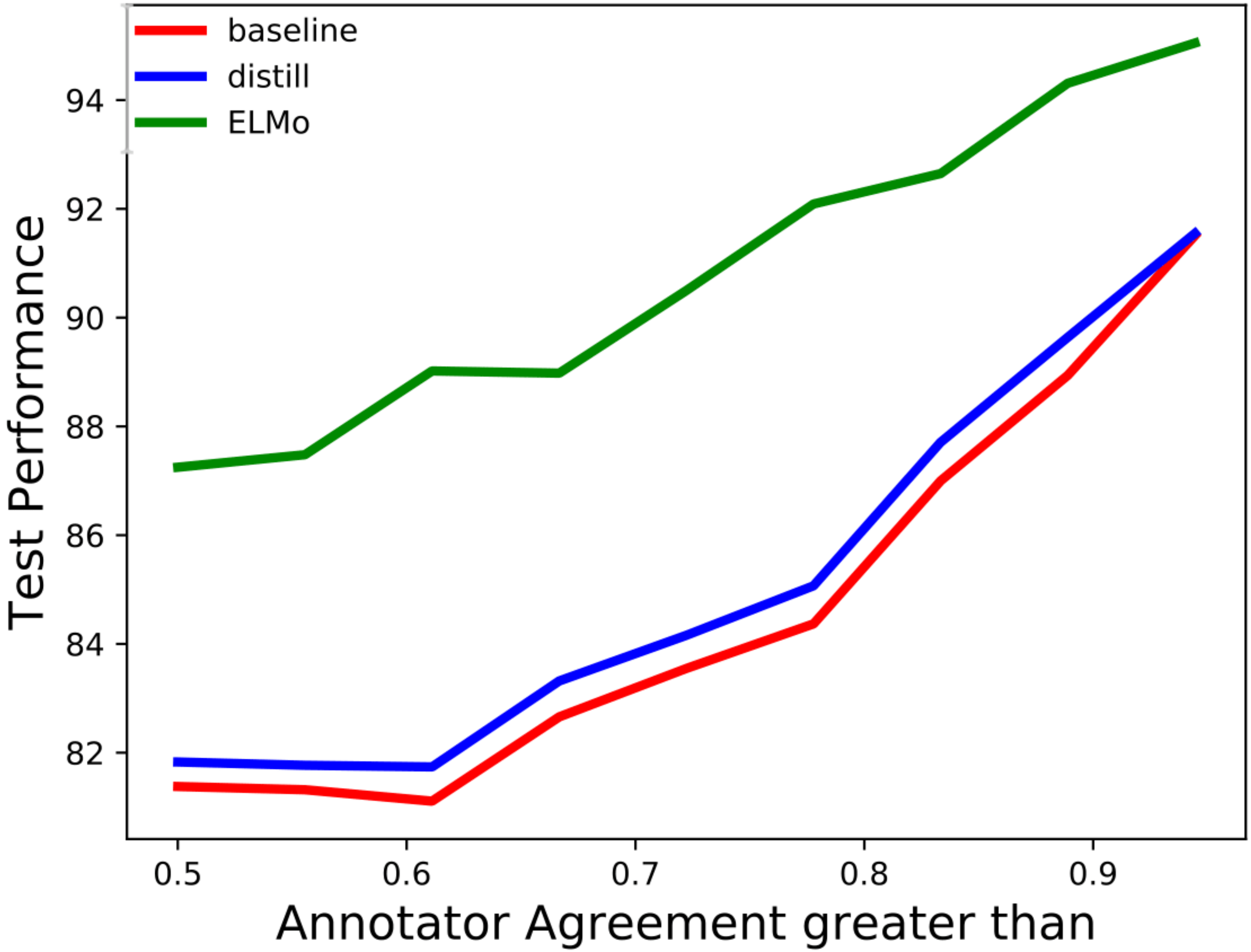
nine crowd-workers label each A-but-B sentence as positive / negative / neutral

Sentiment is Ambiguous!

