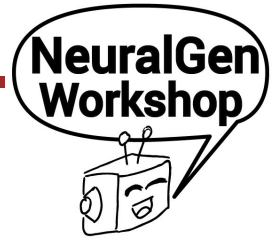


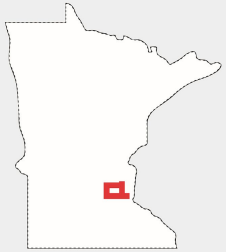
Pretraining for Generation

Alexander Rush

(Zack Ziegler, Luke Melas-Kyriazi, Sebastian Gehrmann)

HarvardNLP / Cornell Tech

NeuralGen Workshop



Minneapolis, MN

NAACL HLT 2019

June 2-7, 2019

Overview

- **Motivation**
- Current and Classical Approaches
- Models
- Experiments
- Challenges

Summarization

London, England (reuters) – Harry Potter star Daniel Radcliffe gains access to a reported $20 million fortune as he turns 18 on monday, but he insists the money won't cast a spell on him. Daniel Radcliffe as harry potter in "Harry Potter and the Order of the Phoenix" to the disappointment of gossip columnists around the world , the young actor says he has no plans to fritter his cash away on fast cars , drink and celebrity parties . " i do n't plan to be one of those people who , as soon as they turn 18 , suddenly buy themselves a massive sports car collection …

Harry Potter star Daniel Radcliffe gets $20m fortune as he turns 18 monday. Young actor says he has no plans to fritter his fortune away. ….

Common Summarization Mistakes

Mammoth wave of snow darkens the sky over everest basecamp. Appearing like a white mushroom cloud roaring, they scurry as their tents flap like feathers in the wind. Cursing and breathing heavily, they wait until the pounding is over.

Gehrmann et al. 2018

Problem

- How can we learn the general properties of long-form language (discourse, reference, etc.) from a specific NLG dataset (summary, data-to-text, image captioning, dialogue, etc.)?

Motivation Long-Form Generation: Lambada

They tuned, discussed for a moment, then struck up a lively jig. Everyone joined in, turning the courtyard into an even more chaotic scene, people now dancing in circles, swinging and spinning in circles, everyone making up their own dance steps. I felt my feet tapping, my body wanting to move. Aside from writing, I 've always loved dancing

Paperno et al. 2016

Lambada: Specialized Structure

LSTM	21.8
Hoang et al (2018)	59.2

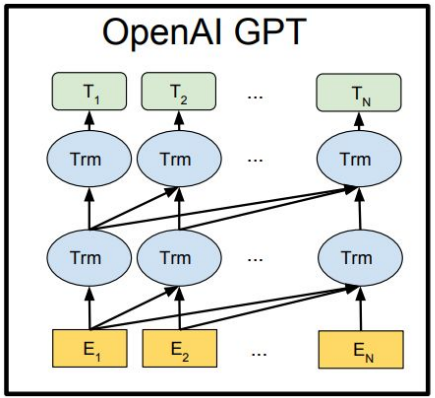
- Specialized attention-based model with kitchen-sink of entity tracking features and multi-task learning.

GPT-2: Impact of Model Scale

LSTM	21.8
Hoang et al (2018)	59.2
GPT-2 117M	45.9
GPT-2 345M	55.5
GPT-2 762M	60.1
GPT-2 1542M	**63.2**

Radford et al. 2019

This Talk: Conditional Generation with Pretraining



- Practical question: how can we use language models to improve the quality of conditional generation tasks?

Peters et al. 2018, Devlin et al. 2018, Radford et al. 2018

Overview

- Motivation
- **Current and Classical Approaches**
- Models
- Experiments
- Challenges

Notation: Conditional Generation

- Pretrained NN module 
- Rand. initialized NN module 
- Conditioning object $\mathbf{x}$
- Generated text $\mathbf{y}$

01101

CNN - On Monday, lead anchor …

The quick brown fox...

Notation: Using pretrained language model

$$p(y_t \mid y_{<t})$$

Pretrained Model

$$p(\boldsymbol{y} \mid \boldsymbol{x})$$

Conditional Model

$$p(\boldsymbol{x} \mid \boldsymbol{y})$$

Reverse Model

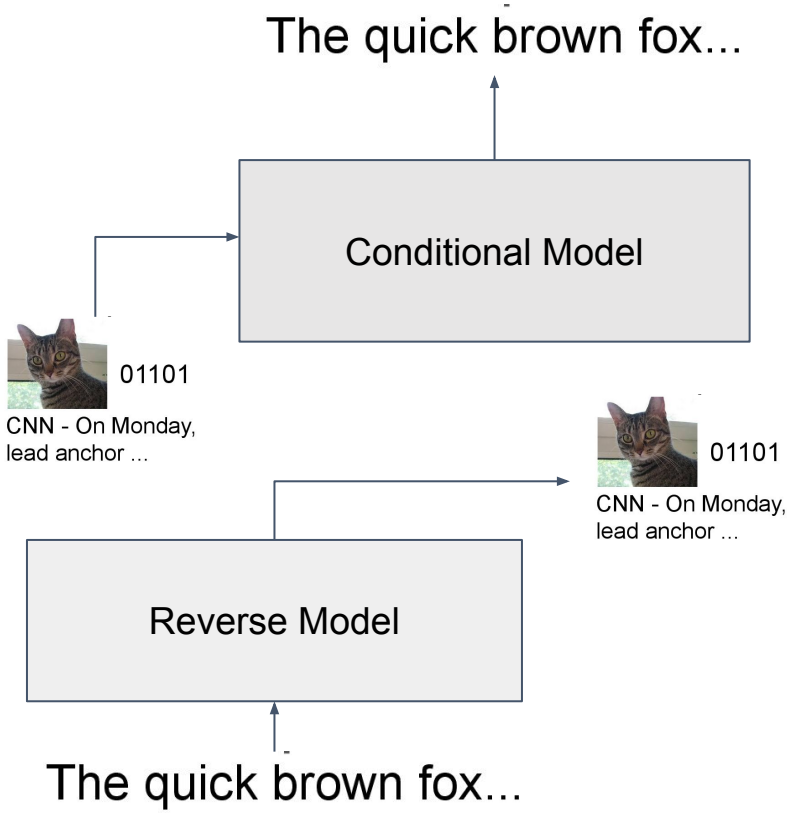
Approach 0: Backtranslation

- Incorporate additional data to approximate joint by heuristic alternating projection.

