Yandex

UNIVERSITEIT VAN AMSTERDAM

Transfer learning
Multi-task learning in NLP

NOVEMBER 8, 2018

Elena Voita
Yandex Research,
University of Amsterdam
lena-voita@yandex-team.ru

Plan

- What and why does the network learn?
- “Model” is never really just “Model”: it’s “Model + data + loss”
- Transfer learning in NLP
- Multi-task learning in NLP
- A piece of understanding
- A piece of practice
- Hack of the day (real-life story inside!)

What and why does the network learn?

RECAP: CNN Language Models

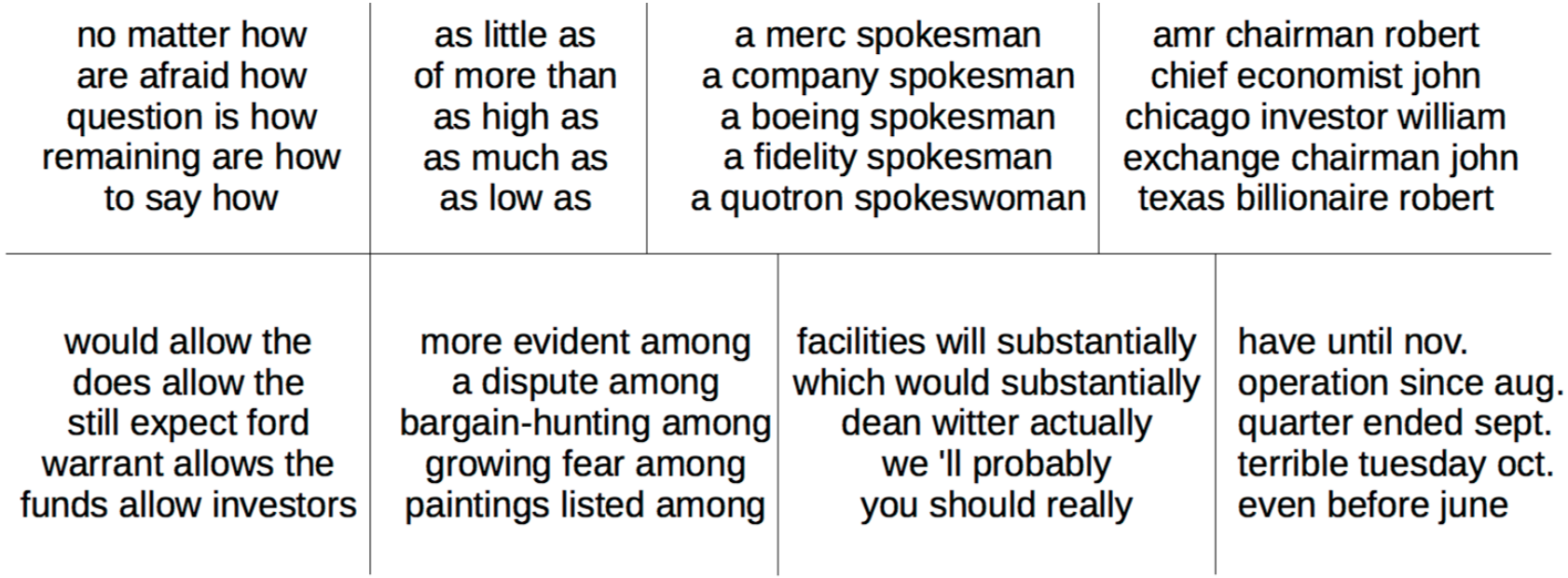
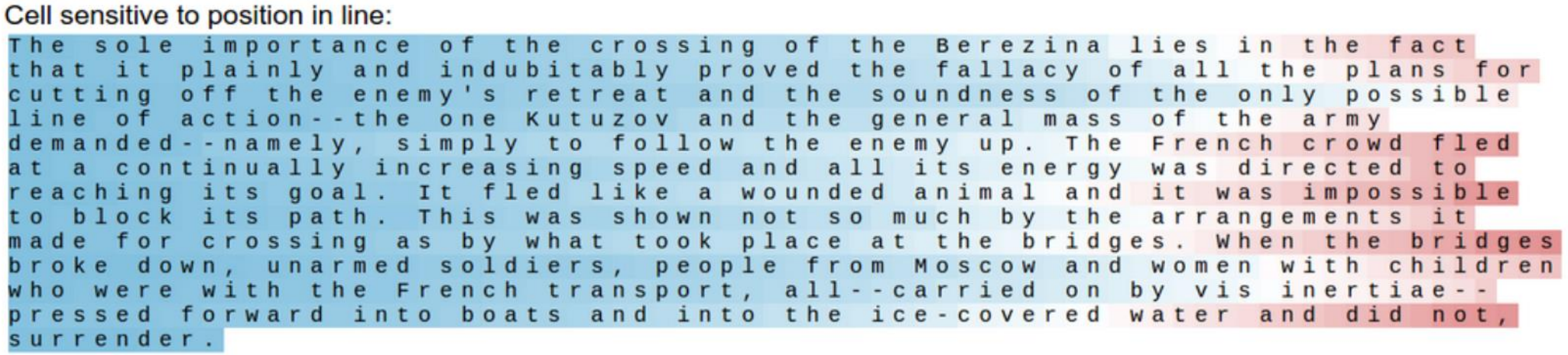


Figure 4: Some example phrases that have highest activations for 8 example kernels (each box), extracted from the validation set of the Penn Treebank. Model trained with 256 kernels for 256-dimensional word vectors.

Pham, et al, EMNLP 2016, https://www.aclweb.org/anthology/D16-1123

RECAP: Visualizing and Understanding Recurrent Networks

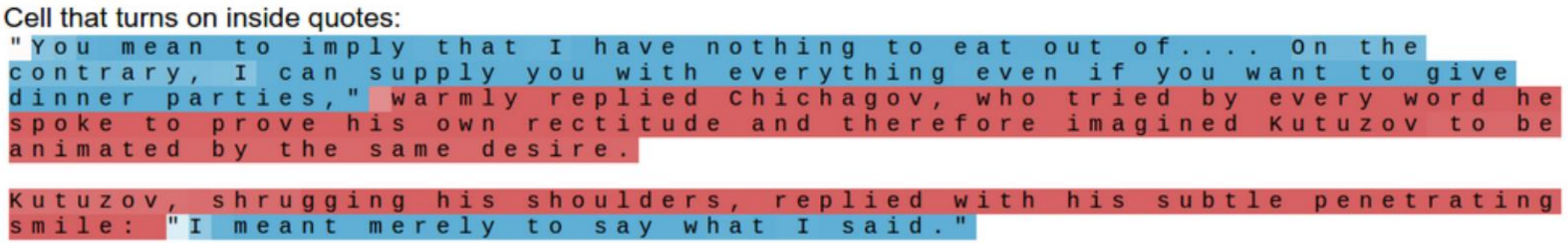
Train char-based RNN (LSTM) language model. The input character sequence (blue/green) is colored based on the *firing* of a randomly chosen neuron in the hidden representation of the RNN.



Karpathy et al, ICLR workshop, 2016 https://arxiv.org/pdf/1506.02078.pdf

RECAP: Visualizing and Understanding Recurrent Networks

Train char-based RNN (LSTM) language model. The input character sequence (blue/green) is colored based on the *firing* of a randomly chosen neuron in the hidden representation of the RNN.



Karpathy et al, ICLR workshop, 2016 https://arxiv.org/pdf/1506.02078.pdf

RECAP: Visualizing and Understanding Recurrent Networks

Train char-based RNN (LSTM) language model. The input character sequence (blue/green) is colored based on the *firing* of a randomly chosen neuron in the hidden representation of the RNN.

Cell that is sensitive to the depth of an expression:

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
  int i;
  if (classes[class]) {
    for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
      if (mask[i] & classes[class][i])
        return 0;
  }
  return 1;
}
```

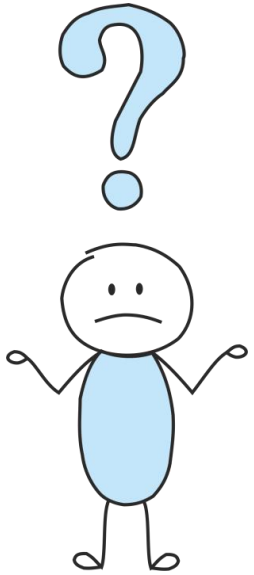
Karpathy et al, ICLR workshop, 2016 https://arxiv.org/pdf/1506.02078.pdf

Do you think that the model architecture is the main aspect in task learning?

Cell that turns on inside quotes:

```
"You mean to imply that I have nothing to eat out of.... On the
contrary, I can supply you with everything even if you want to give
dinner parties," warmly replied Chichagov, who tried by every word he
spoke to prove his own rectitude and therefore imagined Kutuzov to be
animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating
smile: "I meant merely to say what I said."
```



Cell that is sensitive to the depth of an expression:

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
  int i;
  if (classes[class]) {
    for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
      if (mask[i] & classes[class][i])
        return 0;
  }
  return 1;
}
```

Do you think that the model architecture is the main aspect in task learning?

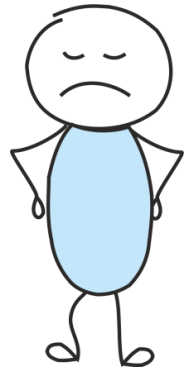
Cell that turns on inside quotes:

```
"You mean to imply that I have nothing to eat out of.... On the
contrary, I can supply you with everything even if you want to give
dinner parties," warmly replied Chichagov, who tried by every word he
spoke to prove his own rectitude and therefore imagined Kutuzov to be
animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating
smile: "I meant merely to say what I said."
```

Cell that is sensitive to the depth of an expression:

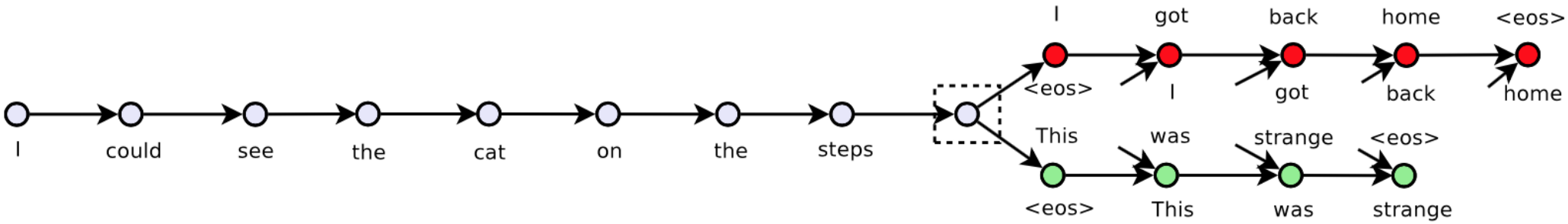
```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
  int i;
  if (classes[class]) {
    for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
      if (mask[i] & classes[class][i])
        return 0;
  }
  return 1;
}
```



“Model” is **never** just “model”!

RECAP: Skip-Thought Vectors

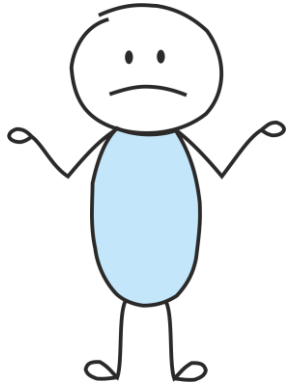
- encode a sentence to predict the sentences around it



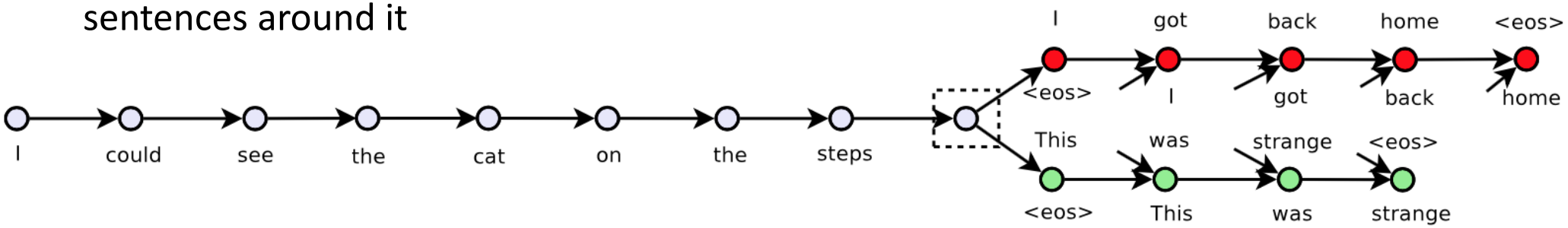
Kiros et al., NIPS 2015 https://papers.nips.cc/paper/5950-skip-thought-vectors.pdf

RECAP: Skip-Thought Vectors

What does this model remind you?



- encode a sentence to predict the sentences around it

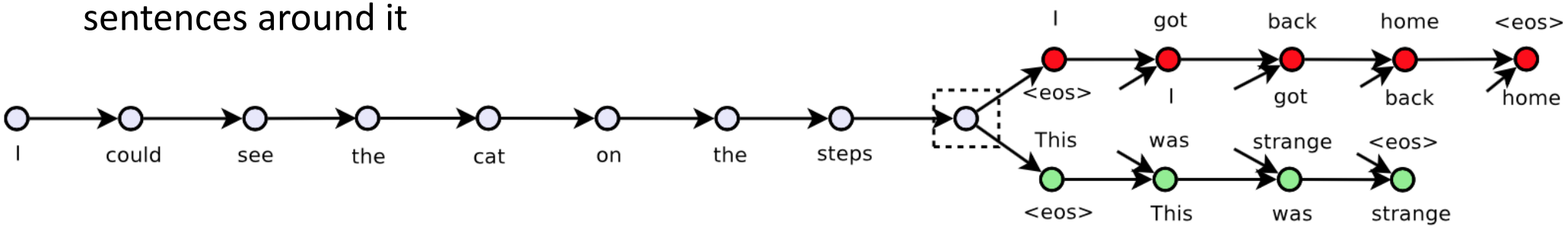


Kiros et al., NIPS 2015 https://papers.nips.cc/paper/5950-skip-thought-vectors.pdf

RECAP: Skip-Thought Vectors

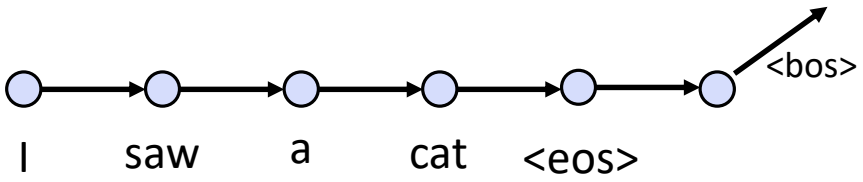
Machine Translation! (no attention)

- encode a sentence to predict the sentences around it

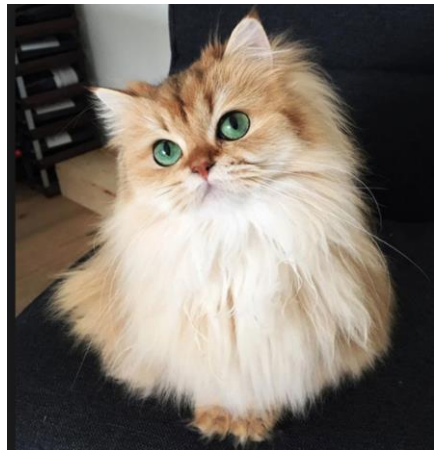


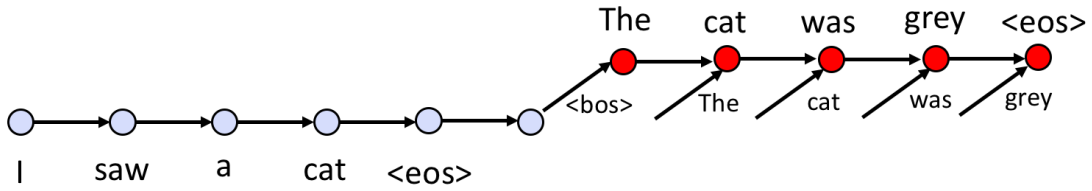
Kiros et al., NIPS 2015 https://papers.nips.cc/paper/5950-skip-thought-vectors.pdf

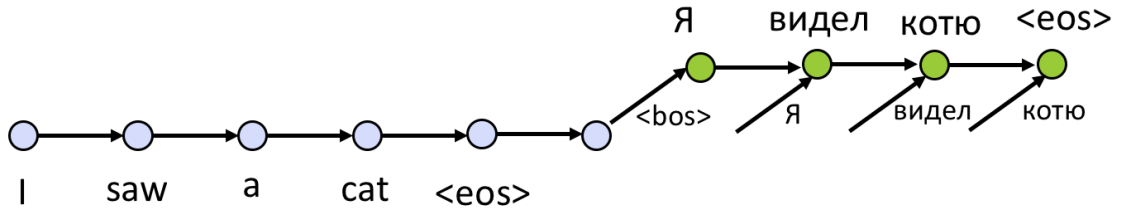
The same model, different tasks



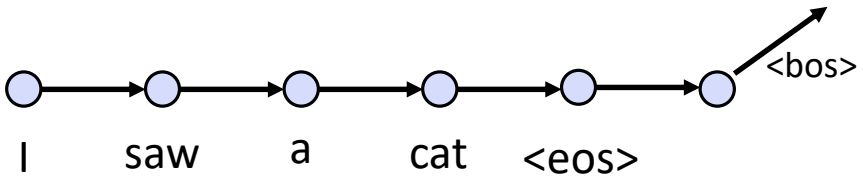
?





