Google Research

Learning from Unlabeled Data

Thang Luong

Labeled data

Labeled data

Unlabeled data

Labeled data

Unlabeled data

Agenda

- Advancing Semi-Supervised Learning
  - *Unsupervised Data Augmentation*: small labeled set
  - *NoisyStudent*: large labeled set

- *Meena* – Towards a human-like open-domain chatbot

- (next time) *ELECTRA* – Efficient Language Pretraining

Semi-Supervised Learning (SSL)

SSL is apparently an important acronym & **overloaded**!

Why now?

The Quiet Semi-Supervised Revolution

Time to dust off that unlabeled data?

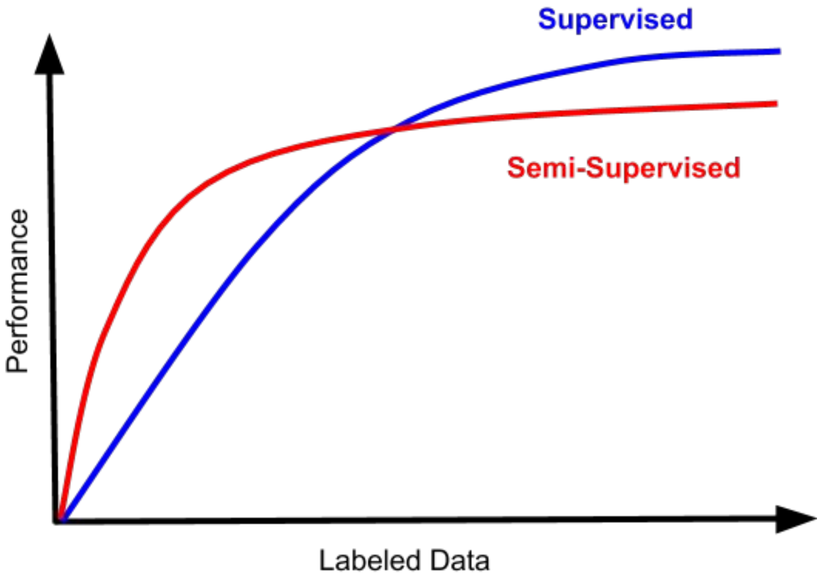
Vincent Vanhoucke Follow

May 15, 2019 · 5 min read ★

Our work “**Unsupervised Data Augmentation (UDA)**” was featured.

https://towardsdatascience.com/the-quiet-semi-supervised-revolution-edec1e9ad8c

Why now?

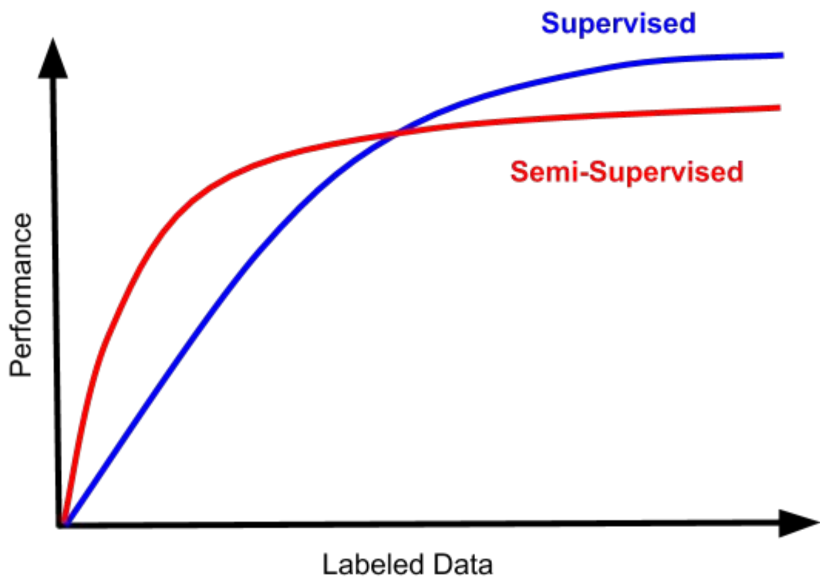


Belief of many ML practitioners

When there is enough labeled data, who cares about SSL?