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x})$$

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}^*)$$

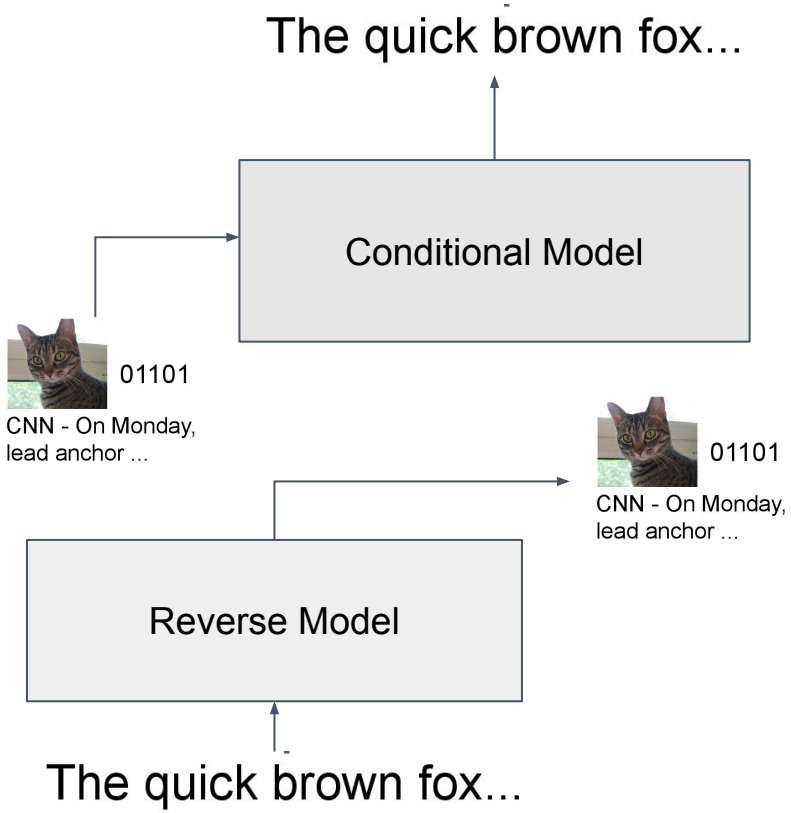
- Dominant approach in NMT. Does not require any pretraining.



Sennrich et al. 2015

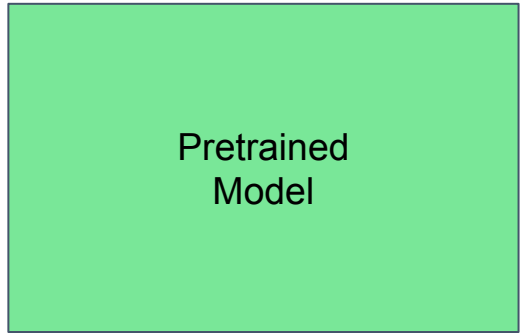
Backtranslation: Challenges

- Requires a reverse model for input modality.
- Requires access to the pretraining dataset.
- Computationally wasteful.

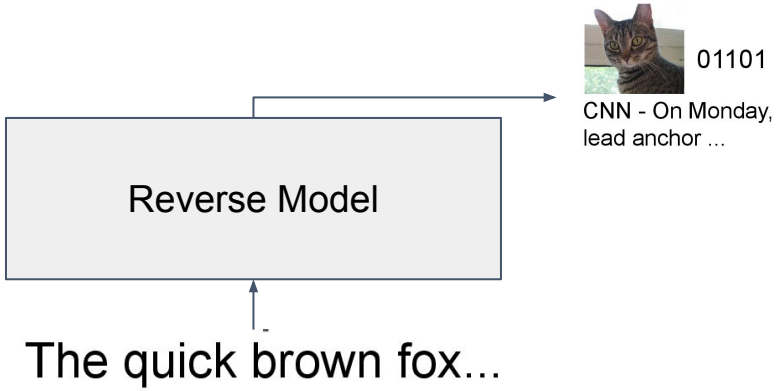


Approach 1: Noisy Channel / Bayes' Rule

$$p(\mathbf{y}|\mathbf{x}) \propto p(\mathbf{y}) \times p(\mathbf{x}|\mathbf{y})$$



- Dominant approach in statistical machine translation.
- Does not require conditional model.



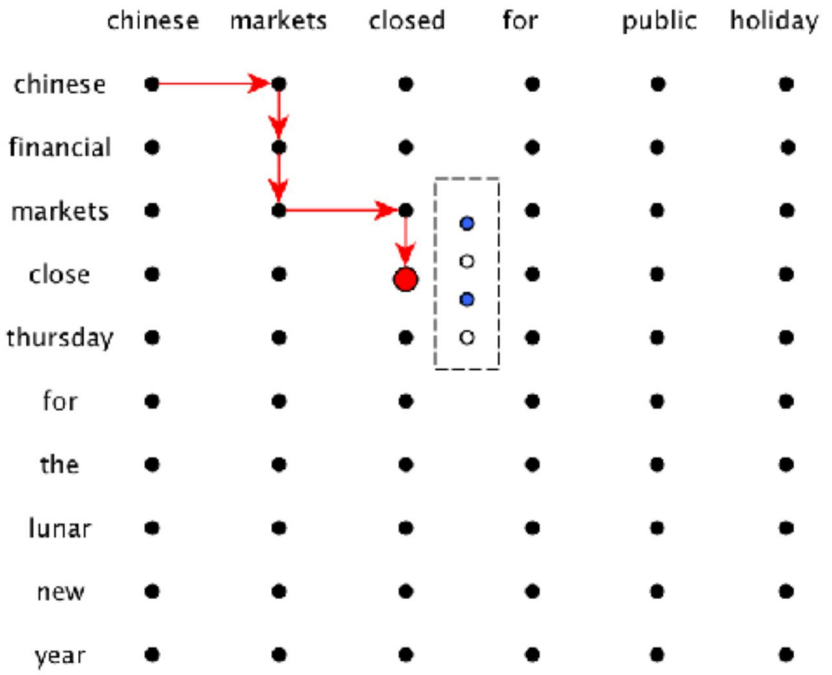
Yu et al. 2017

Neural Noisy Channel

$$p(\mathbf{y}|\mathbf{x}) \propto p(\mathbf{y}) \times p(\mathbf{x}|\mathbf{y})$$

$$\arg\max_{\mathbf{y}} p(\mathbf{y}) \times p(\mathbf{x}|\mathbf{y})$$

- Construct model to facilitate approximate inference.



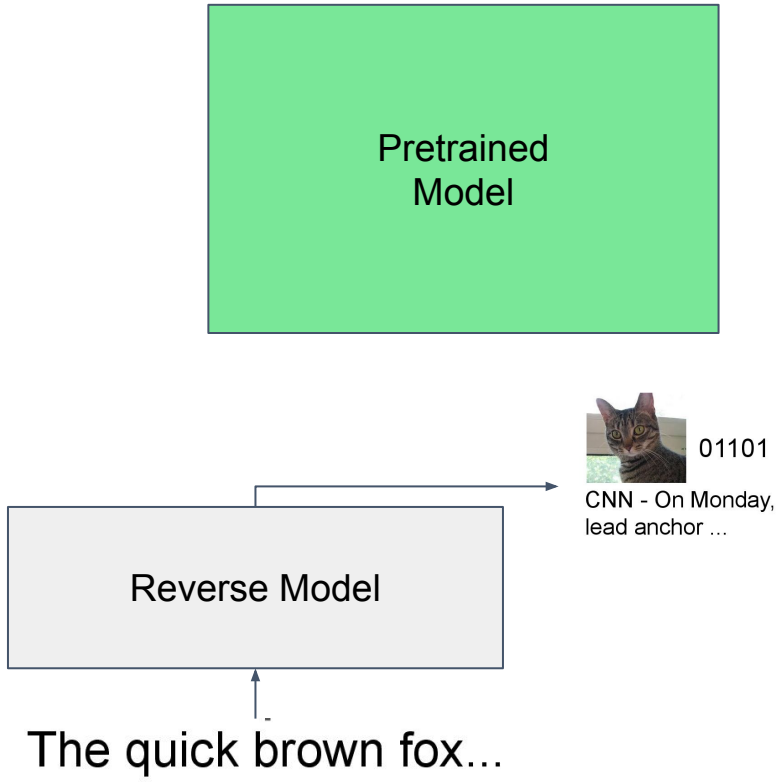
Yu et al. 2017

Noisy Channel: Challenges

- Requires generative model for input modality.
- Challenging MAP inference problem when using deep model.