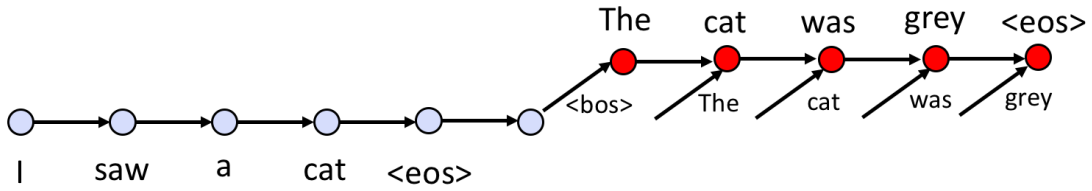


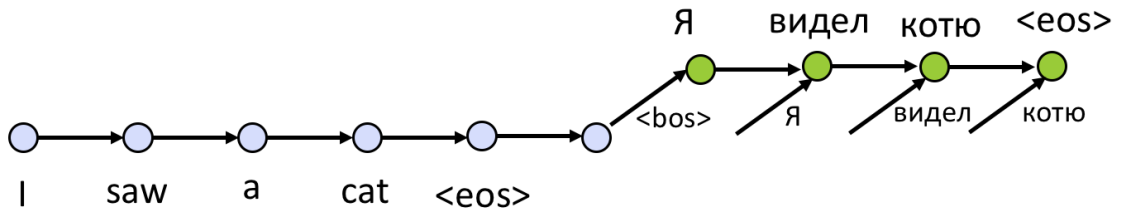
The same model, different tasks



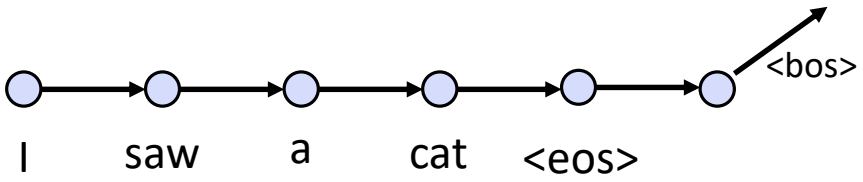
?

Do you think that learned representations are roughly the same?





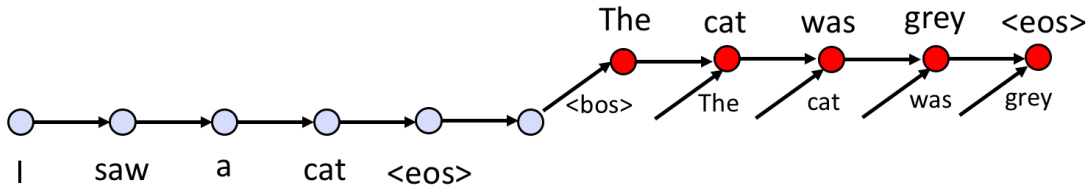
The same model, different tasks

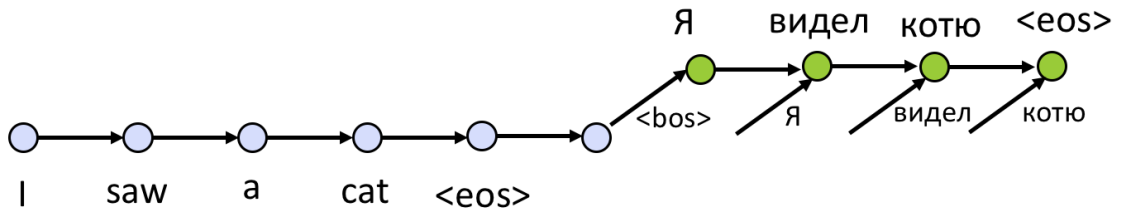


?

Do you think that learned representations are roughly the same?

Not really!





“Model” is never just “model”!

But what is it then?

“Model” is never just “model”!

But what is it then?

“model” = model + data + task

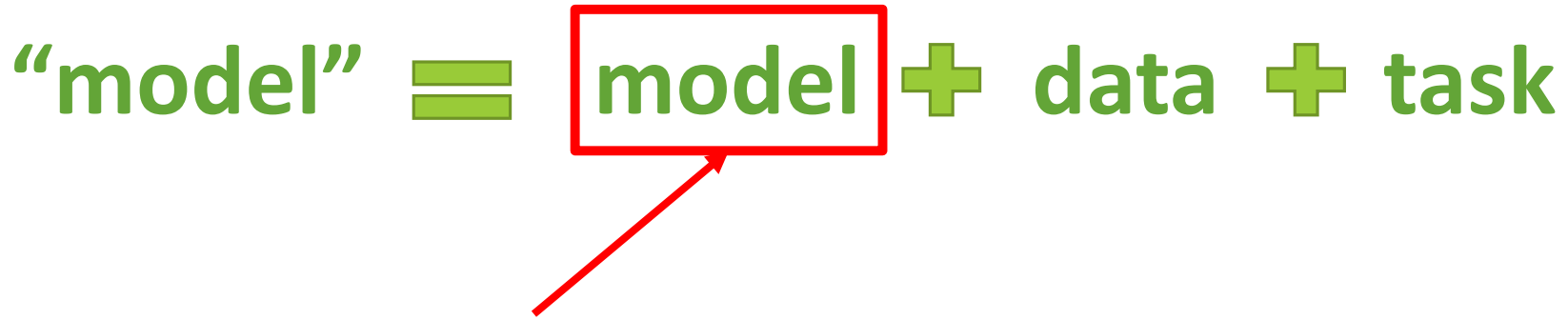
“Model” is never just “model”!

But what is it then?

“model” = model + data + task

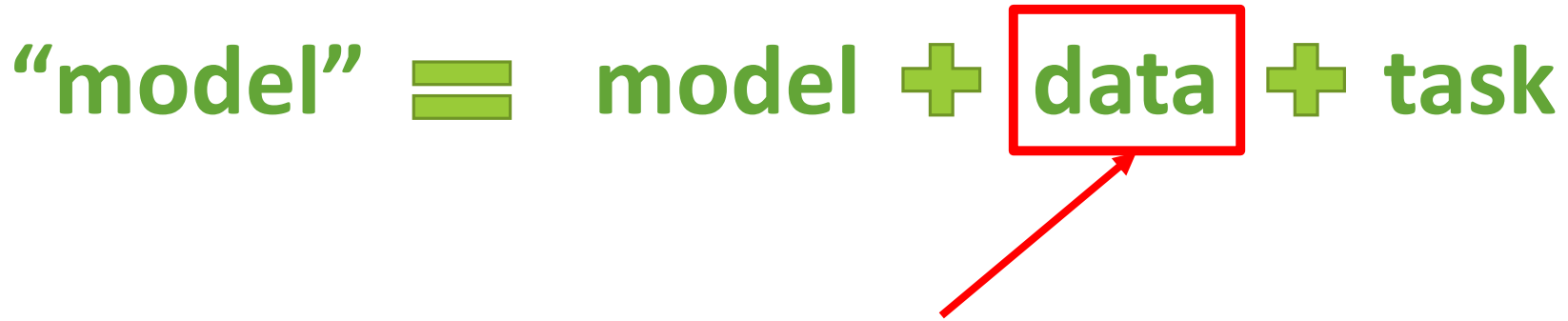
(And also your optimization algorithm, lr schedule, batch size, …
But let’s not be so pessimistic)

“Model” is never just “model”!



CNN/RNN, attention, Transformer, etc,
and their hp

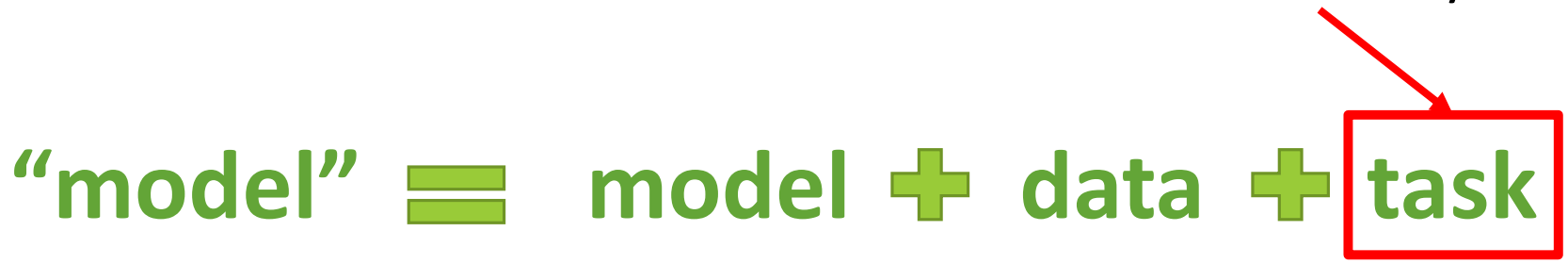
“Model” is never just “model”!



- Domain adaptation (**in the next episode!**)
- Use data for other task (unlabeled data)

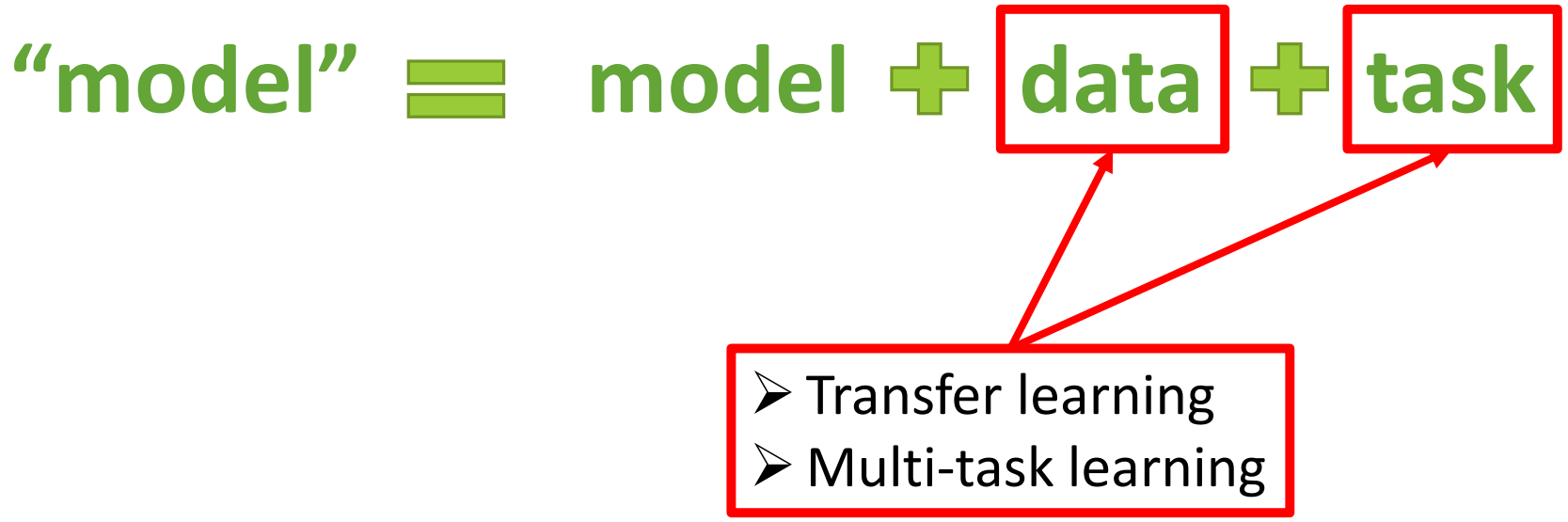
“Model” is never just “model”!

- Use the signal from some task to make the model learn what you want

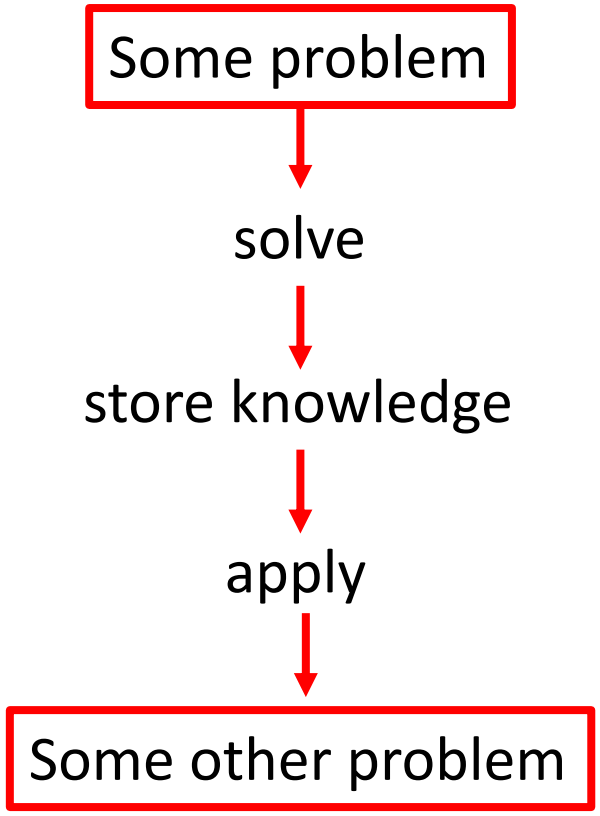


Transfer learning vs Multi-task learning

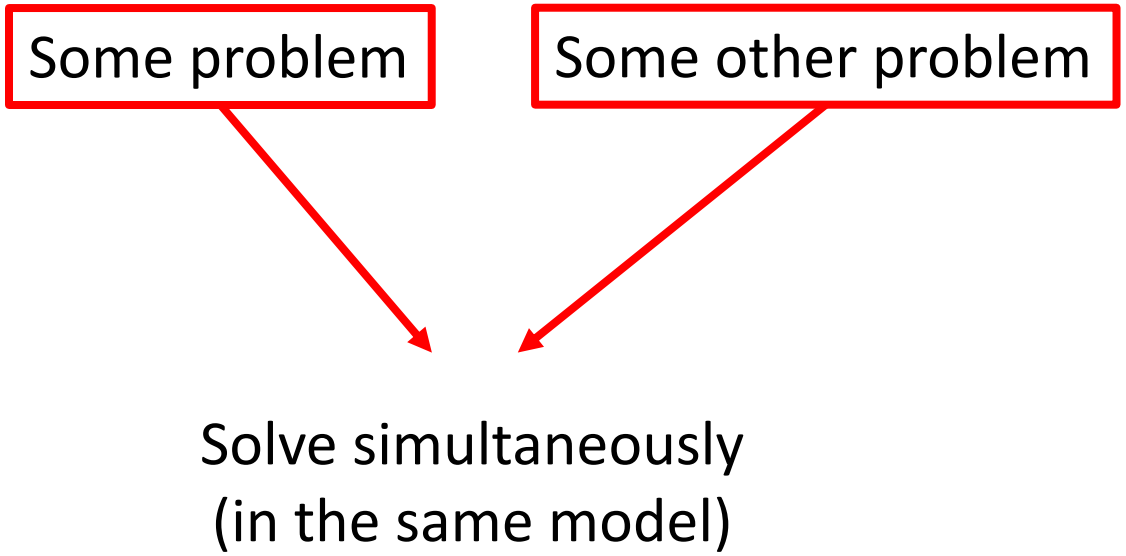
Transfer learning vs Multi-task learning



