Out-of-distribution detection and selective generation for conditional language models

Jie Ren
Research Scientist
Google Research, Brain team 

Joint work with Peter J. Liu*, Jiaming Luo, Yao Zhao,
Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan

*Correspondence to: {jjren, peterjliu}@google.com

Google Research

Sequence-to-sequence – Success guaranteed?

Given data distribution, (x,y) ~ D, and sufficiently powerful conditional language model architecture

1. Collect enough *Train* ~ D.
2. Train/Fine-tune, probably a Transformer.
3. Do well on *Test* ~ D.

Example 1 (Summarization)
x = document
y = summary
D = XSum summarization dataset

Example 2 (Translation)
x = French sentence
y = English sentence
D = WMT15

Google Research

Out-of-distribution: x ~ O != D

In general language models can say **arbitrary things**. They say **worse** things **out-of-distribution**.

Example BBC news (Xsum) summarization model given a DailyMail input:

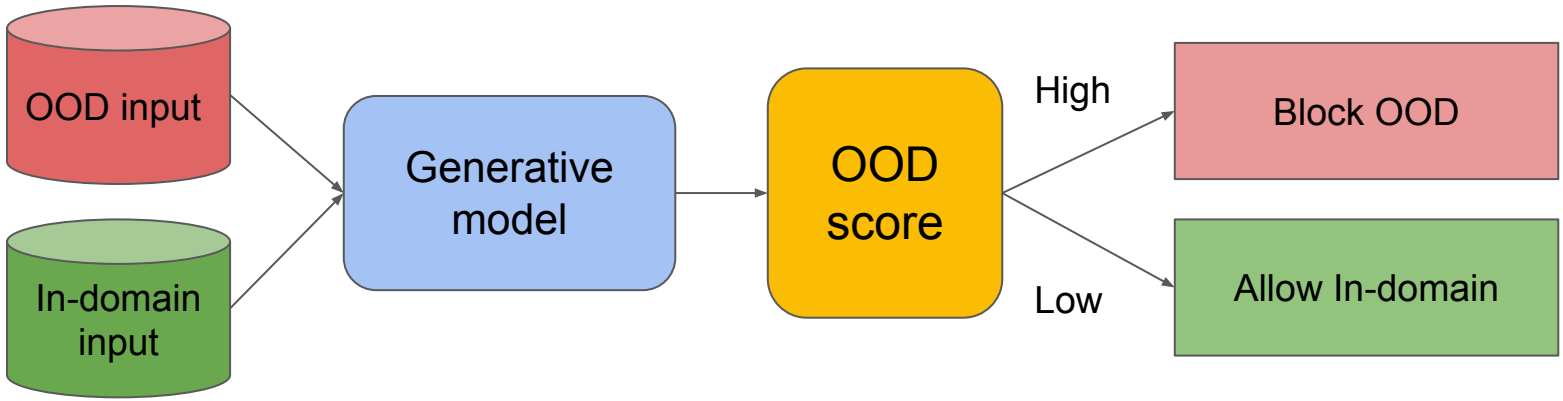
(DailyMail article) A man trying to elude police jumped into a Missouri creek overnight wearing only his under-wear – but his daring gambit did not pay off. Responding officers and firefighters followed the fugitiveinto the murky waters of Brush Creek in Kansas City and fished him out early Friday morning. The 38-year-old suspect has been taken to an area hospital to be treated for injuries to his arm and leg. He mayface charges in connection to a hit-and-run crash. *[truncated]*

(Model summary)
All images are copyrighted.