$$\arg\max_{\mathbf{y}} p(\mathbf{y}) \times p(\mathbf{x}|\mathbf{y})$$

- Distributions often un-calibrated.



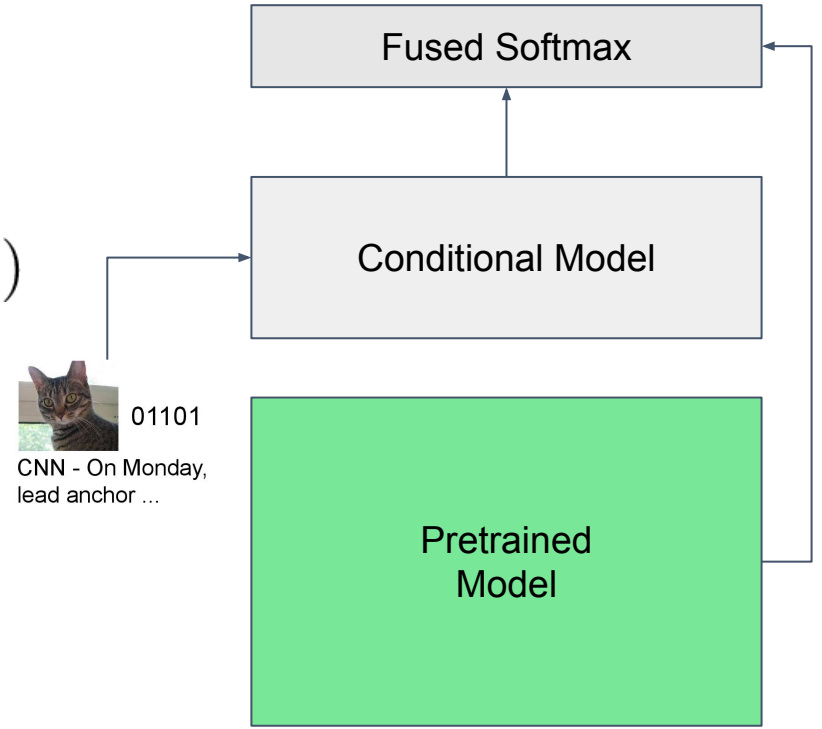
Yu et al. 2017

Approach 2: Simple Fusion

- Assume access to logit representation (pre-softmax).

$$p(y_t \mid y_{<t}, \mathbf{x}) = \text{softmax}(\text{MLP}(\alpha, \beta))$$

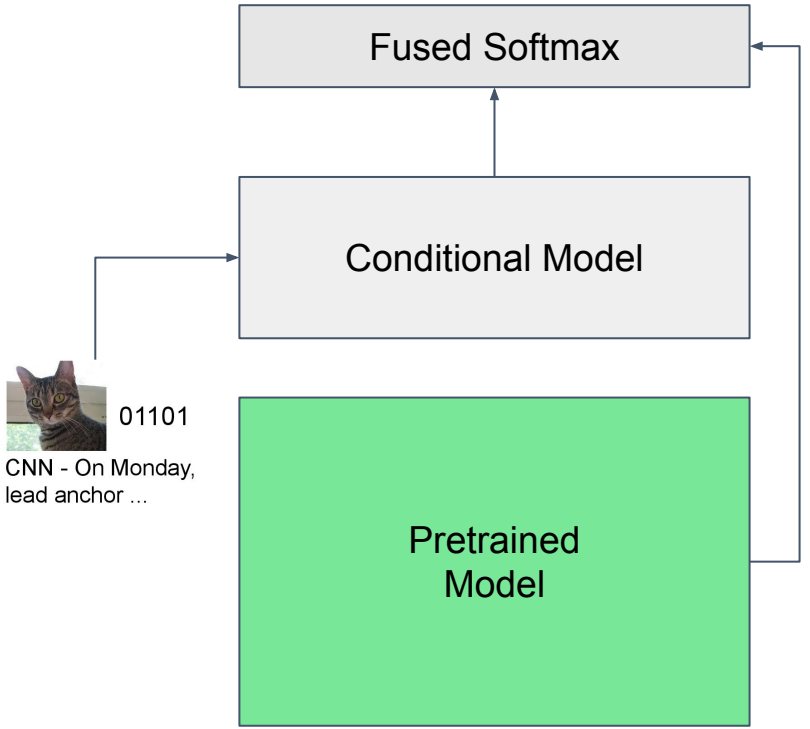
- Learn to smooth between conditional model and pretrained model.
- Several other variants: cold fusion, shallow fusion, deep fusion.



Gulcehre et al. 2015, Stahlberg et al. 2018

Fusion: Challenges

- Conditional model has no access to pretraining.
- Conditional model must relearn aspects of language generation already learned in the pretrained model.



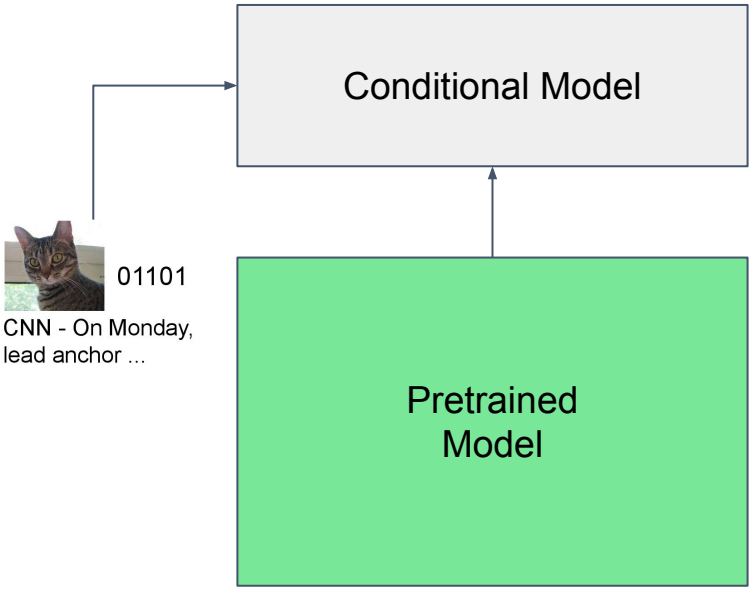
Gulcehre et al. 2015, Stahlberg et al. 2018