Transfer learning



Multi-task learning

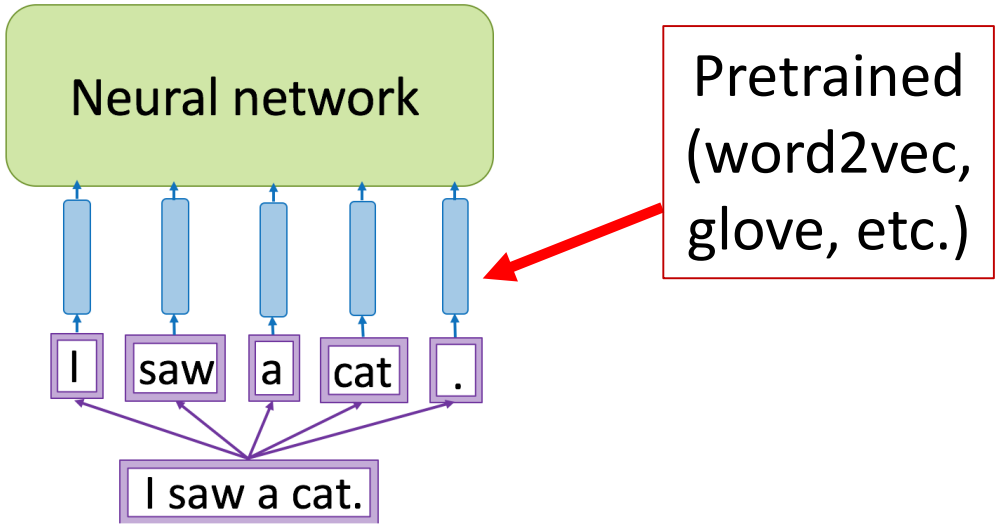


Transfer learning in NLP

RECAP: When do we use pretrained embeddings?

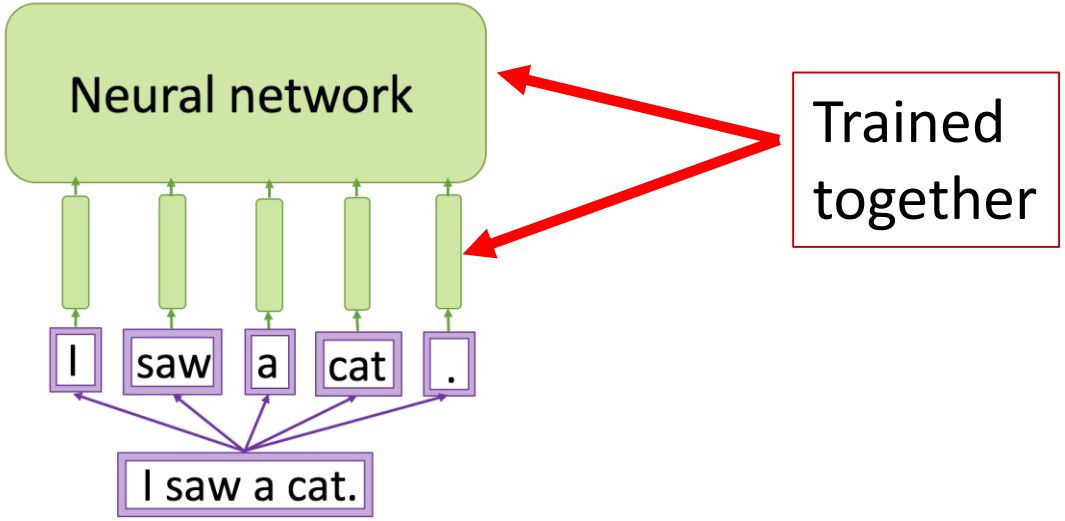
Not enough data or the task is too simple

Use pretrained on the other task



Enough data and a hard task (LM, MT, …)

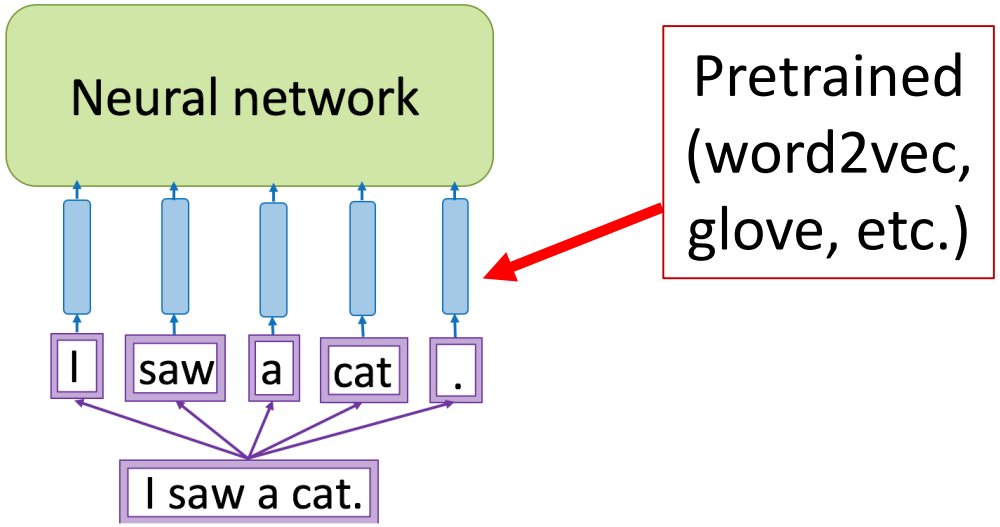
Train with the model



RECAP: When do we use pretrained embeddings?

Not enough data or the task is too simple

Use pretrained on the other task



This is transfer learning!

Transfer knowledge from embedding training to your model

RECAP: Deep contextualized word representations (ELMO)

- Word embedding: concat(embedding of a word, char-cnn)
- Train Bi-LSTM LM on a large corpus
- Use weighted sum of hidden representations from different layers, weights are trained with the task

Why profit?

- char-based information used
- context of a sentence is used
- information from different layers
- task-specific weights for layers

(**ELMO** – **E**mbeddings from **L**anguage **MO**del)

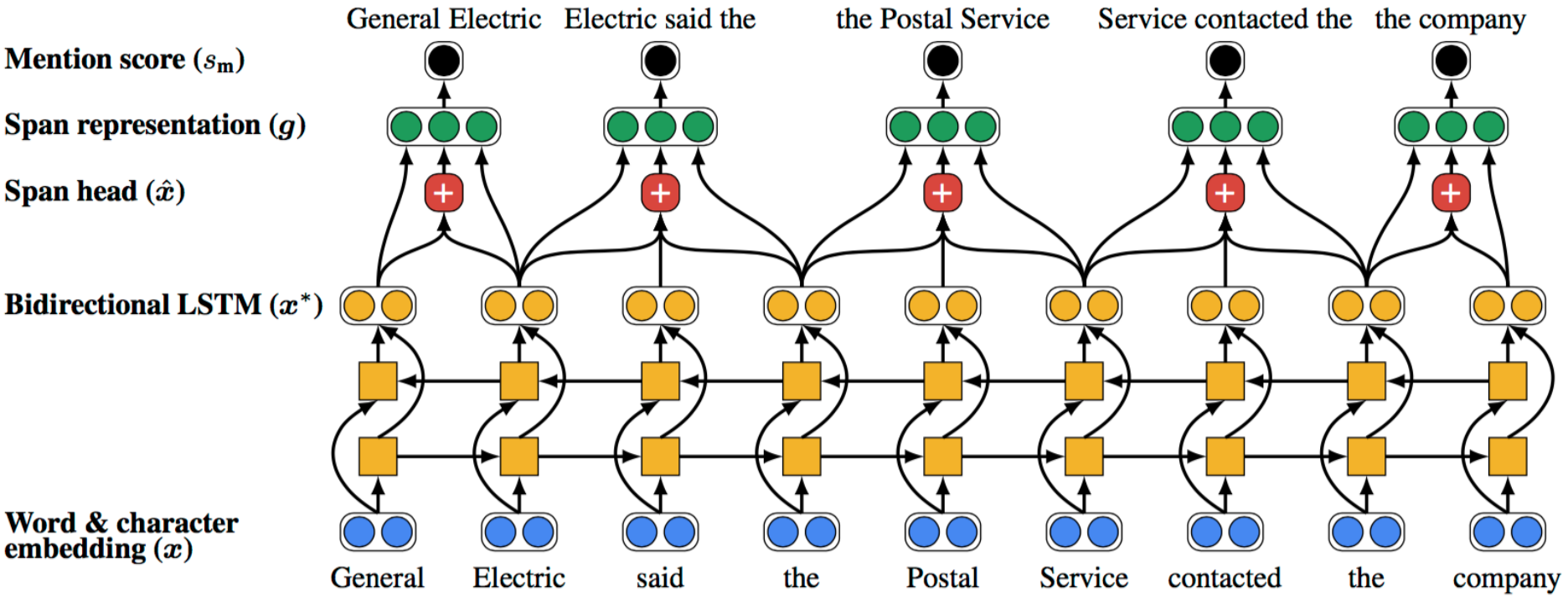
Peters et al, NAACL 2018, http://aclweb.org/anthology/N18-1202

Deep contextualized word representations (ELMO)

TASK	PREVIOUS SOTA		OUR BASELINE	ELMO + BASELINE	INCREASE (ABSOLUTE/ RELATIVE)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 $\pm$ 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 $\pm$ 0.19	90.15	92.22 $\pm$ 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 $\pm$ 0.5	3.3 / 6.8%

Table 1: Test set comparison of ELMo enhanced neural models with state-of-the-art single model baselines across six benchmark NLP tasks. The performance metric varies across tasks – accuracy for SNLI and SST-5; F_1 for SQuAD, SRL and NER; average F_1 for Coref. Due to the small test sizes for NER and SST-5, we report the mean and standard deviation across five runs with different random seeds. The "increase" column lists both the absolute and relative improvements over our baseline.

ELMO for coreference resolution



Lee et al, EMNLP 2017, http://aclweb.org/anthology/D17-1018

Learned in Translation: Contextualized Word Vectors (CoVe)

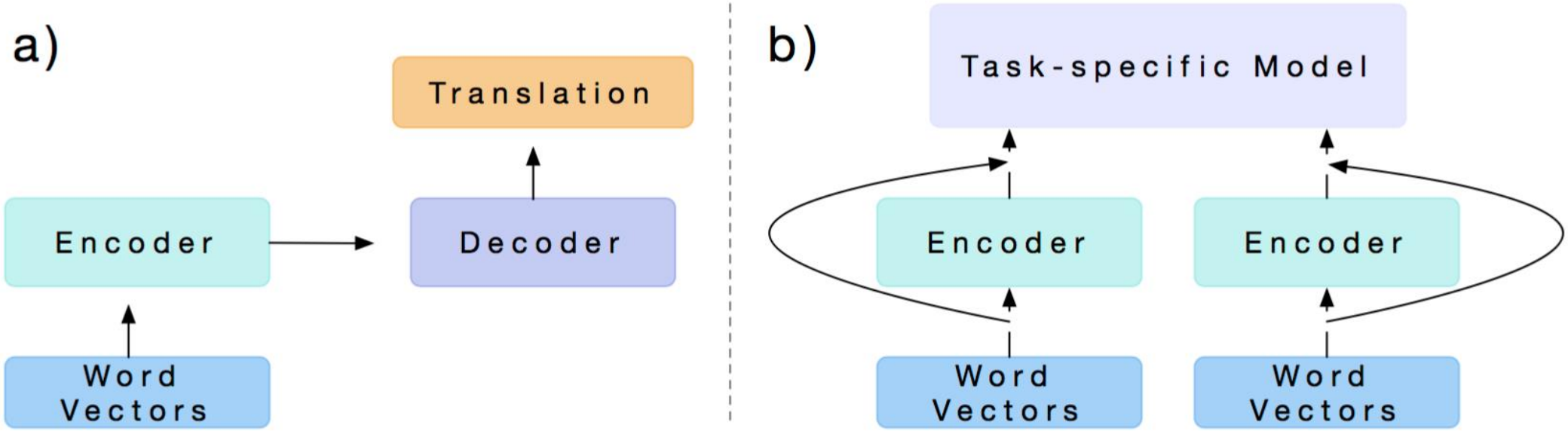
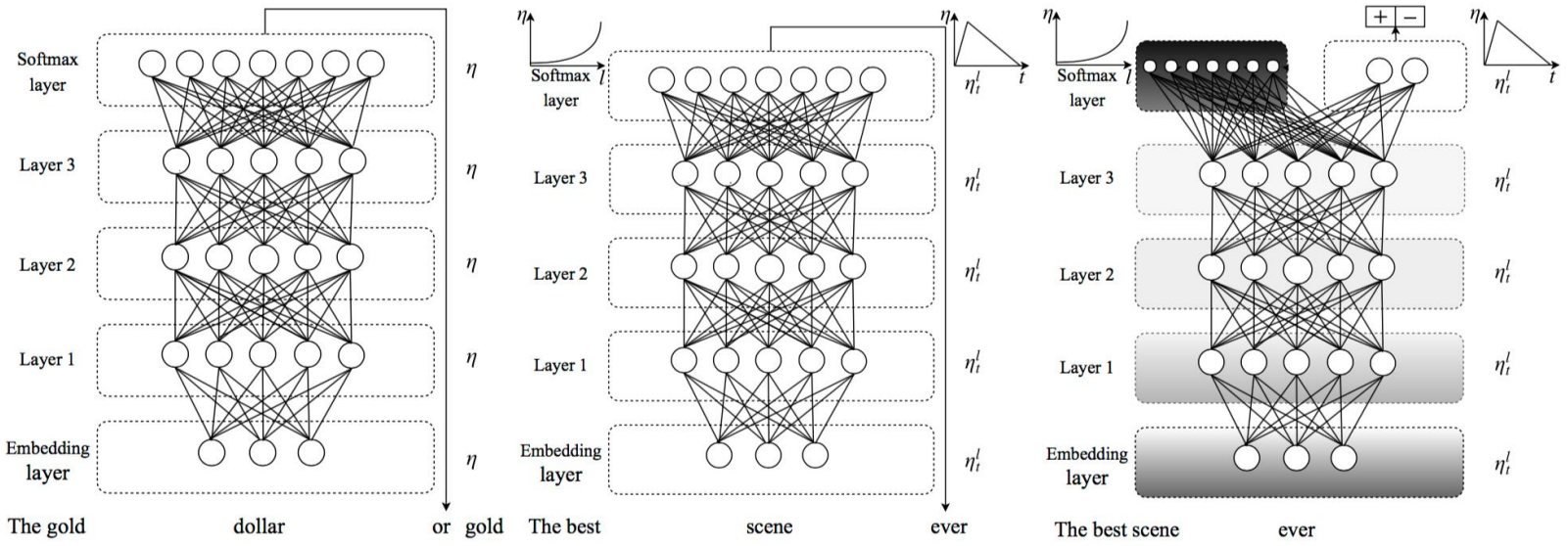


Figure 1: We a) train a two-layer, bidirectional LSTM as the encoder of an attentional sequence-to-sequence model for machine translation and b) use it to provide context for other NLP models.