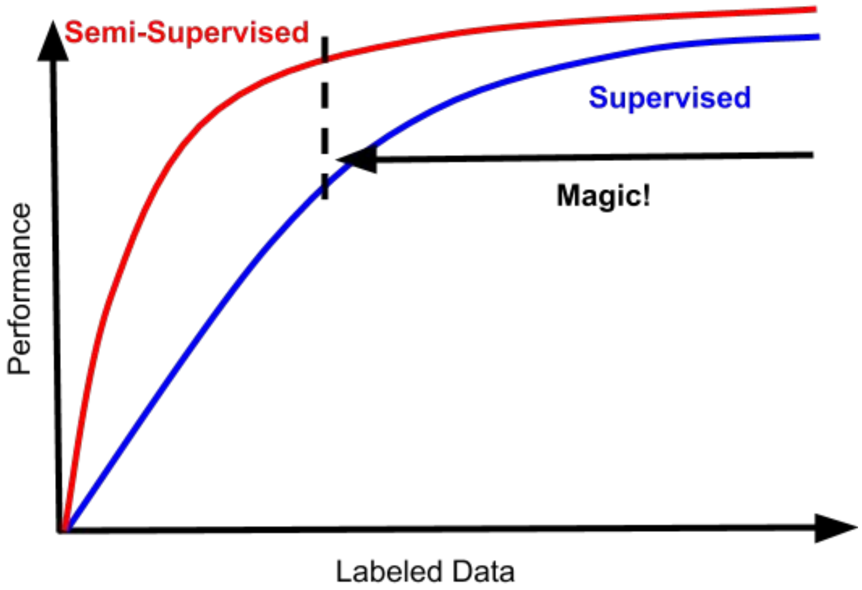
Google Research

Why now?



Belief of many ML practitioners

Past

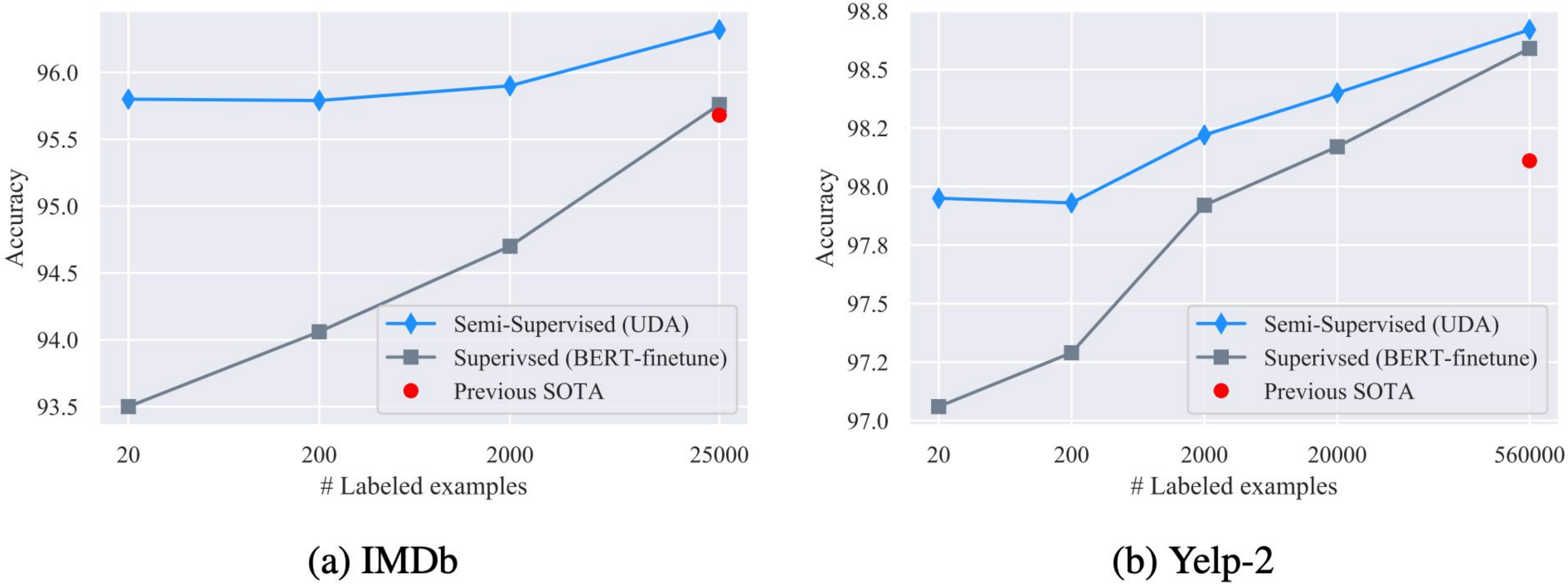


Dream of many SSL researchers

Now

Google Research

In our UDA paper:



Matches Vincent's mental picture: SSL > Supervised!