Approach 3: Representation Learning / Pretraining

- Utilize variable-length representation from model ("embeddings")

$$p(y_t \mid y_{<t}, \mathbf{x}) = \mathrm{softmax}(\mathrm{f}(\alpha_{1:t-1}))$$

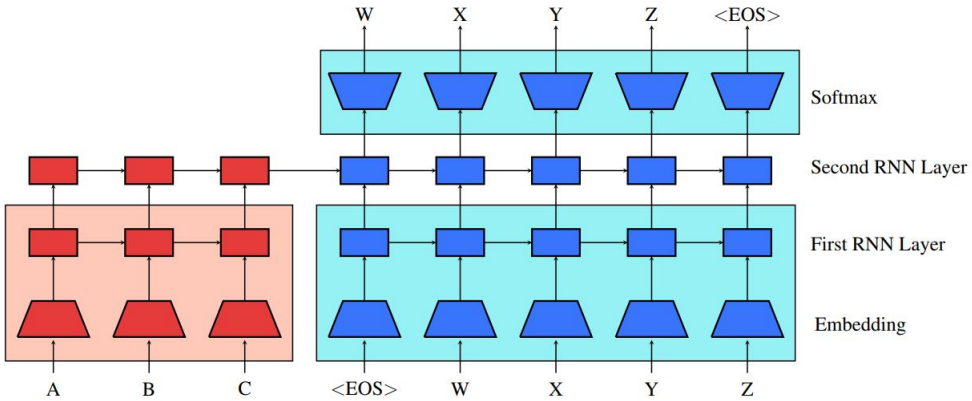
- Dominate approach in NLU applications (BERT/ELMo)

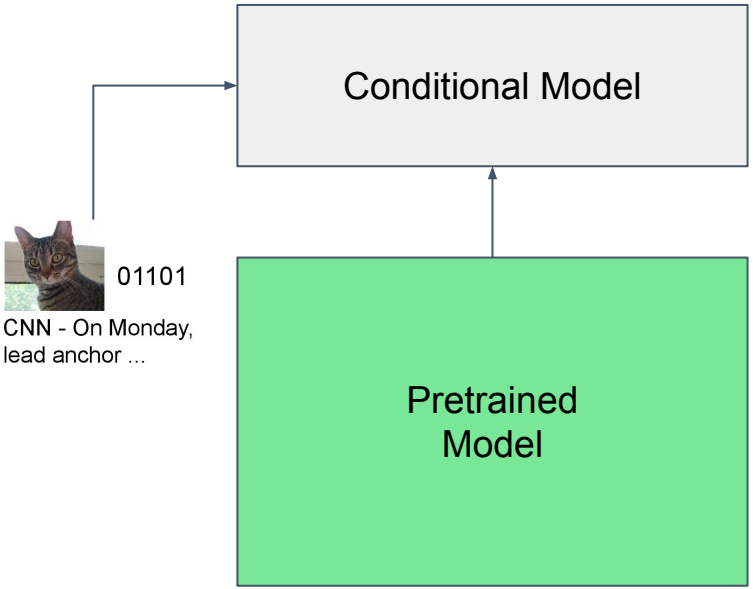


Ramachandran et al 2017, Edunov et al. 2019

Representation Learning: Challenges

- Empirically less effective than simpler fusion approaches.
- Little success (even with word embeddings) for conditional generation tasks.





Ramachandran et al 2017, Edunov et al. 2019

Lessons: Pretraining for Generation

- Simple fusion based approaches seem most robust.
- Approaches requiring reverse models seem intractable.
- Backtranslation likely infeasible for generation.
- Deep pretraining seems to be the most interesting, but ...

	160K	640K	5186K
baseline	21.4	33.1	40.1
SRC-ELMO	26.6	35.6	41.8
SRC-FT	24.3	34.9	40.8
TGT-ELMO	21.3	31.9	40.5
TGT-FT	24.2	31.4	38.8
SRC-ELMO+SHDEMB	29.0	36.2	41.8

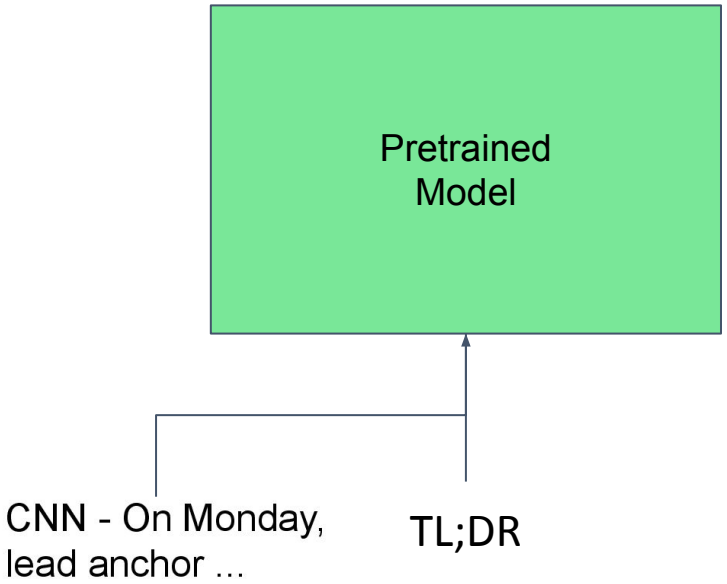
Edunov et al. 2019

Approach 4: Zero-Shot Generation

- Fake conditioning by prepending source with a special control word.

$$p(y_t \mid \boldsymbol{x} \odot y_{<t})$$

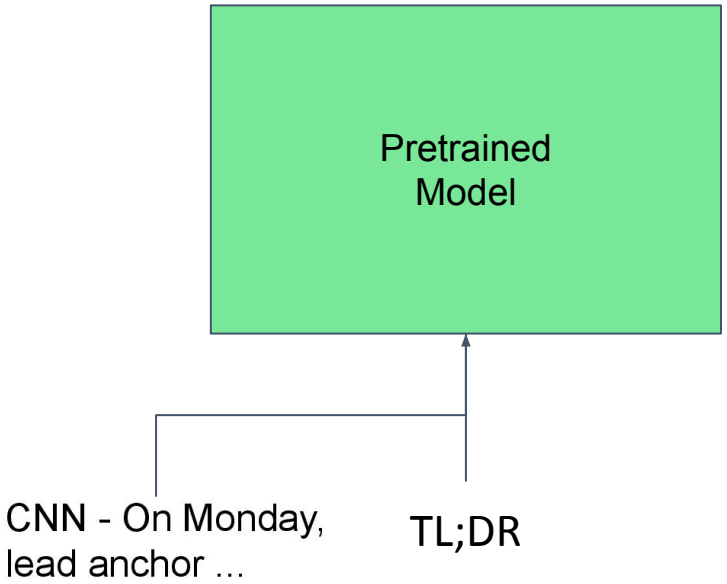
- Produces surprisingly good outputs for a simple trick.