McCann et al, NIPS 2017, https://papers.nips.cc/paper/7209-learned-in-translation-contextualized-word-vectors.pdf

Universal Language Model Fine-tuning for Text Classification



(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

Figure 1: ULMFiT consists of three stages: a) The LM is trained on a general-domain corpus to capture general features of the language in different layers. b) The full LM is fine-tuned on target task data using discriminative fine-tuning ('*Discr*') and slanted triangular learning rates (STLR) to learn task-specific features. c) The classifier is fine-tuned on the target task using gradual unfreezing, '*Discr*', and STLR to preserve low-level representations and adapt high-level ones (shaded: unfreezing stages; black: frozen).

Howard & Ruder, ACL 2018, http://aclweb.org/anthology/P18-1031

Universal Language Model Fine-tuning for Text Classification

Why language model?

- it is able to capture long-term dependencies in language
- it effectively incorporates hierarchical relations
- it can help the model learn sentiments
- large data corpus is easily available for LM

Howard & Ruder, ACL 2018, http://aclweb.org/anthology/P18-1031

Multi-task learning in NLP

Joint training of existing NLP tasks

- Machine Translation: different language pairs together
- MT + POS, NER, parsing, image captioning
- Jointly train for different languages helps for: POS, NER, parsing, document classification, discourse segmentation, sequence tagging
- Video captioning and textual entailment generation
- different parsings together
- different tasks for representation learning
- and much, much more

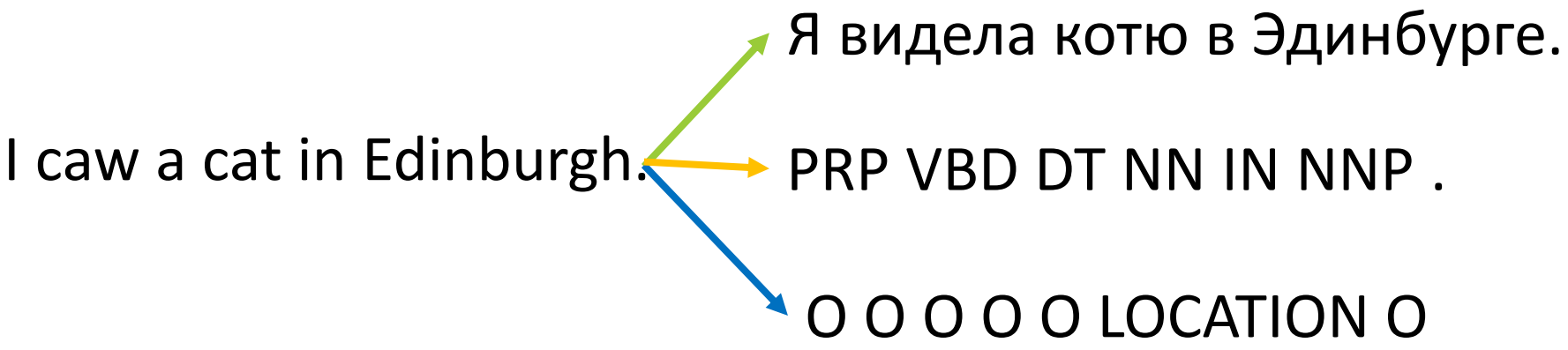
To know more: http://ruder.io/multi-task-learning-nlp

Joint training of existing NLP tasks

- Machine Translation: different language pairs together
- MT + POS, NER, parsing, image captioning **-> MT+POS+NER**
- Jointly train for different languages helps for: POS, NER, parsing, document classification, discourse segmentation, sequence tagging
- **Video captioning and textual entailment generation**
- different parsings together
- **different tasks for representation learning**
- and much, much more **–> a bit of this**

To know more: http://ruder.io/multi-task-learning-nlp

Exploiting Linguistic Resources for Neural Machine Translation Using Multi-task Learning



Tasks:

- MT
- POS
- NER

POS demo online: http://cogcomp.org/page/demo_view/pos
NER demo online: http://nlp.stanford.edu:8080/ner/process

Niehues et al, WMT 2017, http://www.statmt.org/wmt17/pdf/WMT08.pdf

Adventures of a PhD student in Edinburgh

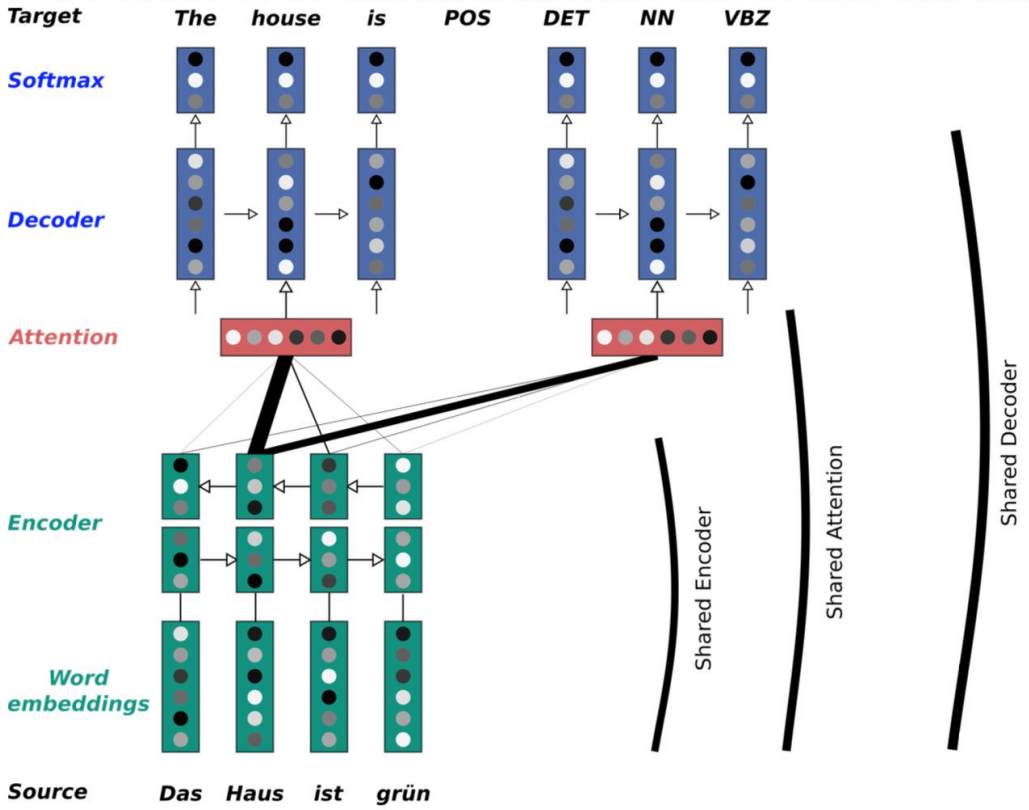
By the way, it's
the Edinburgh University's
library cat

(Can't believe it? Look here: https://twitter.com/edinlibrarycat)

THE UNIVERSITY *of* EDINBURGH

Exploiting Linguistic Resources for Neural Machine Translation Using Multi-task Learning

Figure 1: Overview on the different architectures used for multi-task learning



Tasks:

- MT
- POS
- NER

Niehues et al, WMT 2017, http://www.statmt.org/wmt17/pdf/WMT08.pdf

Exploiting Linguistic Resources for Neural Machine Translation Using Multi-task Learning

Task(s)	Arch.	Valid	Test	
		dev 2010	tst2013	tst2014
MT	-	29.91/62.16/51.06	30.85/62.27/51.16	26.12/58.73/55.17
POS + MT	shrd Enc	30.62/62.77/48.35	31.97/62.72/49.69	27.08/58.99/54.50
	shrd Att	30.51/62.27/49.09	31.76/62.68/49.59	26.86/58.84/53.88
	shrd Dec	30.36/62.34/49.28	31.26/62.31/50.35	26.52/58.48/54.00
Adapted NE + POS + MT	shrd Enc	30.70/62.96/48.60	32.30/63.25/49.22	27.78/59.74/53.49

Table 1: Results of multi-task learning architectures on the machine translation task (BLEU/BEER/characTER)

Tasks:

- MT
- POS
- NER

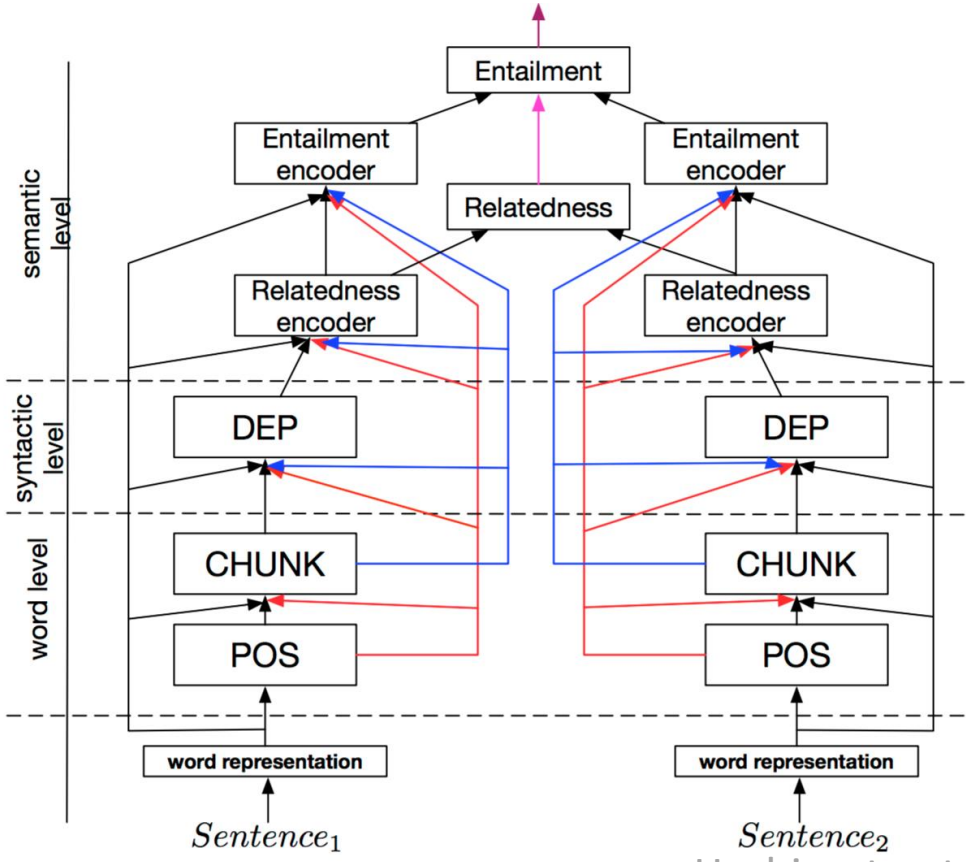
Niehues et al, WMT 2017, http://www.statmt.org/wmt17/pdf/WMT08.pdf

Exploiting Linguistic Resources for Neural Machine Translation Using Multi-task Learning

German	sie ist kein Geburtsfehler.
Reference	it's not a birth defect.
Baseline	she's not born.
Multi-task	it's not a birth error.
German	das bedeutet, dass 8 von 10 Entscheidungen...
Reference	that means that eight out of 10 of the decisions...
Baseline	that means that eight of 10 of 10 choices...
Multi-task	that means that eight of 10 decisions...
German	...["Benjamin Franklin" von Walter Isaacson]["John Adams" von David McCullough]...
Reference	..."Benjamin Franklin" by Walter Isaacson]["John Adams" by David McCullough]...
Baseline	...[Benjamin Franklin, from Walter Franklin"][The "John Adams"]...
Multi-task	..."Benjamin Franklin" from Walter Isaacson],["John Adams" from David McCullough...

Niehues et al, WMT 2017, http://www.statmt.org/wmt17/pdf/WMT08.pdf

A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks



Textual entailment

↑

Semantic relatedness

↑

Dependency parsing

↑

Chunking

↑

POS tagging

Hashimoto et al, EMNLP 2017, https://www.aclweb.org/anthology/D17-1206

Task example: Chunking

Text chunking consists of dividing a text in syntactically correlated parts of words. For example, the following sentence:

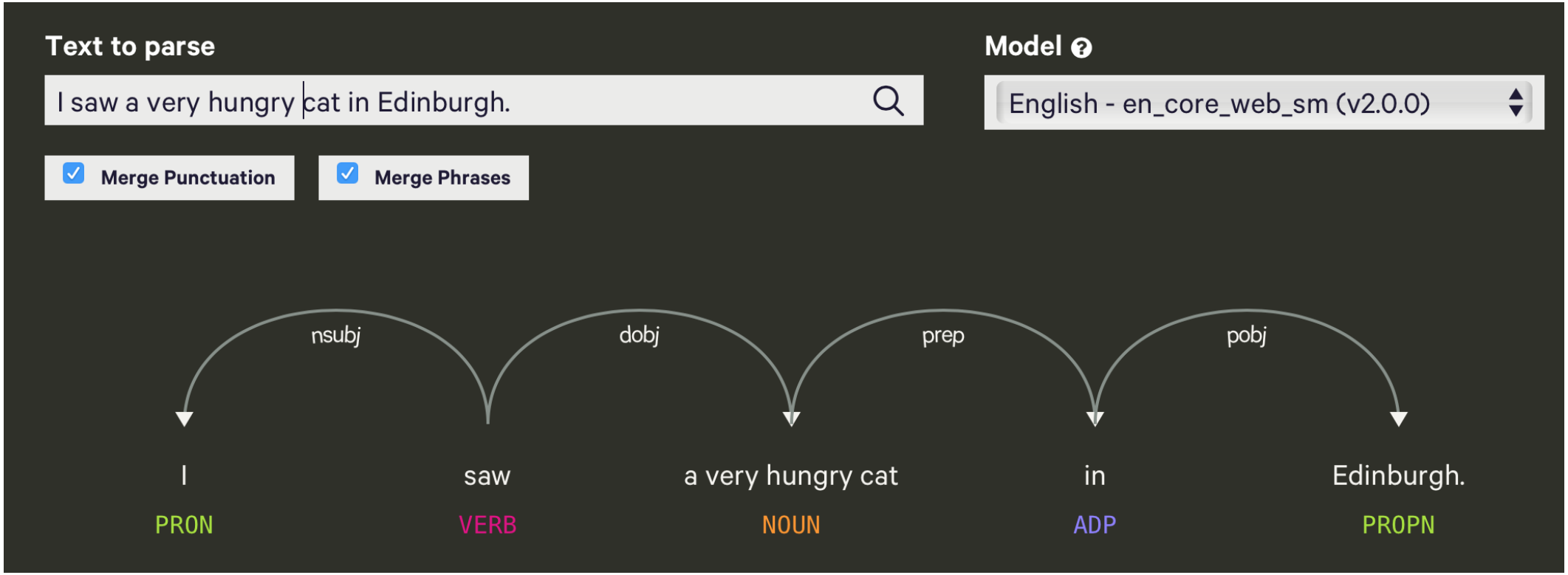
> He reckons the current account deficit will narrow to only # 1.8 billion in September .

can be divided as follows:

```
[NP He ] [VP reckons ] [NP the current account deficit ] [VP will narrow ] [PP to ] [NP only # 1.8 billion ] [PP in ] [NP Sep
tember ] .
```

Example from: http://deepdive.stanford.edu/example-chunking

Task example: Dependency parsing



Dependency parsing demo: https://explosion.ai/demos/displacy

Task example: Sematic relatedness

Relatedness score	Example
1.6	A: "*A man is jumping into an empty pool*" B: "*There is no biker jumping in the air*"
2.9	A: "*Two children are lying in the snow and are making snow angels*" B: "*Two angels are making snow on the lying children*"
3.6	A: "*The young boys are playing outdoors and the man is smiling nearby*" B: "*There is no boy playing outdoors and there is no man smiling*"
4.9	A: "*A person in a black jacket is doing tricks on a motorbike*" B: "*A man in a black jacket is doing tricks on a motorbike*"