Google Research

Two defensive strategies: OOD detection

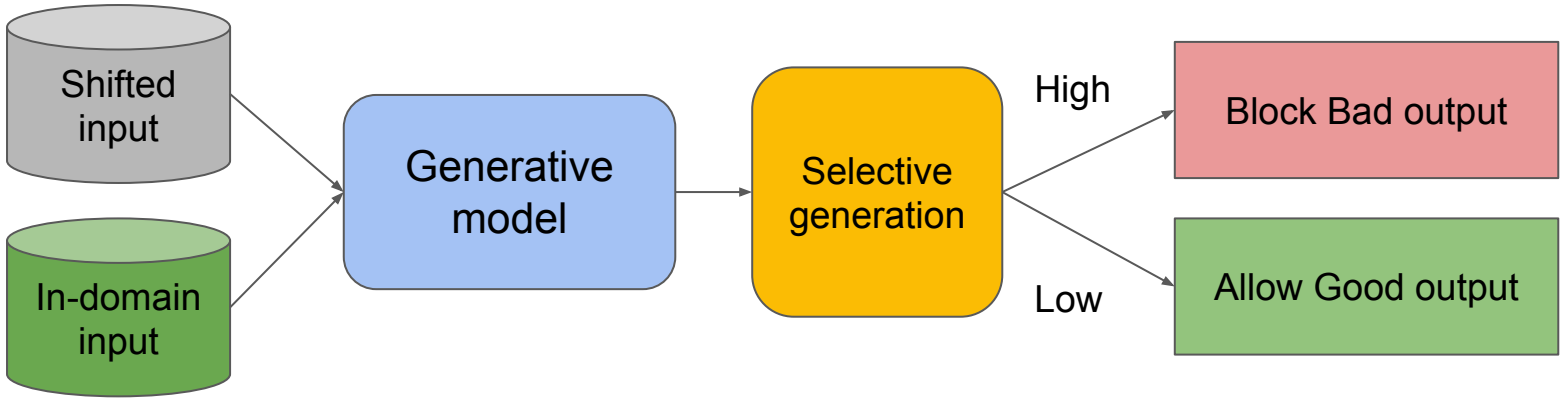
1. Most conservative: **Detect Out-of-distribution** and block/reject.
   e.g. model trained on BBC articles, do not use on any non-BBC