Radford et al. 2019

Zero Shot: Challenges

- Only works with textual inputs.
- Requires a combinatorial search to find source.
- Seed word is problem specific.



Radford et al. 2019

Overview

- Motivation
- Current and Classical Approaches
- **Models**
- Experiments
- Challenges

Pretraining Models

Consider three different approaches to deep pretraining.

- Representation Learning: Repr-Transformer
- Combination through Context-Attn
- Pseudo Self Attention

Differ in usage of the source data.

Assumption: Self-attention Models

$$\mathrm{SA}(Y) = \mathrm{softmax}((YW_q)(YW_k)^\top)(YW_v)$$

$$p(y_t \mid y_{<t})$$

Pretrained self-attention model

Pretrained Model

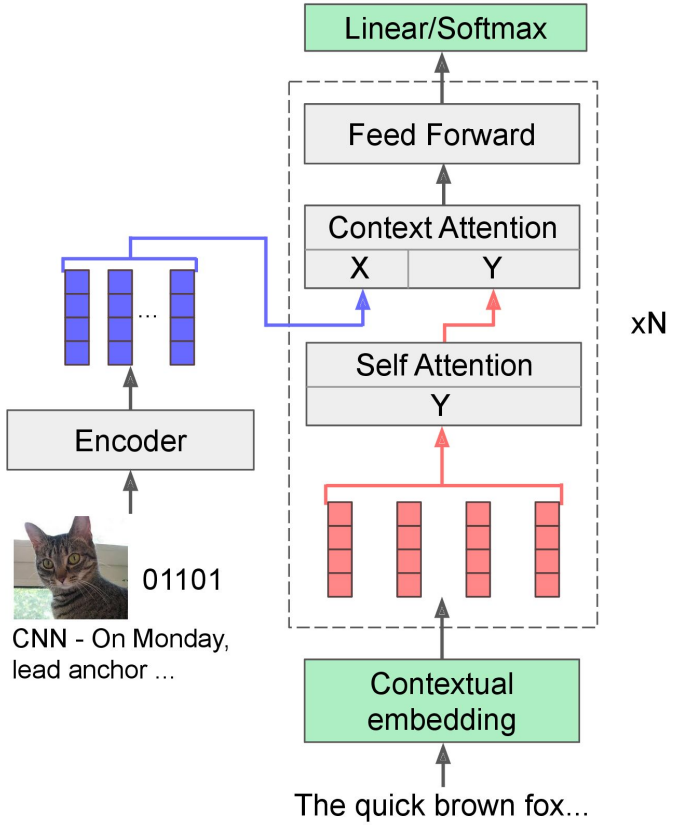
$$p(y_t \mid y_{<t}, \mathbf{x})$$

Extended transformer model

Conditional Model

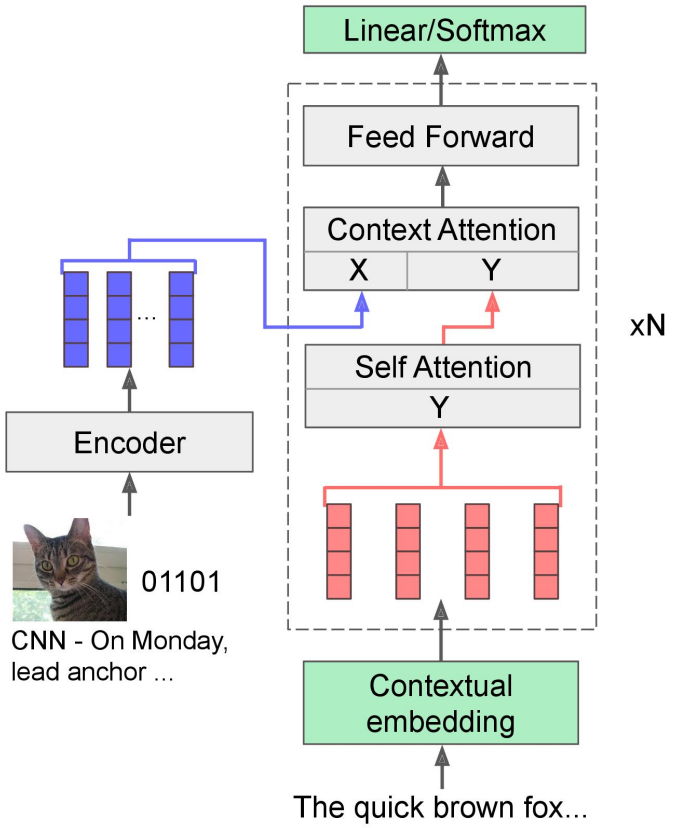
Representation Learning: Repr-Transformer

- Utilize pretraining to provide contextual embeddings to a conditional transformer.
- Transformer used as “conditional head” to the pretrained LM.



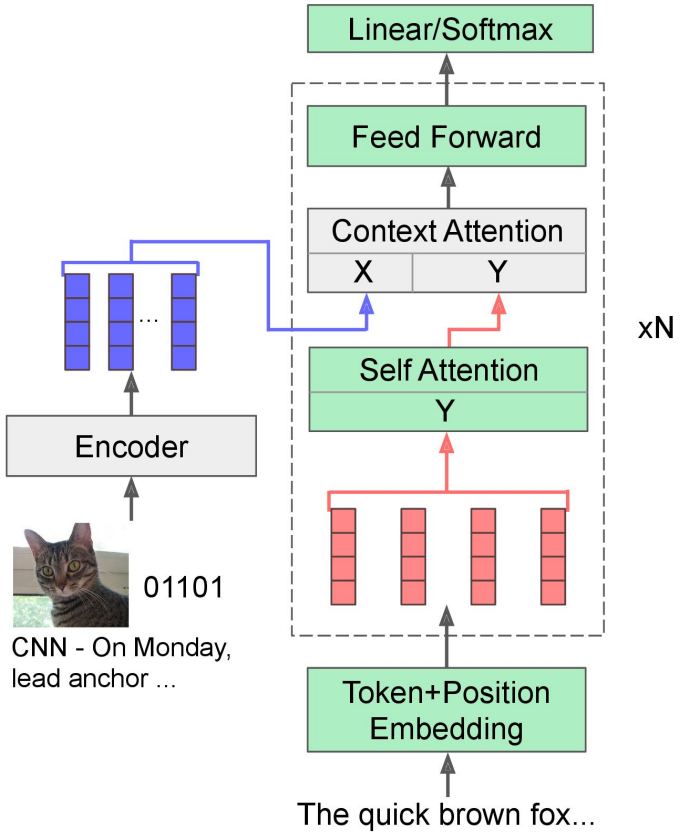
(Layer norm and residual connections omitted)

Intuition


Linear/Softmax
Feed Forward
Context Attention
X
Y
Self Attention
Y
xN
Encoder
01101
CNN - On Monday, lead anchor ...
Contextual embedding
The quick brown fox...