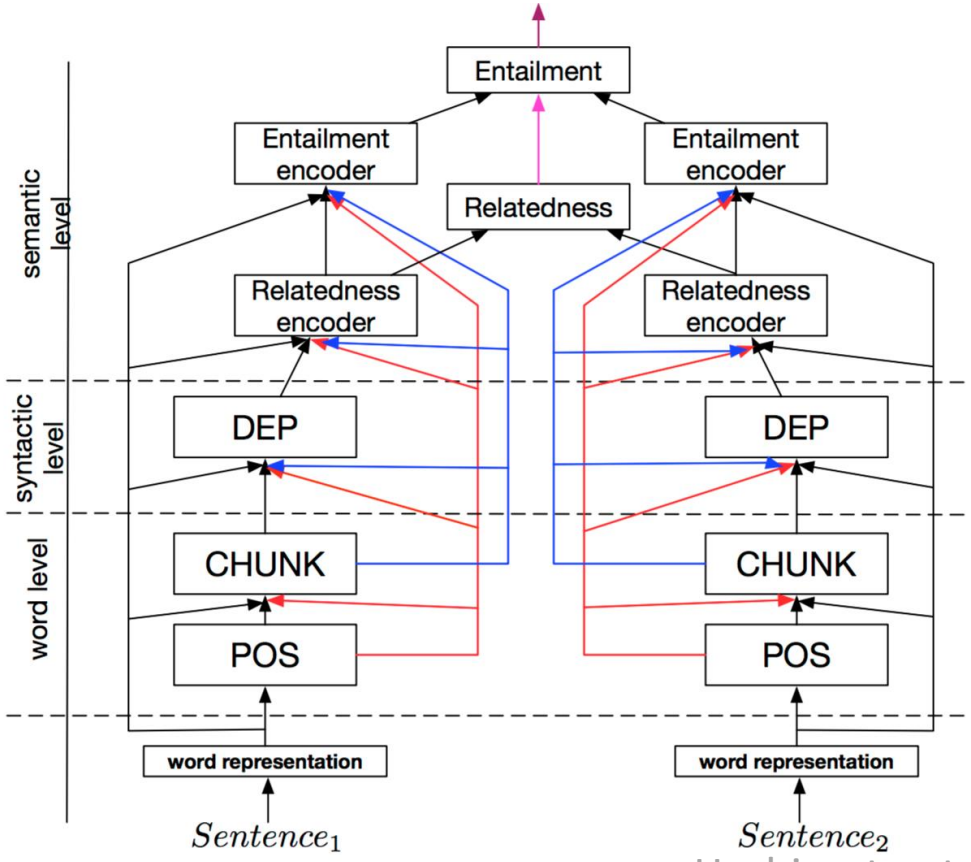
Examples from: http://alt.qcri.org/semeval2014/cdrom/pdf/SemEval2014001.pdf

Task example: Textual entailment

Entailment label	Example
ENTAILMENT	A: "*Two teams are competing in a football match*" B: "*Two groups of people are playing football*"
CONTRADICTION	A: "*The brown horse is near a red barrel at the rodeo*" B: "*The brown horse is far from a red barrel at the rodeo*"
NEUTRAL	A: "*A man in a black jacket is doing tricks on a motorbike*" B: "*A person is riding the bicycle on one wheel*"

Examples from: http://alt.qcri.org/semeval2014/cdrom/pdf/SemEval2014001.pdf

A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks



Textual entailment

↑

Semantic relatedness

↑

Dependency parsing

↑

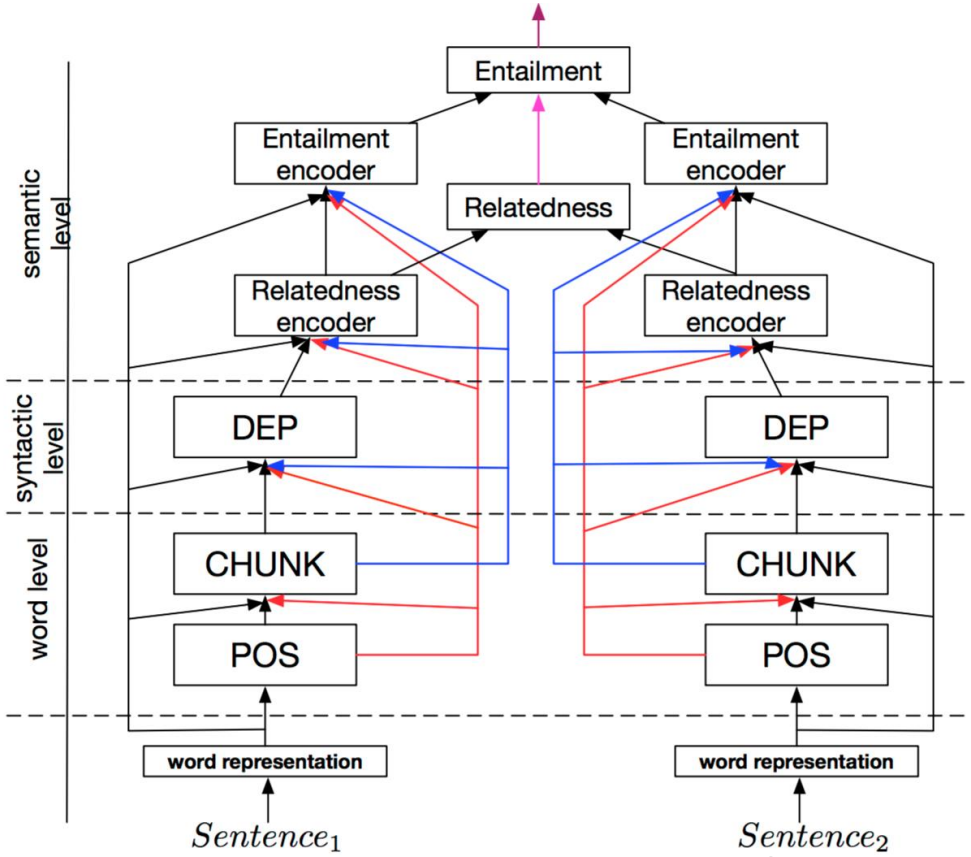
Chunking

↑

POS tagging

Hashimoto et al, EMNLP 2017, https://www.aclweb.org/anthology/D17-1206

A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks

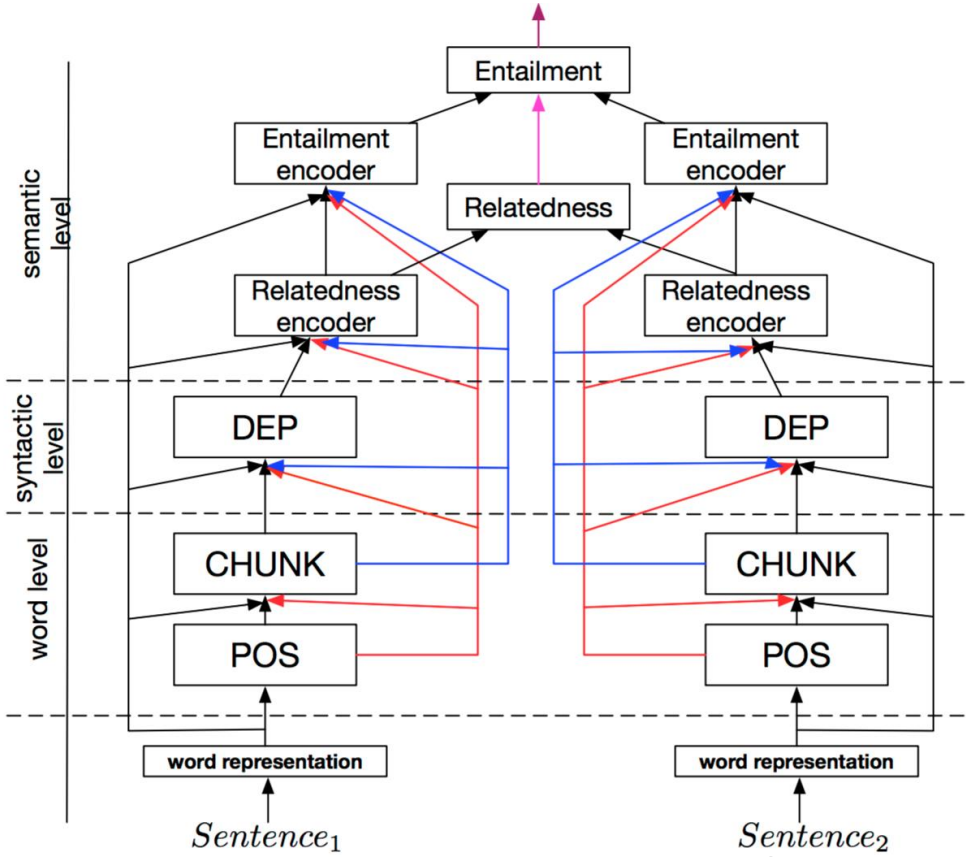


$$J_1(\theta_{\text{POS}}) = -\sum_s \sum_t \log p(y_t^{(1)} = \alpha | h_t^{(1)}) + \lambda \|W_{\text{POS}}\|^2 + \delta \|\theta_e - \theta_e'\|^2,$$

Don't want to forget what we've learned on higher tasks

Hashimoto et al, EMNLP 2017, https://www.aclweb.org/anthology/D17-1206

A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks



$$J_2(\theta_{\text{chk}}) = -\sum_s \sum_t \log p(y_t^{(2)} = \alpha | h_t^{(2)}) + \lambda \|W_{\text{chk}}\|^2 + \delta \|\theta_{\text{POS}} - \theta'_{\text{POS}}\|^2,$$

Don't want to forget what we've learned on higher tasks

Hashimoto et al, EMNLP 2017, https://www.aclweb.org/anthology/D17-1206

A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks

		Single	JMT_{all}	JMT_{AB}	JMT_{ABC}	JMT_{DE}	JMT_{CD}	JMT_{CE}
A ↑	POS	97.45	97.55	97.52	97.54	n/a	n/a	n/a
B ↑	Chunking	95.02	n/a	95.77	n/a	n/a	n/a	n/a
C ↑	Dependency UAS	93.35	94.67	n/a	94.71	n/a	93.53	93.57
	Dependency LAS	91.42	92.90	n/a	92.92	n/a	91.62	91.69
D ↓	Relatedness	0.247	0.233	n/a	n/a	0.238	0.251	n/a
E ↑	Entailment	81.8	86.2	n/a	n/a	86.8	n/a	82.4

Table 1: Test set results for the five tasks. In the relatedness task, the lower scores are better.

Hashimoto et al, EMNLP 2017, https://www.aclweb.org/anthology/D17-1206

Multi-Task Video Captioning with Video and Entailment Generation

Baselines:

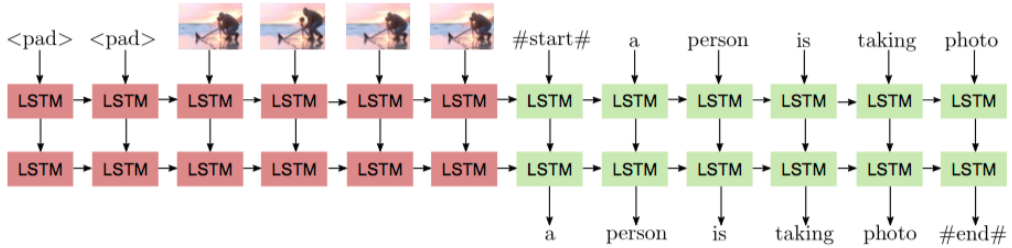


Figure 2: Baseline sequence-to-sequence model for video captioning: standard encoder-decoder LSTM-RNN model.

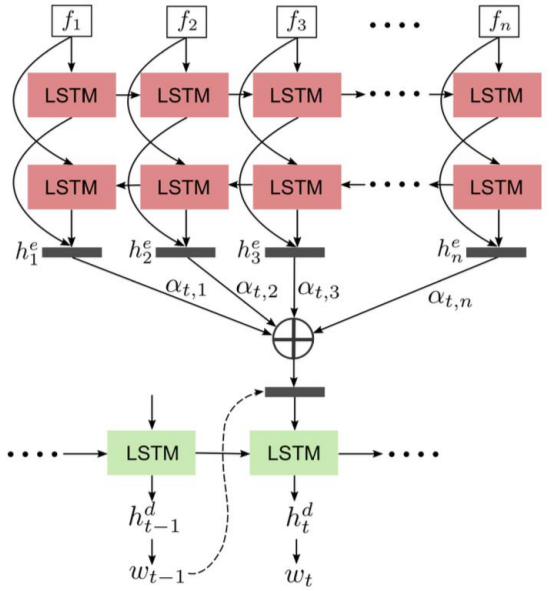
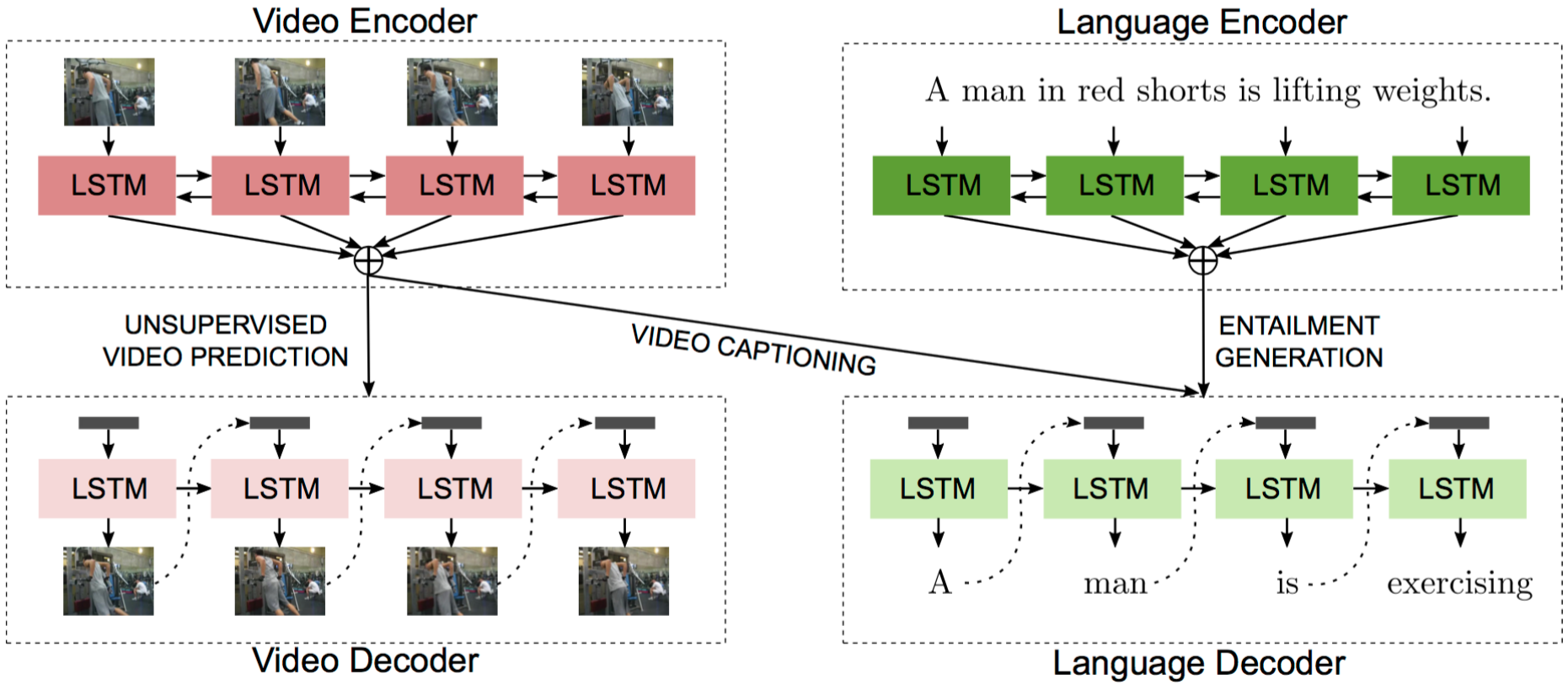


Figure 3: Attention-based sequence-to-sequence baseline model for video captioning (similar models also used for video prediction and entailment generation).