Google Research

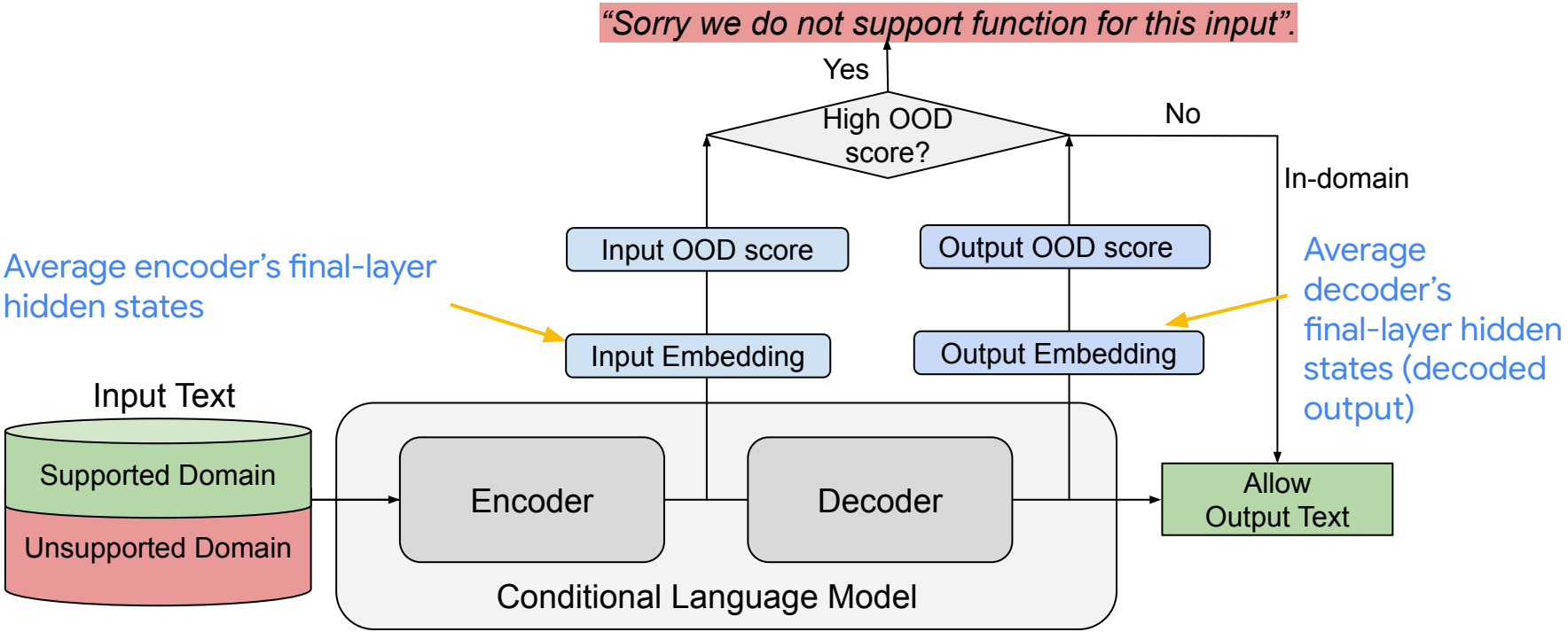
Two defensive strategies: OOD detection

1. Most conservative: **Detect Out-of-distribution** and block/reject.
   e.g. model trained on BBC articles, do not use on any non-BBC
2. More useful: Figure out how bad the response will be and ***selectively generate*** (analogous to selective classification).
   e.g. also use model on CNN articles when output is good



Google Research

Our Method: Detecting OOD using CLM's embeddings



Google Research

OOD detection – summarization (AUC-ROC)

(a) Summarization

Measure	Near Shift OOD		Far Shift OOD		
	cnn_dailymail	newsroom	reddit_tifu	forumsum	samsum
			INPUT OOD		
MD	0.651	0.799	0.974	0.977	0.995
RMD	0.828	0.930	0.998	0.997	**0.999**
Binary logits	**0.997**	0.959	**1.000**	**0.999**	0.998
			OUTPUT OOD		
Perplexity (baseline)	0.424	0.665	0.909	0.800	0.851
NLI score (baseline)	0.440	0.469	0.709	0.638	0.743
MD	0.944	0.933	0.985	0.973	0.985
RMD	0.958	0.962	0.998	0.993	0.998
Binary logits	0.989	**0.982**	**1.000**	0.998	0.997

NLI score measures the factual consistency by treating the input document as a premise and the generated summary as a hypothesis

Google Research

OOD detection – translation (AUC-ROC)

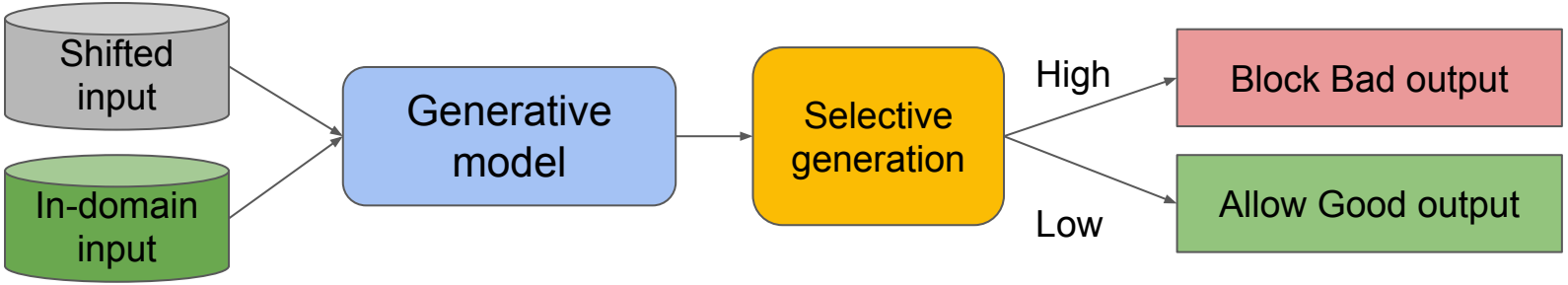
(b) Translation

Measure	WMT			OPUS					MTNT
	nt2014	ndd2015	ndt2015	law	medical	Koran	IT	sub	
INPUT OOD									
MD	0.534	0.671	0.670	0.511	0.704	0.737	0.828	0.900	0.668
RMD	0.798	0.866	0.863	0.389	0.840	0.957	**0.959**	**0.969**	0.943
Binary logits	**0.864**	**0.904**	**0.904**	0.485	0.813	**0.963**	0.928	0.950	**0.963**
OUTPUT OOD									
Perplexity (baseline)	0.570	0.496	0.494	0.392	0.363	0.657	0.343	0.359	0.633
COMET (baseline)	0.484	0.514	0.525	0.435	0.543	0.632	0.619	0.518	0.724
Prism (baseline)	0.445	0.504	0.505	0.459	0.565	0.716	0.604	0.577	0.699
MD	0.609	0.733	0.739	0.482	0.784	0.838	0.900	0.935	0.794
RMD	0.786	0.858	0.861	0.355	**0.845**	0.939	0.951	0.959	0.922
Binary logits	0.822	0.860	0.865	0.507	0.783	0.942	0.890	0.910	0.931

COMET: translation quality prediction model trained on human evaluation data
Prism: perplexity score from multilingual NMT model trained on 99.8M sentence pairs in 39 languages

Google Research

Selective generation



In practice, useful to use a model on a slightly shifted input distribution.