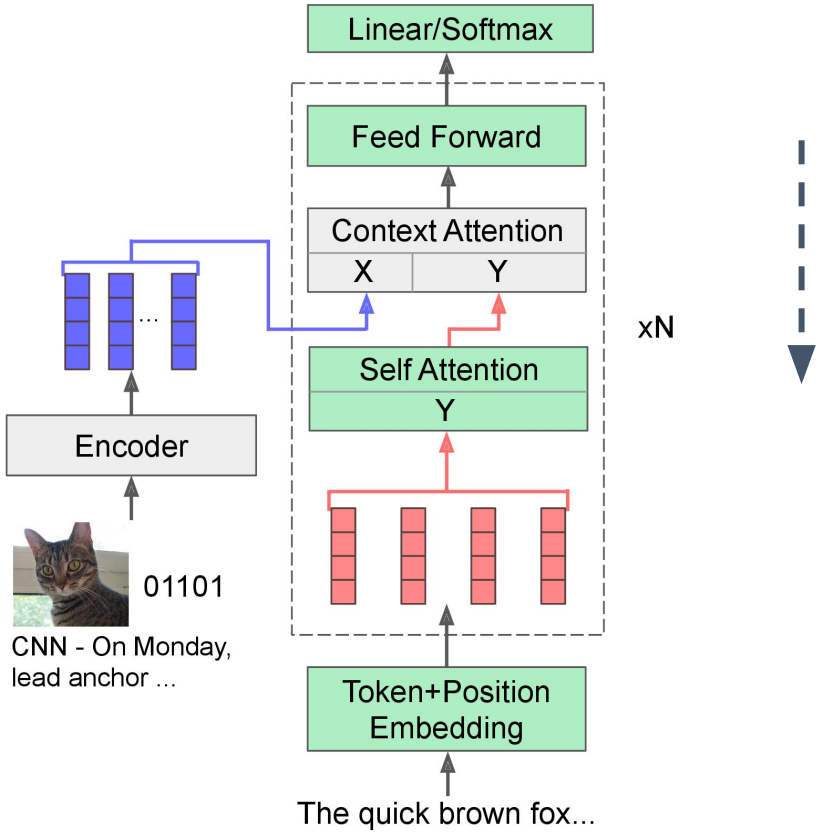
Context-Attn

- Assume that pretrained model has the same form as the head.
- Can initialize conditional transformer with self attention and feed forward layers.



(Layer norm and residual connections omitted)

Intuition

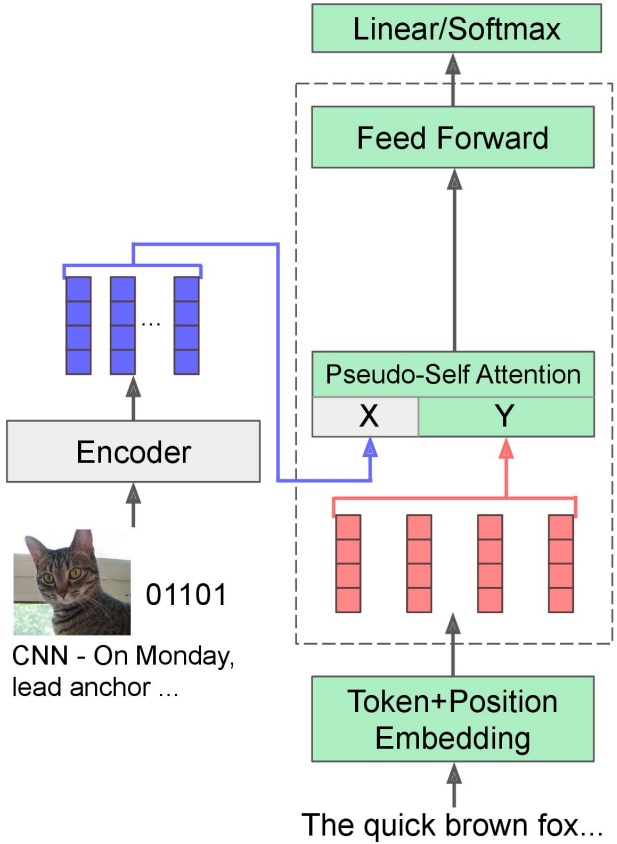

Linear/Softmax
Feed Forward
Context Attention
X
Y
Self Attention
Y
xN
Encoder
01101
CNN - On Monday, lead anchor ...
Token+Position Embedding
The quick brown fox...

Pseudo-Self Attention

- Train a model to inject conditioning directly into pretrained network.

$$\mathrm{PSA}(X, Y) = \mathrm{softmax}((YW_q)\begin{bmatrix} XU_k \\ YW_k \end{bmatrix}^\top)\begin{bmatrix} XU_v \\ YW_v \end{bmatrix}$$

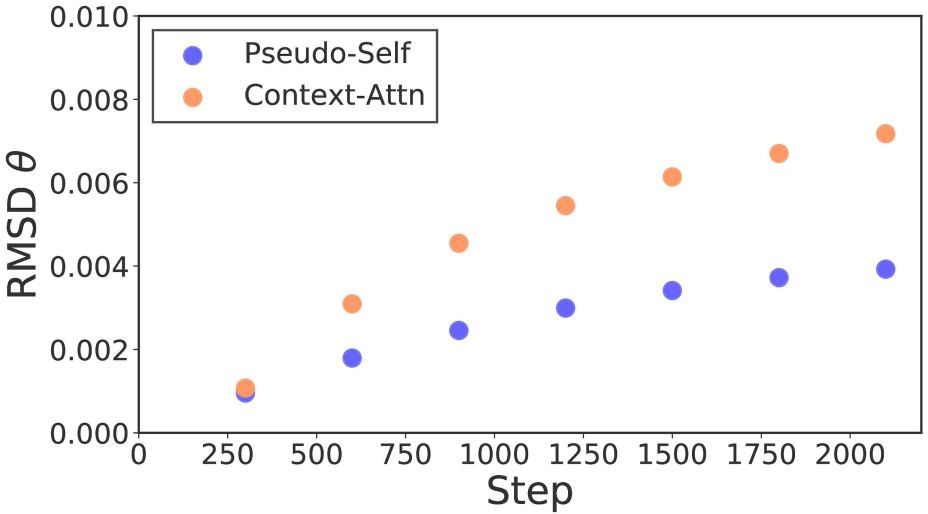
- Learn to project conditioning as additional attention keys.

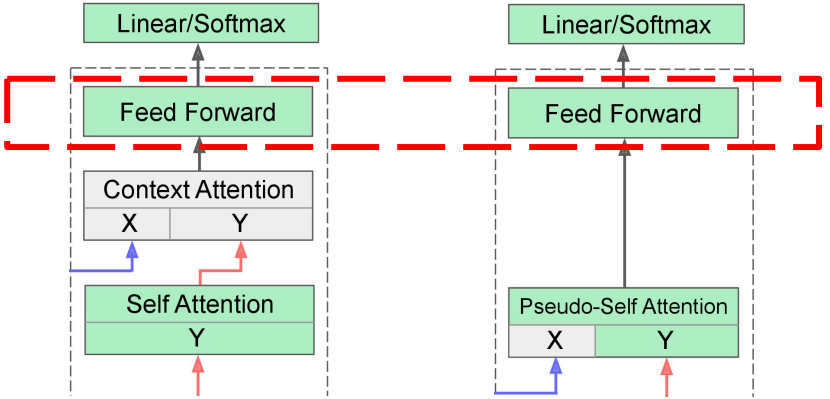


(Layer norm and residual connections omitted)

How do the methods differ?