Pasunuru & Bansal, ACL 2017, http://aclweb.org/anthology/P17-1117

Multi-Task Video Captioning with Video and Entailment Generation



Pasunuru & Bansal, ACL 2017, http://aclweb.org/anthology/P17-1117

Multi-Task Video Captioning with Video and Entailment Generation

Models	METEOR	CIDEr-D	ROUGE-L	BLEU-4
PREVIOUS WORK				
LSTM-YT (V) (Venugopalan et al., 2015b)	26.9	-	-	31.2
S2VT (V + A) (Venugopalan et al., 2015a)	29.8	-	-	-
Temporal Attention (G + C) (Yao et al., 2015)	29.6	51.7	-	41.9
LSTM-E (V + C) (Pan et al., 2016b)	31.0	-	-	45.3
Glove + DeepFusion (V) (E) (Venugopalan et al., 2016)	31.4	-	-	42.1
p-RNN (V + C) (Yu et al., 2016)	32.6	65.8	-	49.9
HNRE + Attention (G + C) (Pan et al., 2016a)	33.9	-	-	46.7
OUR BASELINES				
Baseline (V)	31.4	63.9	68.0	43.6
Baseline (G)	31.7	64.8	68.6	44.1
Baseline (I)	33.3	75.6	69.7	46.3
Baseline + Attention (V)	32.6	72.2	69.0	47.5
Baseline + Attention (G)	33.0	69.4	68.3	44.9
Baseline + Attention (I)	33.8	77.2	70.3	49.9
Baseline + Attention (I) (E) ⊗	35.0	84.4	71.5	52.6
OUR MULTI-TASK LEARNING MODELS				
⊗ + Video Prediction (1-to-M)	35.6	88.1	72.9	54.1
⊗ + Entailment Generation (M-to-1)	35.9	88.0	72.7	54.4
⊗ + Video Prediction + Entailment Generation (M-to-M)	**36.0**	**92.4**	**72.8**	**54.5**

Table 1: Primary video captioning results on Youtube2Text (MSVD), showing previous works, our several strong baselines, and our three multi-task models. Here, V, G, I, C, A are short for VGGNet, GoogLeNet, Inception-v4, C3D, and AlexNet visual features; E = ensemble. The multi-task models are applied on top of our best video captioning baseline ⊗, with an ensemble. All the multi-task models are statistically significant over the baseline (discussed inline in the corresponding results sections).

Pasunuru & Bansal, ACL 2017, http://aclweb.org/anthology/P17-1117

Multi-Task Video Captioning with Video and Entailment Generation

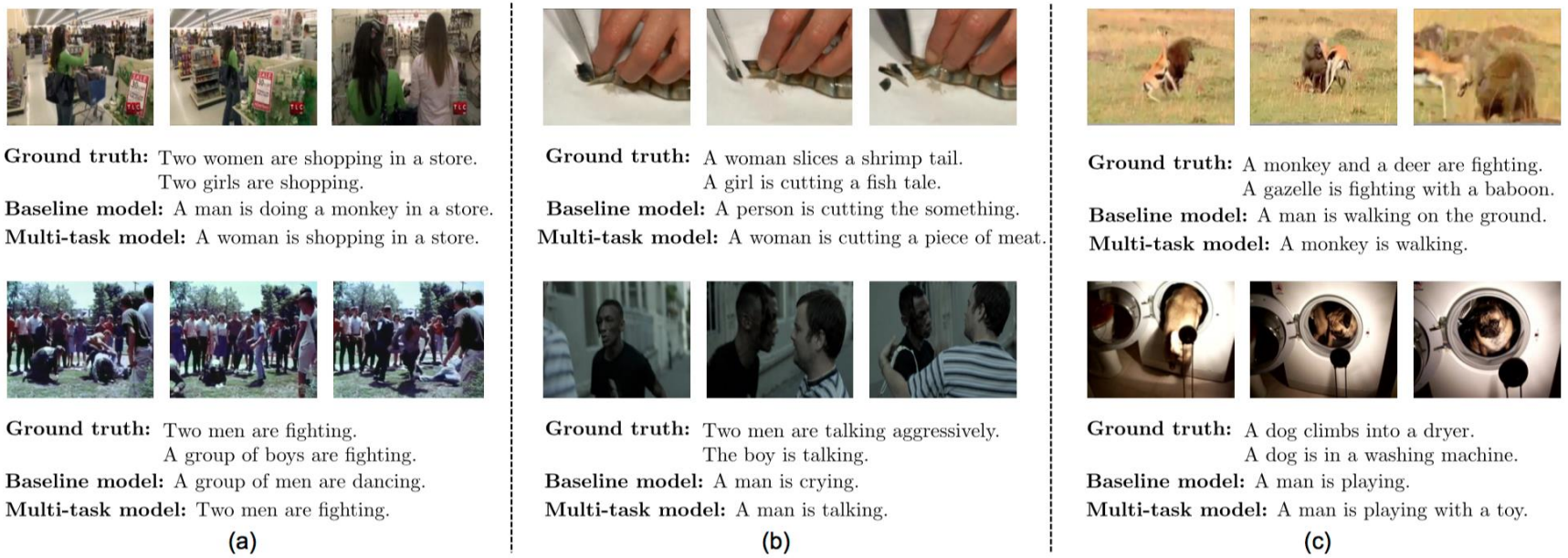


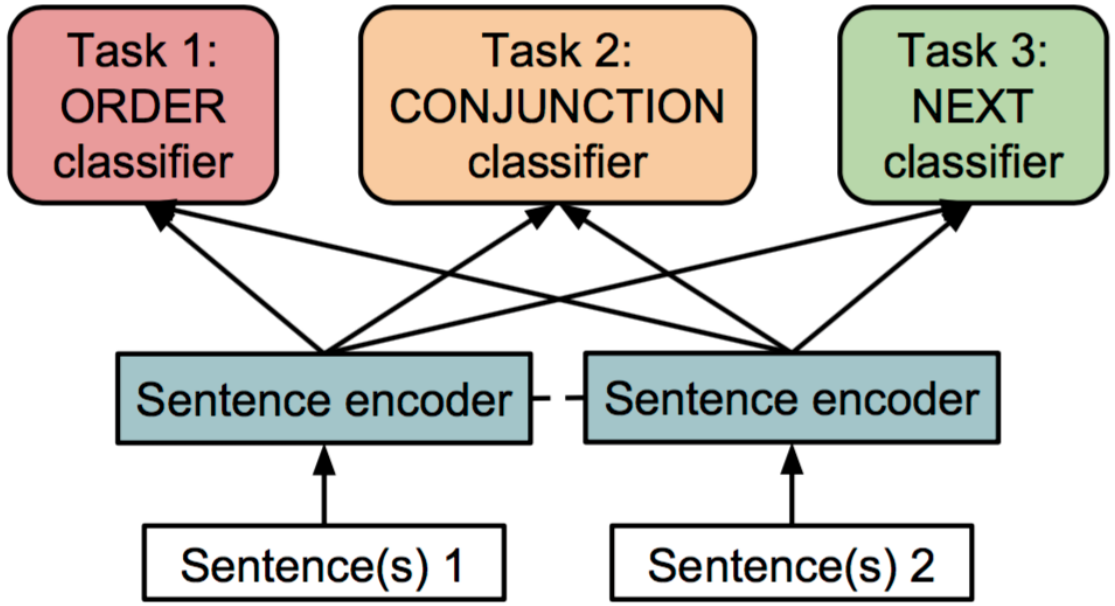
Figure 5: Examples of generated video captions on the YouTube2Text dataset: (a) complex examples where the multi-task model performs better than the baseline; (b) ambiguous examples (i.e., ground truth itself confusing) where multi-task model still correctly predicts one of the possible categories (c) complex examples where both models perform poorly.

Pasunuru & Bansal, ACL 2017, http://aclweb.org/anthology/P17-1117

Discourse-Based Objectives for Fast Unsupervised Sentence Representation Learning

If for sentence embedding information about neighboring sentences is useful, let's predict something about them:

- Binary Ordering of Sentences
- Next Sentence (classifier)
- Conjunction Prediction (predict a conjunction phrase if the second sentence starts from any)

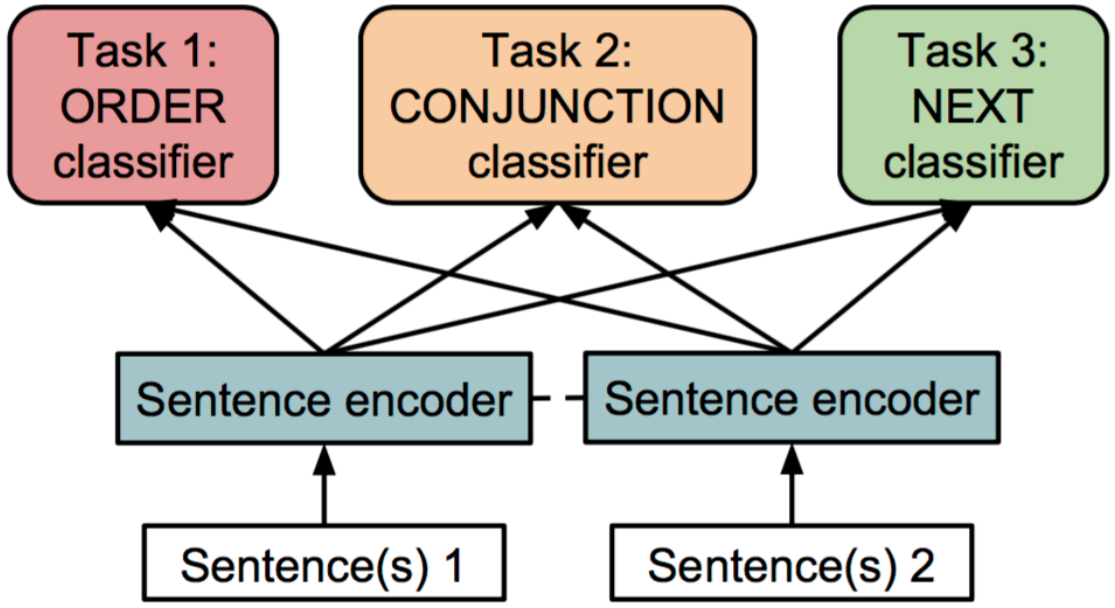


Jernite et al., 2017, https://arxiv.org/pdf/1705.00557.pdf

Discourse-Based Objectives for Fast Unsupervised Sentence Representation Learning

Sentence Pair	Label	Relation
A strong one at that. *Then I became a woman.*	Y	elaboration
I saw flowers on the ground. *I heard birds in the trees.*	N	list
It limped closer at a slow pace. *Soon it stopped in front of us.*	N	spatial
I kill Ben, you leave by yourself. *I kill your uncle, you join Ben.*	Y	time

Table 1: The binary ORDER objective. Discourse relation labels are provided for the reader, but are not available to the model.

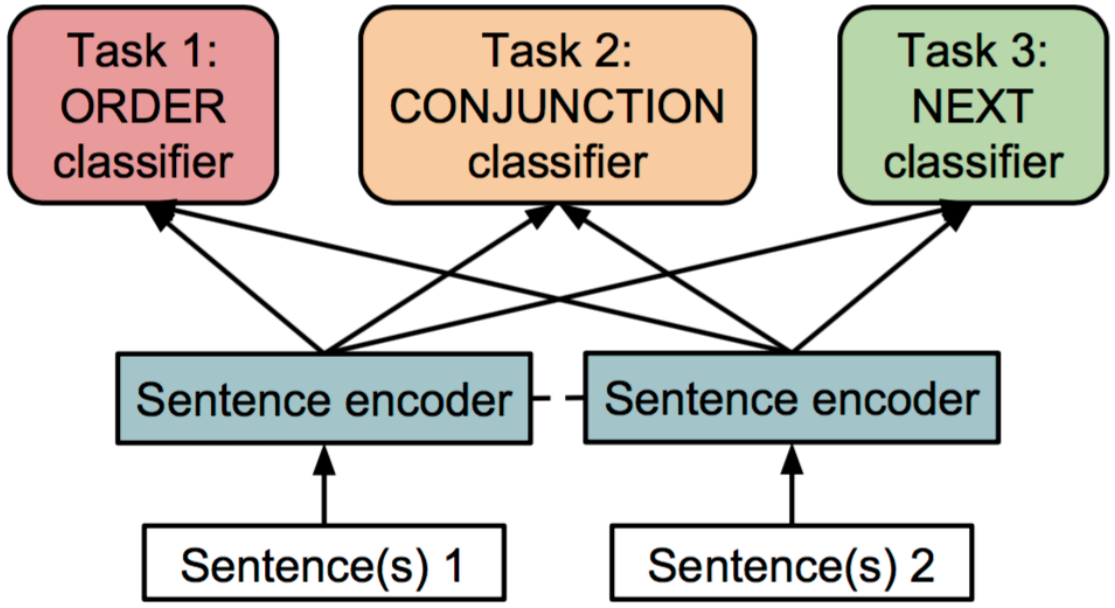


Jernite et al., 2017, https://arxiv.org/pdf/1705.00557.pdf

Discourse-Based Objectives for Fast Unsupervised Sentence Representation Learning

Context
No, not really. *I had some ideas, some plans.* *But I never even caught sight of them.*
Candidate Successors
1. *There's nothing I can do that compares that.* 2. *Then one day Mister Edwards saw me.* **3. *I drank and that was about all I did.*** 4. *And anyway, God's getting his revenge now.* 5. *He offered me a job and somewhere to sleep.*

Table 2: The NEXT objective.