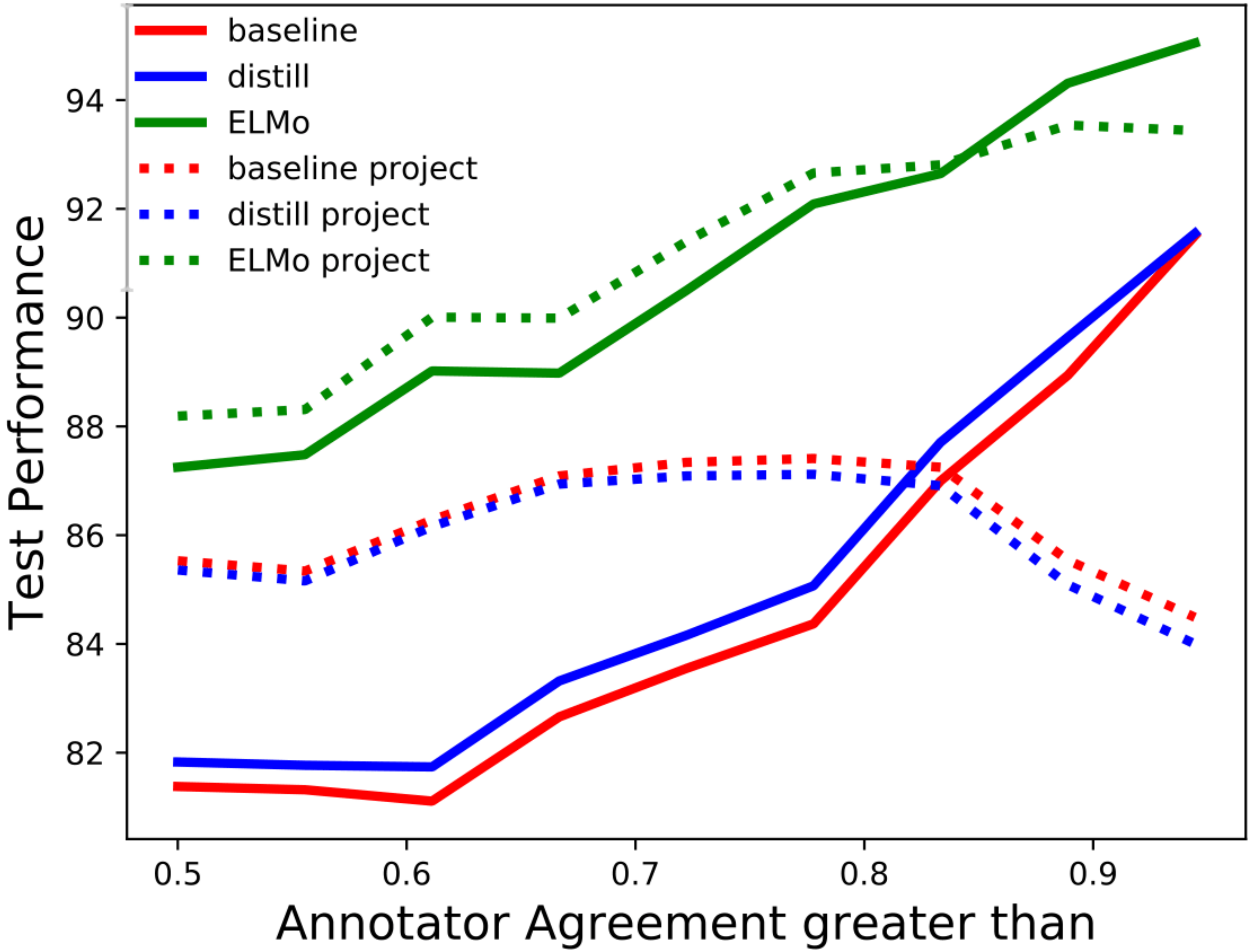
beautiful film, but those who have read the book will be disappointed

nine crowd-workers label each A-but-B sentence as positive / negative / neutral

we test our models on subsets of varying agreement



Consistent trends on all levels of agreement



Projection degrades accuracy on high agreement sentences!

Key Takeaways

- Carefully perform sentiment classification research
  - variation across runs - average across several seeds
  - ambiguous sentences - benchmark on subsets of varying annotator agreement
- ELMo embeddings **implicitly** learn logic rules for sentiment classification

Code + Data

github.com/martiansideofthemoon/logic-rules-sentiment

Key Takeaways

Thank You!

- Carefully perform sentiment classification research
  - variation across runs - average across several seeds
  - ambiguous sentences - benchmark on subsets of varying annotator agreement
- ELMo embeddings **implicitly** learn logic rules for sentiment classification

Code + Data

github.com/martiansideofthemoon/logic-rules-sentiment