- **Key Idea:** Train models to preserve as much of the original weight structure as possible.





Overview

- Motivation
- Current and Classical Approaches
- Models
- **Experiments**
- Challenges

Adaptive Conditional Generation Tasks

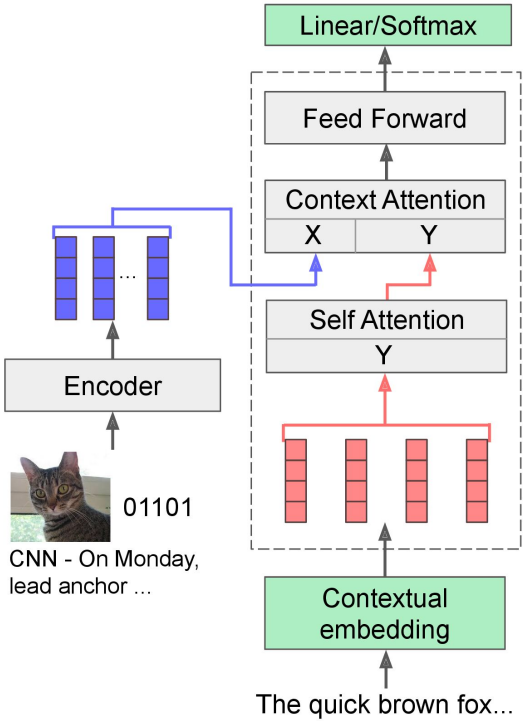
Conditional Generation Tasks

- Task 1: Class-Conditional Generation
- Task 2: Document Summarization
- Task 3: Story Generation
- Task 4: Image Paragraph Captioning

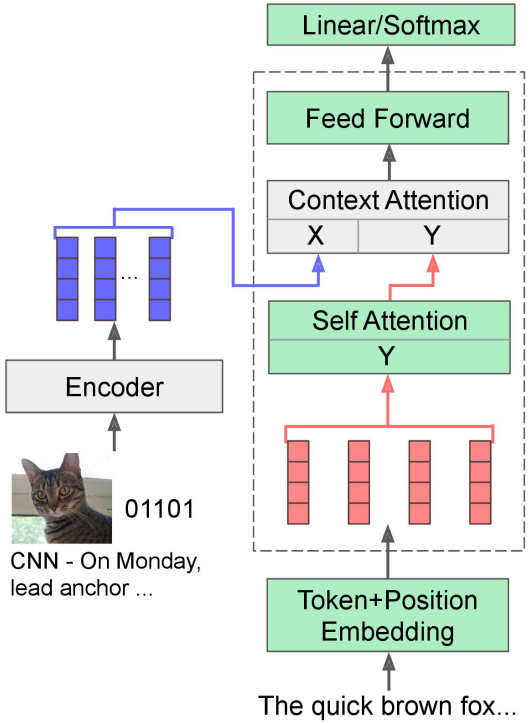
Metrics:

- Perplexity (general quality of the language)
- Task-Specific Quality

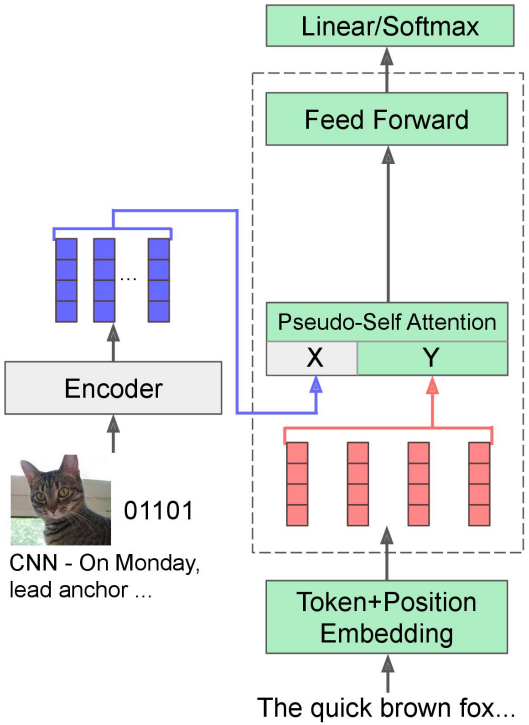
Deep Pretraining for Adaptation: Three Approaches



Repr-Trans



Context-Attn



Pseudo-Self

Task 1: Class-Conditional Generation (IMDB)

Positive movie review?

When I saw the preview of this film, I thought it was going to be a horrible movie. I was wrong. The film has some of the funniest and most escapist scenes I've seen in a long time. The acting is superb. The story is decent, but the direction and editing may have been a bit harsh at times.

Model	PPL ↓	Cls Acc ↑
Test set	-	90.1
GPT-2	41.21	-
Simple Fusion	38.31	65.1
Transformer	105.43	**92.7**
Repr-Trans	39.69	72.7
Context-Attn	40.74	88.8
Pseudo-Self	**34.80**	92.3

~10 million training tokens (tgt)

Task 2: Document Summarization (CNN/DM)

London, England (reuters) – Harry Potter star Daniel Radcliffe gains access to a reported $20 million fortune as he turns 18 on monday, but he insists the money won't cast a spell on him. Daniel Radcliffe as harry potter in "Harry Potter and the Order of the Phoenix" to the disappointment of gossip columnists around the world , the young actor says he has no plans to fritter his cash away on fast cars , drink and celebrity parties . " i do n't plan to be one of those people who , as soon as they turn 18 , suddenly buy themselves a massive sports car collection …

Harry Potter star Daniel Radcliffe gets $20m fortune as he turns 18 monday. Young actor says he has no plans to fritter his fortune away.

Model	R1 ↑ / R2 ↑ / RL ↑	PPL ↓
PGenerator+BU	41.22 / 18.68 / 38.34	-
ELMo+SHDEMB†	41.56 / 18.94 / **38.47**	-
BERT+Two-Stage†	41.38 / **19.34** / 38.37	-
CopyTransformer	39.94 / 17.73 / 37.09	8.21
Repr-Trans	37.09 / 13.77 / 33.99	13.58
Context-Attn	40.59 / 18.17 / 37.24	6.68
Pseudo-Self	40.72 / 18.38 / 37.46	**6.43**
Pseudo-Self+BU	**41.62** / 18.66 / 38.46	**6.43**

~30 million training tokens (tgt)

Task 2b: TL;DR Summarization

not necessarily my lucky day , but some kids this is how it went was sitting out on the dock at a local lake with a friend sharing some beers . little boy aged 2-3 yrs old walks up with a wooden stick and starts poking at the water . it was windy out and the dock was moving , and sure enough the kid leans over just enough to topple head first into the water . i had already pulled my phone out and wallet out just in case i was to accidentally fall in so i went straight over and hopped in . saw his little hand reaching up and tossed him straight back onto the dock . walked him to his dad who didn ' t speak any english and was very confused why i had his son soaking wet . left later that day and saw the kid back on the dock ! it blew my mind.:

TL;DR saved a 2 year old from drowning at a lake because i was drinking beers with a friend .