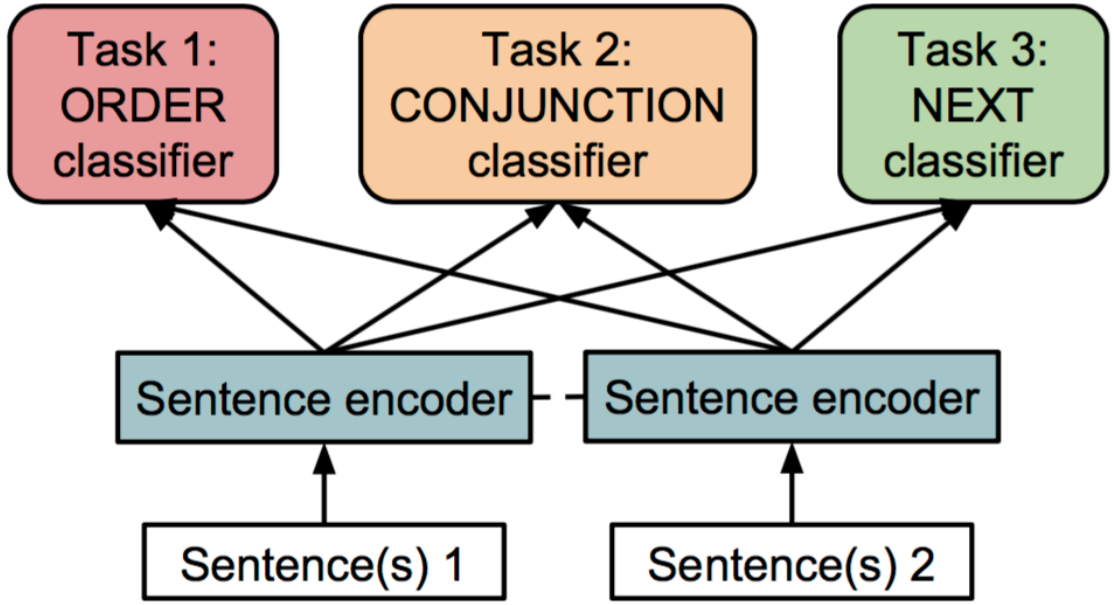


Jernite et al., 2017, https://arxiv.org/pdf/1705.00557.pdf

Discourse-Based Objectives for Fast Unsupervised Sentence Representation Learning

Sentence Pair	Label
He had a point. *For good measure, I pouted.*	RETURN *(Still)*
It doesn't hurt at all. *It's exhilarating.*	STRENGTHEN *(In fact)*
The waterwheel hammered on. *There was silence.*	CONTRAST *(Otherwise)*

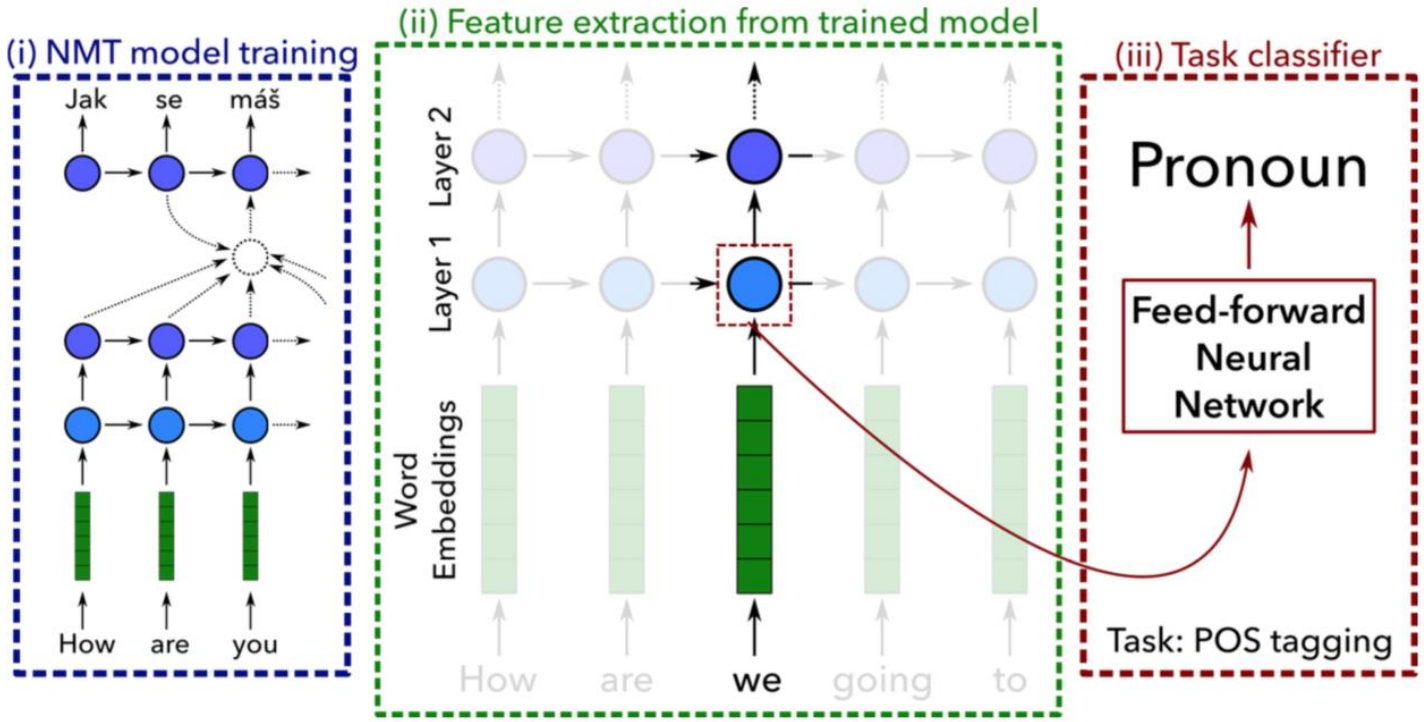
Table 3: The CONJUNCTION objective. Discourse relation labels are shown with the text from which they were derived.



Jernite et al., 2017, https://arxiv.org/pdf/1705.00557.pdf

A piece of understanding:
How much information about some task the model already has?

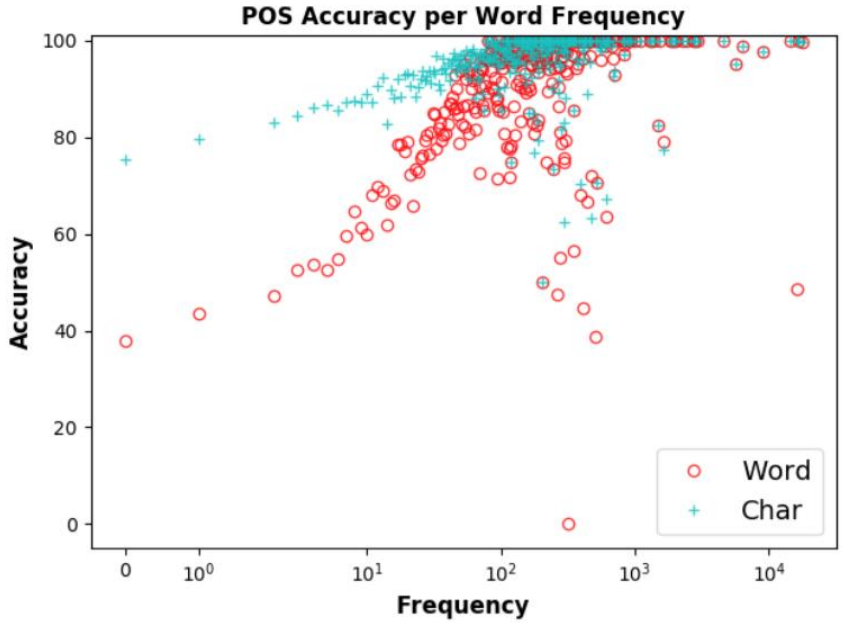
What do Neural Machine Translation Models Learn about Morphology?



- NMT system trained on parallel data
- features extracted from pre-trained model
- classifier trained using the extracted features

Belinkov et al, ACL 2017, http://aclweb.org/anthology/P17-1080

What do Neural Machine Translation Models Learn about Morphology?



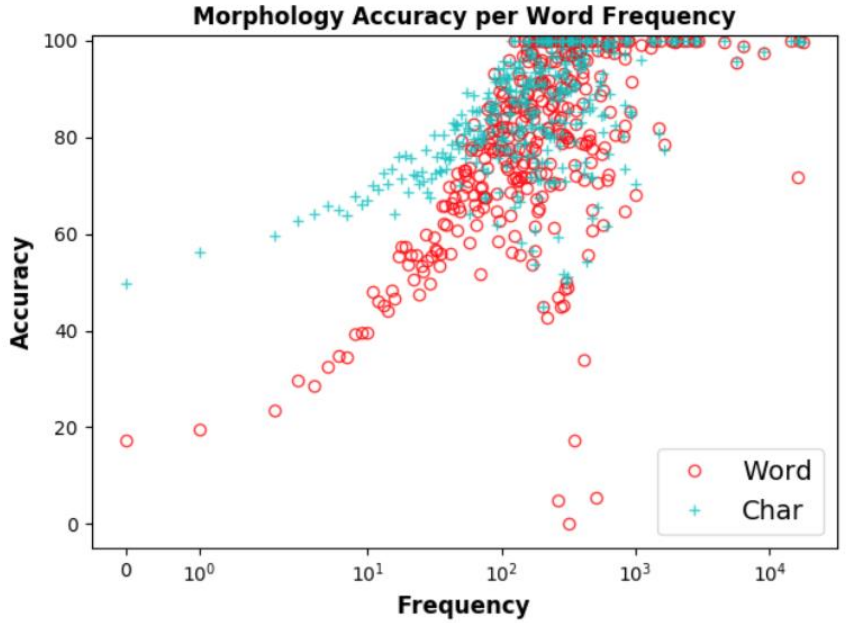


Figure 2: POS and morphological tagging accuracy of word-based and character-based models per word frequency in the training data. Best viewed in color.

Belinkov et al, ACL 2017, http://aclweb.org/anthology/P17-1080

What do Neural Machine Translation Models Learn about Morphology?

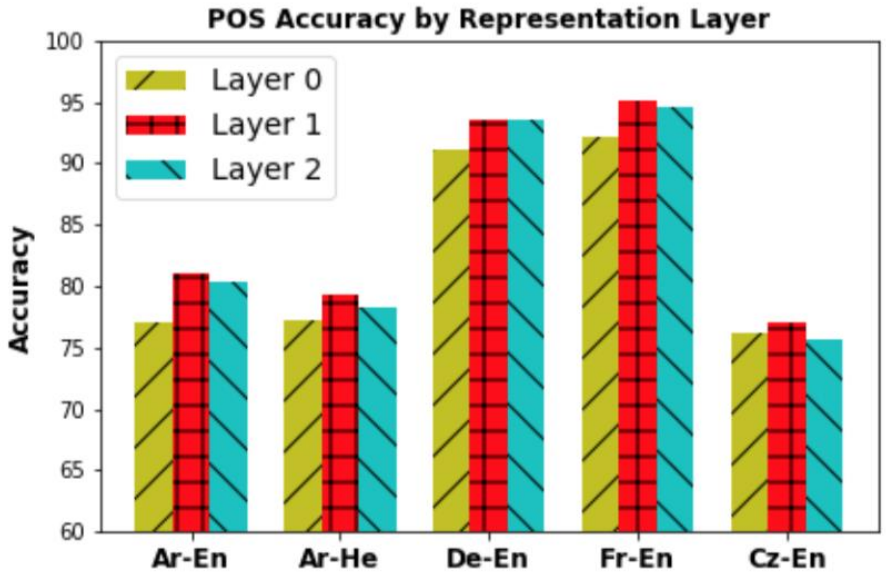


Figure 6: POS tagging accuracy using representations from layers 0 (word vectors), 1, and 2, taken from encoders of different language pairs.

What representations contain more information about POS tags?

Why is that?

Belinkov et al, ACL 2017, http://aclweb.org/anthology/P17-1080

What do Neural Machine Translation Models Learn about Morphology?

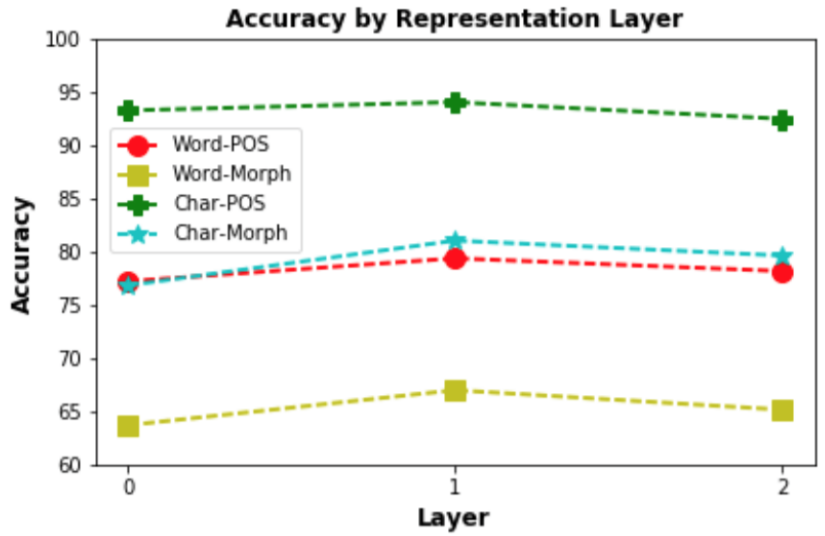


Figure 7: POS and morphological tagging accuracy across layers. Layer 0: word vectors or char-based representations before the encoder; layers 1 and 2: representations after the 1st and 2nd layers.

Representations at layer1 contain more information about morph tags than embeddings

Why is that?

Belinkov et al, ACL 2017, http://aclweb.org/anthology/P17-1080

What do Neural Machine Translation Models Learn about Morphology?

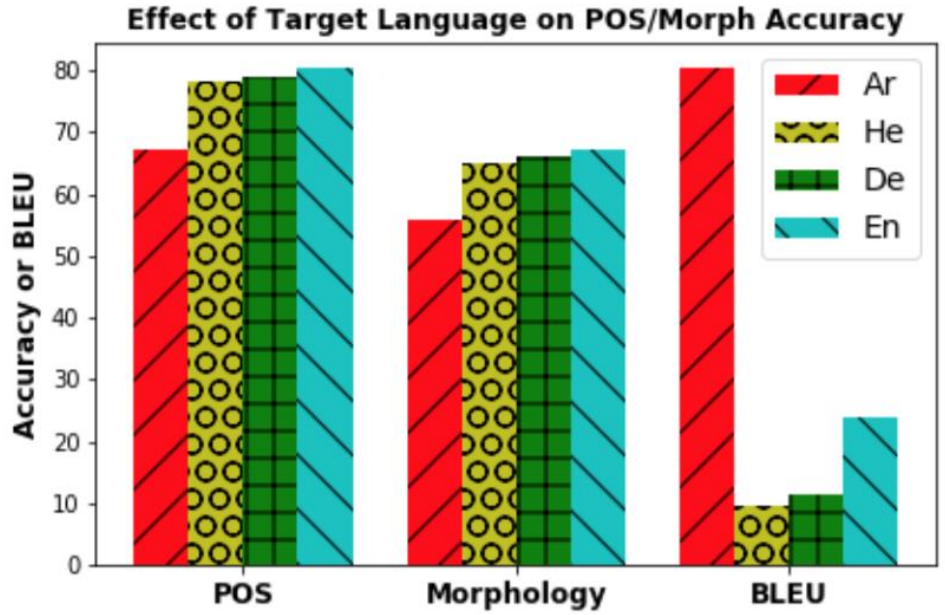


Figure 8: Effect of target language on representation quality of the Arabic source.

How does the target language affect the learned source language representations?

Does translating into a morphologically-rich language require more knowledge about source language morphology?

Belinkov et al, ACL 2017, http://aclweb.org/anthology/P17-1080

A piece of practice:
When do we need transfer or multi-task learning?

When do we need transfer or multi-task learning?

Not enough data

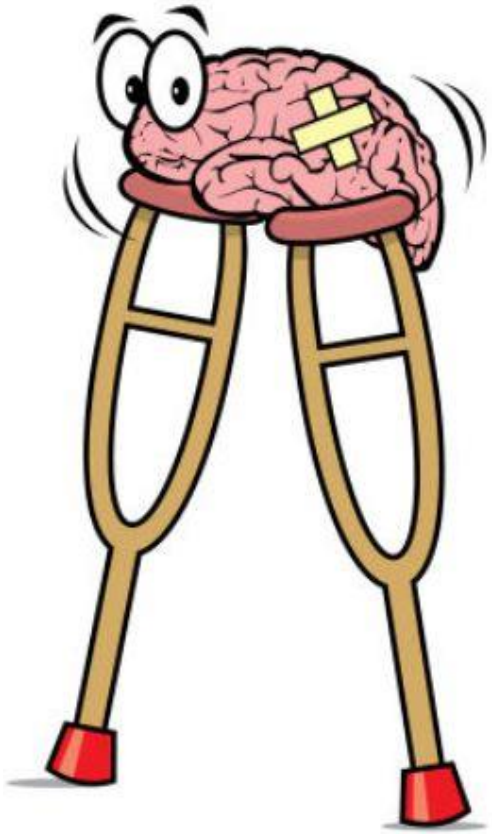
Hard to get a good model

Try transfer learning/multi-task learning to add some data and training signal

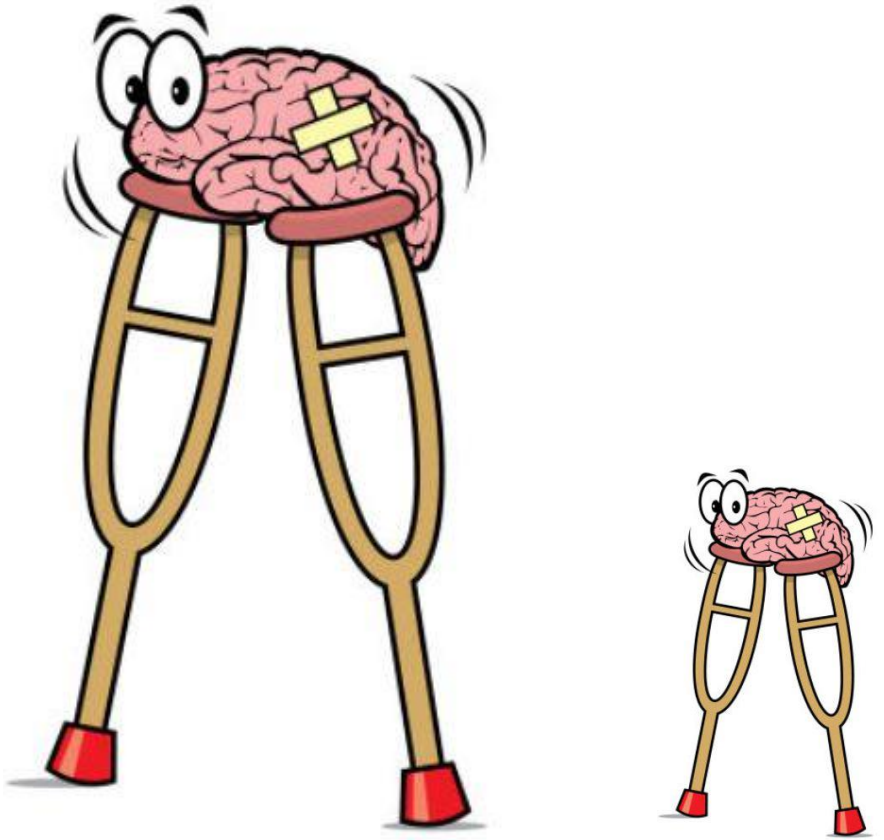
Plenty of data

Don’t worry, be happy

Hack of the day



Hack of the day



Batch size matters!