Same for vision (CIFAR, SVHN)

Google Research

Unsupervised Data Augmentation (UDA) for Consistency Training



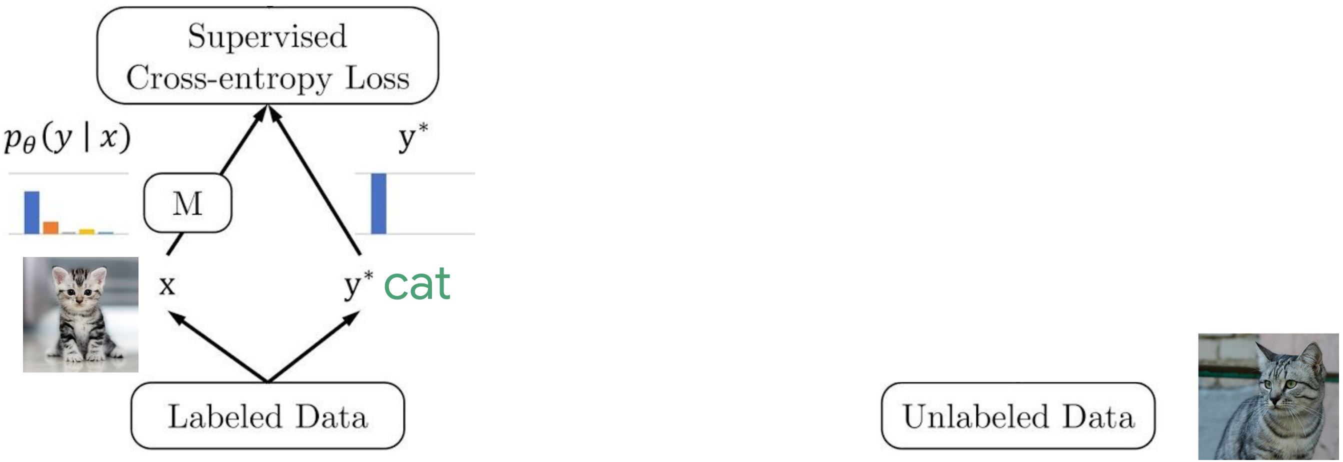


Qizhe Xie Zihang Dai Eduard Hovy Thang Luong Quoc Le

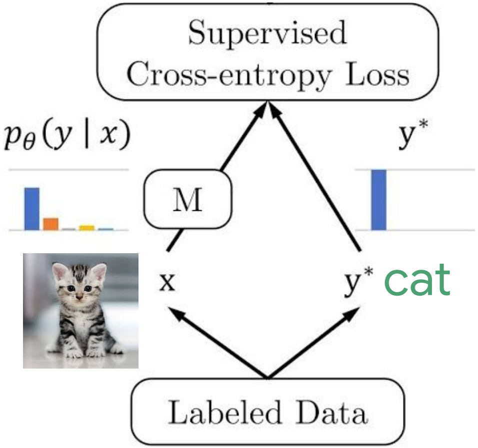
Paper: https://arxiv.org/abs/1904.12848
Code: https://github.com/google-research/uda