- First-place system uses pretrained conditional generation.

Task 3: Story Generation (WritingPrompt)

A portal to a fantasy-like land opens in the middle of New York City and exiles start coming through .

Tannen blinked . Nothingness greeted him ; he was still dreaming of the massive portal before him . How long had it been ? Would it be ... ? How long had it been since he saw it ? That was impossible , and yet , how did it end ? ...

Model	PPL ↓	Rank Acc. ↑
Transformer	29.80	**80.6**
Repr-Trans	**21.16**	77.8
Context-Attn	N/A*	9.3
Pseudo-Self	21.21	80.3

Fan et al. 2018

Task 3: Story Generation (WritingPrompt)

A portal to a fantasy-like land opens in the middle of New York City and exiles start coming through .

Tannen blinked . Nothingness greeted him ; he was still dreaming of the massive portal before him . How long had it been ? Would it be ... ? How long had it been since he saw it ? That was impossible , and yet , how did it end ? ...

Model	Grammaticality	Non-redundancy	Consistency	Typicality
Test set	71.3 ± 4.3	87.2 ± 3.2	85.1 ± 3.4	74.4 ± 4.1
Transformer	55.4 ± 4.7	60.5 ± 4.6	53.7 ± 4.7	39.7 ± 4.6
Repr-Trans	**62.1** ± 4.4	**71.0** ± 4.1	**57.1** ± 4.5	**43.7** ± 4.5
Pseudo-Self	**65.2** ± 4.6	**69.3** ± 4.5	**61.3** ± 4.7	**48.4** ± 4.8

~300 million training tokens (tgt)

Task 4: Image Paragraph Captioning

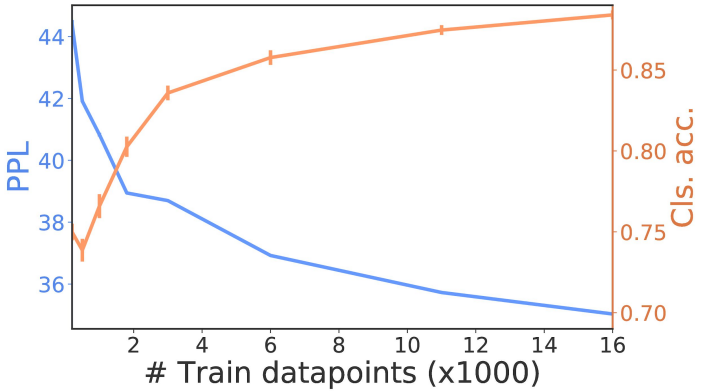
Two people are sitting on a bench. The elephant is sitting on the dirt. The man is sitting on top of the elephant. The woman is wearing a white shirt. The man is wearing a black shirt. There is a tree behind the elephant. There are trees on the ground. There are trees in the background.

Model	CIDEr ↑	B4 ↑
LSTM Baseline	11.1	7.3
Krause et al.	13.5	8.7
Chatterjee et al.	20.9	**9.4**
Melas-Kyriazi et al.	22.7	8.7
Transformer, Repr-Trans	19.3	7.2
Transformer, Context-Attn	22.6	7.6
Transformer, Pseudo-Self	**24.0**	8.3

(All results use cross-entropy. Reinforcement Learning approaches perform better on this task.)

<1 million training tokens (tgt)

Adapting in Low-Data Settings



Pretraining (1.8K)

I fell in love with this film in 1985. It's a quintessential short film that explores the everyday lives of the human condition. The main character of the movie is a man named Donald (Husband George). He buys a home and captures a great deal of information about the businessmen who live and work in his neighborhood.

No Pretraining (1.8K)

"Set's that I liked this movie. I have seen I remember the original movie is one of the music that it is great movie. I've seen this film and one of the whole movie is like this movie. It is so bad, I watched the top of this movie. i would see the movie was bad, I have seen it. This movie, it's a TV main movie is about the plot, relaxing. I

Bigger Models?

- All experiments run with smallest available GPT-2 (117M)
- Bigger model recently released at 345M.

Model	PPL ↓	Cls Acc ↑
Pseudo-Self 117M	34.80	92.3
Pseudo-Self 345M	30.26	92.4

Concurrent Work

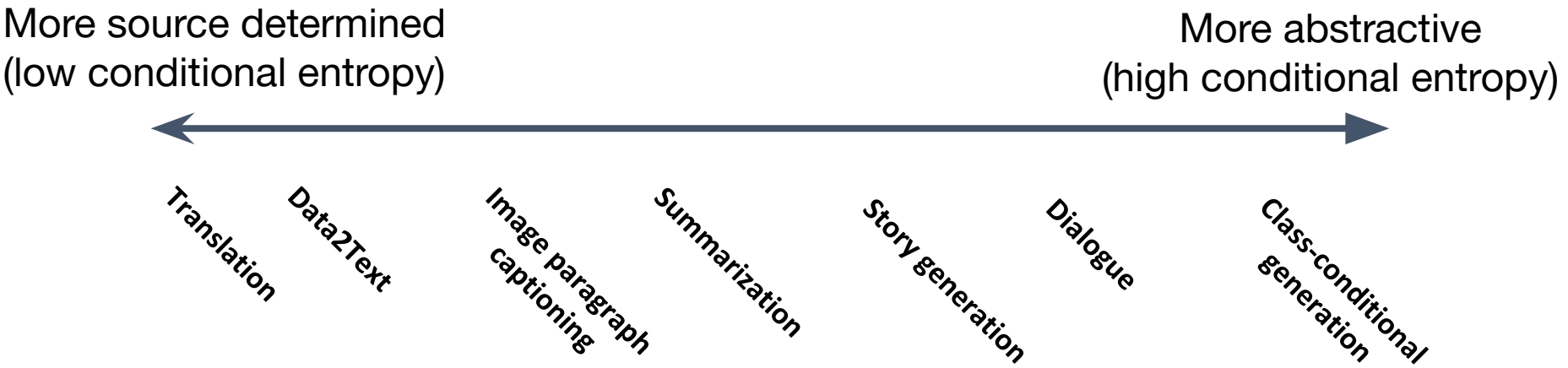
- Large-Scale Transfer Learning for Natural Language Generation - Golovanov et al 2019.

- Use roughly the same model for dialogue tasks.

Overview

- Motivation
- Current and Classical Approaches
- Models
- Experiments
- **Future Challenges**

Open Questions



- Pseudo-Self approach well suited for open-ended conditional generation.
- Application to low conditional entropy tasks?

Conclusions

- *Pseudo self attention* for general conditional generation with pretrained LMs
- Strong automatic and human eval results across diverse long-form conditional generation tasks
- Application to low conditional entropy tasks? Connection with source-side pretraining?