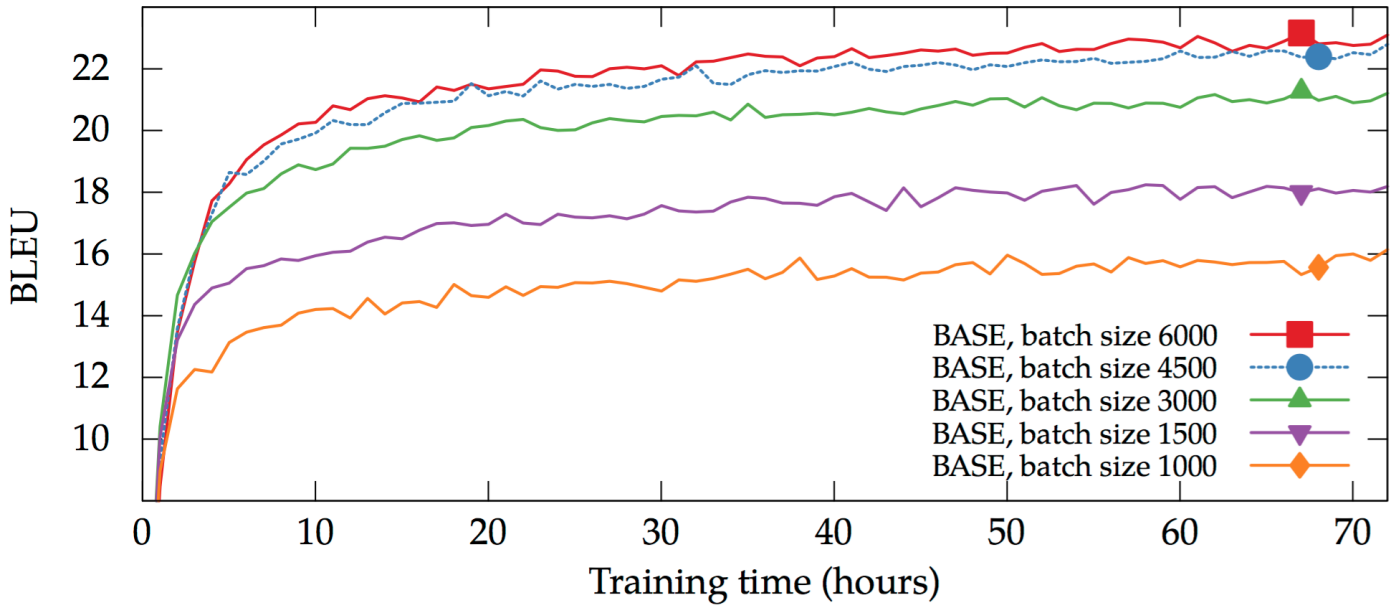
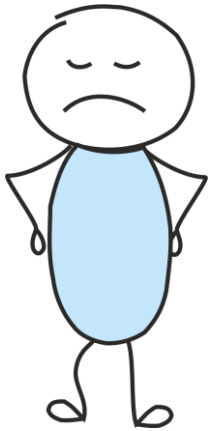


Figure 5: Effect of the batch size with the BASE model. All trained on a single GPU.

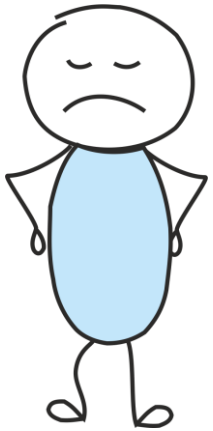
Popel & Bojar, 2018, https://arxiv.org/pdf/1804.00247.pdf

Notes from the life of the Transformer

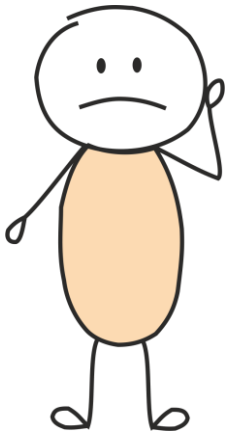


Batch size 1000

Notes from the life of the Transformer

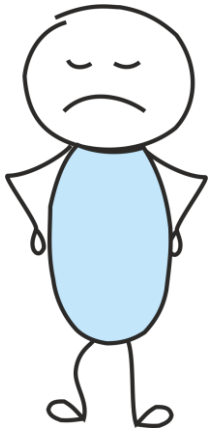


Batch size 1000

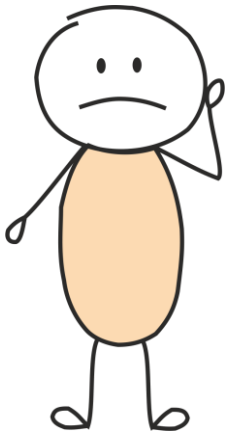


Batch size 3000

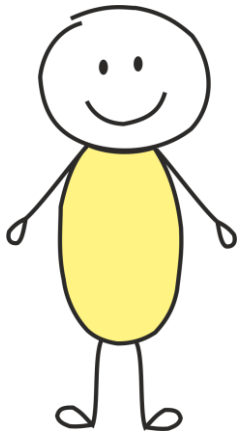
Notes from the life of the Transformer



Batch size 1000



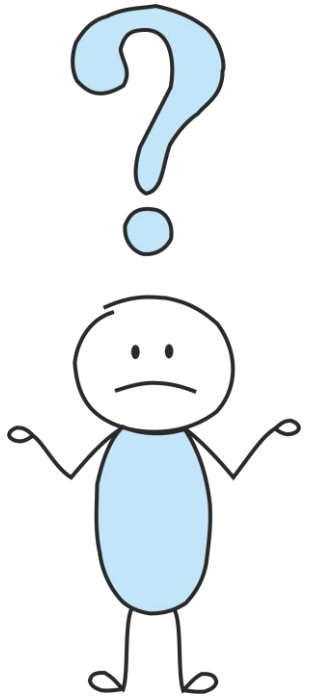
Batch size 3000



Batch size 6000

Hack of the day

How to use this to get profit?

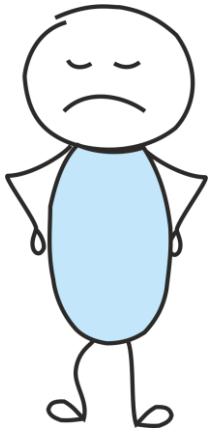


Hack of the day

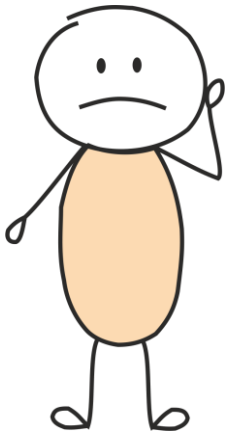
How to use this to get profit?

Virtual batch!

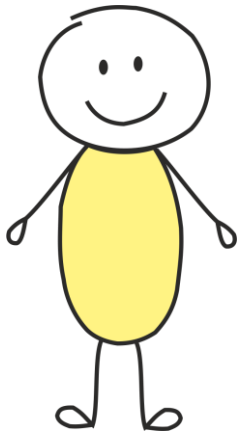
Notes from the life of the Transformer



Batch size 1000

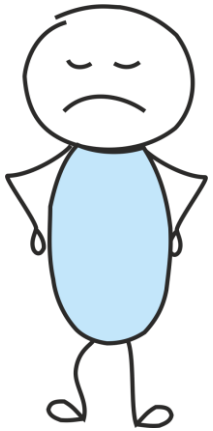


Batch size 3000

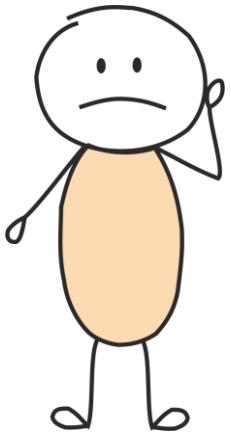


Batch size 6000

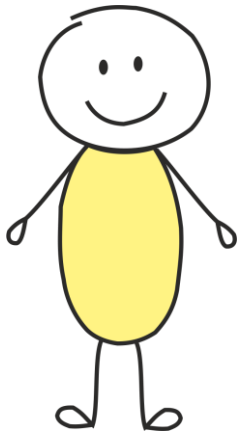
Notes from the life of the Transformer



Batch size 1000



Batch size 3000

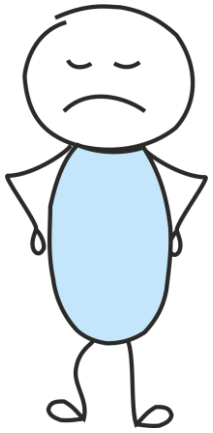


Batch size 6000

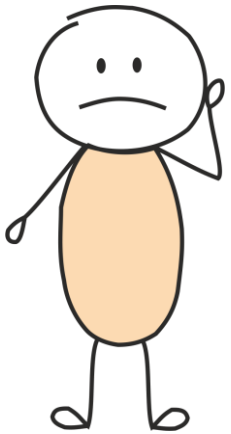
Virtual batch

(dark magic)

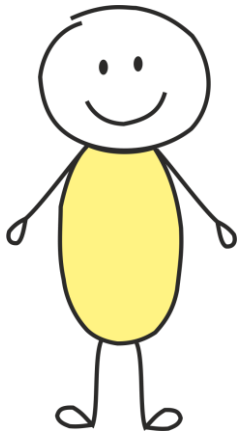
Notes from the life of the Transformer



Batch size 1000



Batch size 3000



Batch size 6000

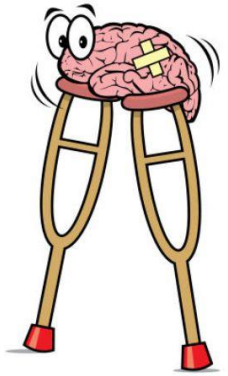
Virtual batch

(dark magic)

Aka ”грязные продакшн-штучки”

Mini-hack: fine-tuning

- Train your model on all the data/domains you have
- Fine-tune of the task/domain you really need



What happens when a researcher has nothing to do (or thinks that he has nothing to do)

Нейрогоголь

Однажды коллеги с ТВ-3 к предстоящей премьере «Гоголь. Страшная месть» решили написать рассказ, вдохновившись произведениями автора. Но чтобы слог рассказа походил на классика, нужно было привлечь технологии.

Специалисты из отдела машинного интеллекта и исследований Яндекса создали нейросеть, которую сначала обучили на более 30 тысячах книг русской прозы, а затем отдельно «дообучили» на Гоголе. В результате она научилась «творить» целые параграфы текста в стиле Николая Васильевича. Полученную нейросеть применили к написанной Сергеем Лукьяненко сюжетной канве будущего рассказа, которая благодаря этому обросла деталями и даже получила новые сюжетные ходы.

Читайте рассказ и смотрите в кино с 30 августа «Гоголь. Страшная месть»

Приводим текст рассказа в оригинале. Орфография и пунктуация нейросети сохранены.

Full text: https://www.kinopoisk.ru/action/neirogogol/

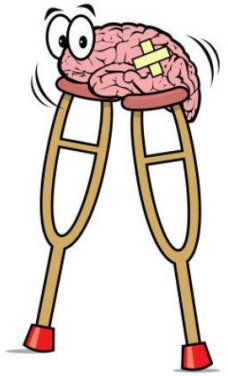
What happens when a researcher has nothing to do (or thinks that he has nothing to do)

Дурной договор

Жил в селе пекарь: возле церкви, близ церкви Иоанна-Галеевского (она была большая, церковь). Церковь деревянная, почерневшая, убранная зеленым мохом, с тремя конусообразными куполами, уныло стояла почти на краю села. Заметно было, что за время обмелели или, лучше сказать, совсем потухали последние ее свечи. свечи теплились пред темными образами. Свет от них освещал только иконостас и слегка середину церкви. Отдаленные углы притвора были закутаны мраком.

Харитон приходил в церковь и становился обыкновенно около дверей. Многие из них были заперты, прочие двери с замками и малеваниями поминутно открывались и показывали, что там никого нет. С другой стороны двери в двери глядели несколько бесцветных взглядов света, как будто бы в них не было огня.

Он вспоминал о красавице жене своей, хорошенькой, и к ней, казалось, уже чересчур привыкнуть нельзя. Ах, как было бы хорошо провести эту ночь вместе с нею. Но увы! Увы, покамест, вместо сего, грезит мысль о ней, как будто об одной только внезапной, временной помощи, и тайная тихая грусть подступает к нему. «Экая судьба!» — подумал про себя Харитон и, пришедши к себе в комнату одеться и лег в постель. Долго боролся он с бессонницею, наконец, пересилил ее. Опять какой-то сон, какой-то пошлый, гадкой сон. Боже, умилосердись: хотя на минуту, хотя на одну минуту покажи ее! Мою милую супругу, которую я еще недавно держал в своих объятьях. Но занемогла она родами дочери, и нет больше моей милой, без которой, может быть, и не будет никогда ничего. Сердце мое только так и ломается от горя.



Full text: https://www.kinopoisk.ru/action/neirogogol/

That’s all for today!

Always yours,
Yandex Research

That’s all for today!

In the next episode:

- Domain adaptation + a real-life story from MT team

Always yours,
Yandex Research

That’s all for today!

In the next episode:

- Domain adaptation + a real-life story from MT team

ТЕКСТ ПЕСНИ | Перевод

Hey Jude, don't make it bad.
Take a sad song and make it better.
Remember to let her into your heart,
Then you can start to make it better.

Hey Jude, don't be afraid.
You were made to go out and get her.
The minute you let her under your skin,
Then you begin to make it better.

Always yours,
Yandex Research

That’s all for today!

In the next episode:

- Domain adaptation + a real-life story from MT team

ТЕКСТ ПЕСНИ | Перевод

Hey Jude, don't make it bad.
Take a sad song and make it better.
Remember to let her into your heart,
Then you can start to make it better.

Hey Jude, don't be afraid.
You were made to go out and get her.
The minute you let her under your skin,
Then you begin to make it better.

ТЕКСТ ПЕСНИ | Перевод

Эй, Джуд, не делай все плохо.
Возьми грустную песню и сделай ее
лучше.
Не забывай впускать ее в свое сердце,
Тогда ты можешь начать делать это лучше.

Эй, Джуд, не бойся.
Тебя заставили выйти и забрать ее.
Как только ты подпустишь ее к себе,
Ты начнешь делать все лучше.

Always yours,
Yandex Research