Combined scores improve the correlation over perplexity

(a) Summarization

Measure	In-domain	All
Single Score		
Perplexity (baseline)	0.256	0.300
NLI score (baseline)	0.337	0.381
Input RMD	0.015	0.336
Output RMD	0.053	0.385
Combined Score		
PR_{sum}(ppx, input RMD)	0.186	0.358
PR_{sum}(ppx, output RMD)	0.250	0.415
Linear Reg. (ppx, input & output)	0.235	**0.422**

12% increase

Quality := Human evaluation rating

(b) Translation

Measure	In-domain	All
Single Score		
Perplexity (baseline)	0.309	0.286
COMET (baseline)	0.184	0.336
Prism (baseline)	0.184	0.301
Input RMD	0.147	0.195
Output RMD	0.086	0.170
Combined Score		
PR_{sum}(ppx, input RMD)	0.321	**0.361**
PR_{sum}(ppx, output RMD)	0.323	0.356
Linear Reg. (ppx, input & output)	0.318	0.352

8% increase

Quality := BLEURT Sellam et al., 2020

Google Research

Selective generation (OOD): worst-scoring

Document: A crisp fan who gets through 42 bags in a week has discovered a skull-shaped deep-fried potato snack in one of his packets. Barry Selby, 54, who lives with his dog in Poole, Dorset, was eating a bag of cheese and onion crisps when he made the bizarre discovery, which appears to be a profile of a human skull. The floor-fitter has decided to keep the two inches tall by two-and-a-half inches wide snack as he believes it is far more impressive than other oddly-shaped examples he has seen on the internet.Scroll down for video. Spooky find: Barry Selby was eating a bag of Tesco cheese and onion crisps when he found the 'skull' snack. [truncated]

Model Summary: All images are copyrighted

Document: Last week she was barely showing – but Demelza Poldark is now the proud mother to theshow's latest addition. Within ten minutes of tomorrow night's episode, fans will see Aidan Turner's dash-ing Ross Poldark gaze lovingly at his new baby daughter. As Sunday night's latest heartthrob, womenacross the country have voiced their longing to settle down with the brooding Cornish gentleman – butunfortunately it seems as if his heart is well and truly off the market. Scroll down for video. Last weekshe was barely showing – but Demelza Poldark is now the proud mother to the show's latest addition. [truncated]

Model Summary: It's all change in the world of Poldark

(BBC summarization model on CNN/DM examples)

Google Research

Selective generation (OOD): best-scoring

Document: Rangers boss Stuart McCall says he is already working on a dossier of signing targets fornext season - even though he may not be around to parade them. The interim Ibrox manager still does notknow if he will be in charge beyond the current campaign after being lured back to his old club to kick-start their faltering promotion bid. So far, everything is going to plan with Gers second in the ScottishChampionship table and destined for a semi-final play-off slot. Stuart McCall says he is already lookingat transfer targets for next season, though he may not be at Rangers. [truncated]

Model Summary: Stuart McCall says he is already looking at transfer targets for next season, though hemay not be at Rangers

Document: An Alberta student who'd accidentally left his headlights on all day was greeted by what mayhave been the world's friendliest note from a stranger when he returned to his car. But Derek Murray, a University of Alberta law student, found more than just the note that cold November day in Edmonton–healso found an extension cord and battery charger left by the stranger to bring his dead Acura back to life. [truncated]

Model Summary: A Canadian student who accidentally left his headlights on all day was greeted bywhat may have been the world's friendliest note from a stranger when he returned to his car.

(BBC summarization model on CNN/DM examples)

Google Research

Conclusion

- OOD Detection
  - **simple**, **lightweight** (no ensembles, no external models), **high-performance** detector for seq2seq/CLMs – first for summarization
- Selective generation
  - extended idea of selective prediction to text generation
  - **safer** way to deploy CLMs **under distribution shift**

Full paper: https://arxiv.org/abs/2209.15558

Contact: {jjren, peterjliu}@google.com

Google Research