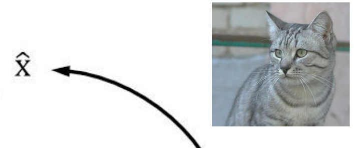
Google Research

Consistency Training in Semi-Supervised Learning

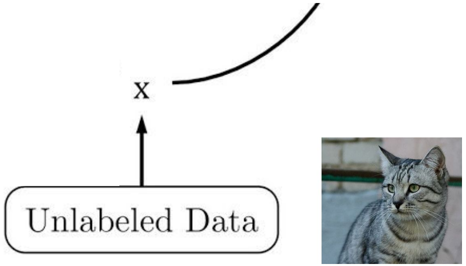


Consistency Training in Semi-Supervised Learning



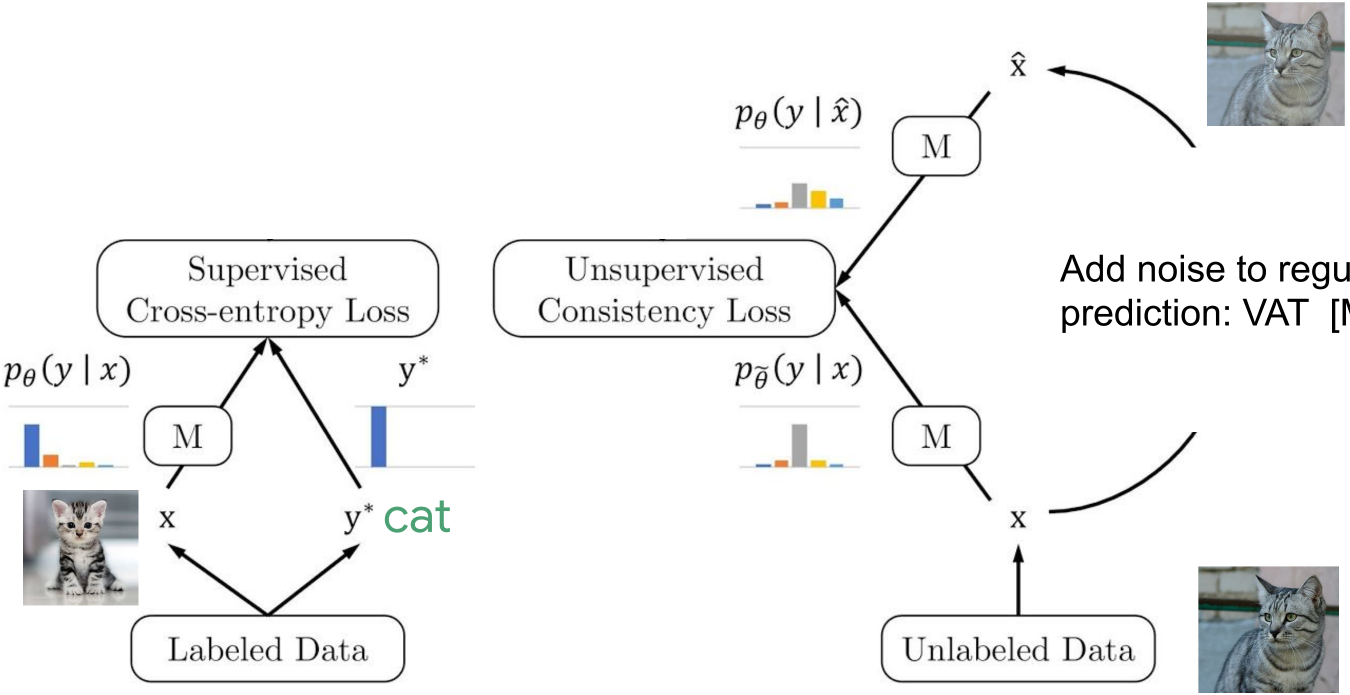


Add noise to regularize model prediction: VAT [Miyato et al., 2018]



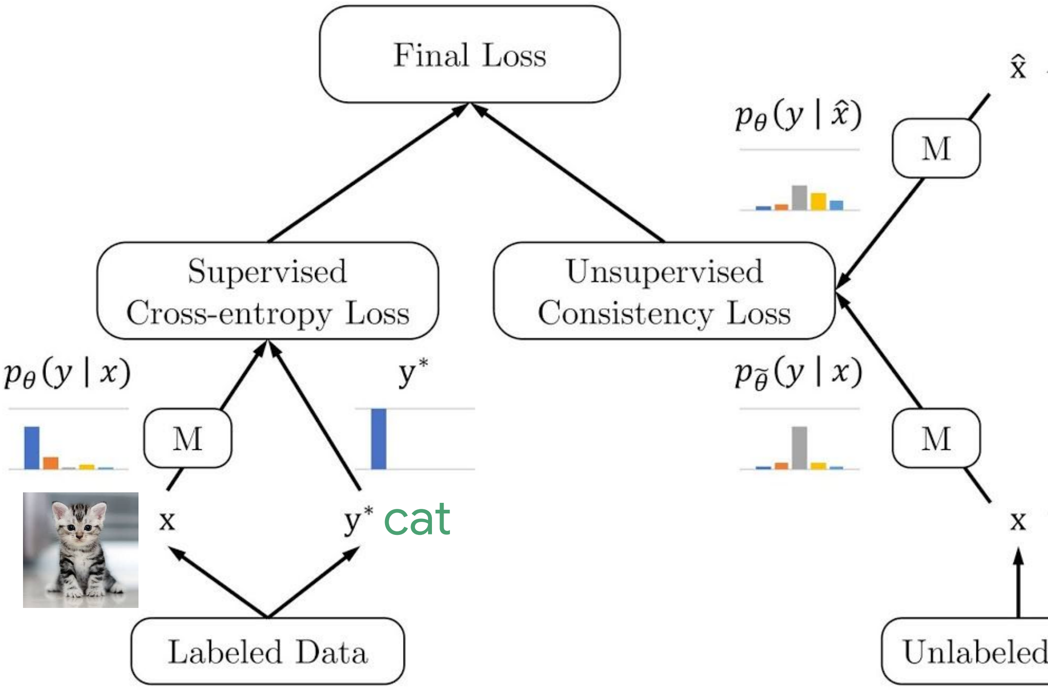
Google Research

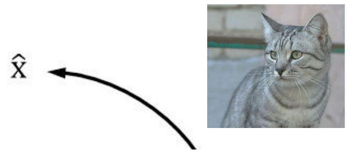
Consistency Training in Semi-Supervised Learning



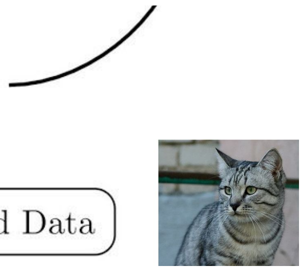
Add noise to regularize model prediction: VAT [Miyato et al., 2018]

Consistency Training in Semi-Supervised Learning

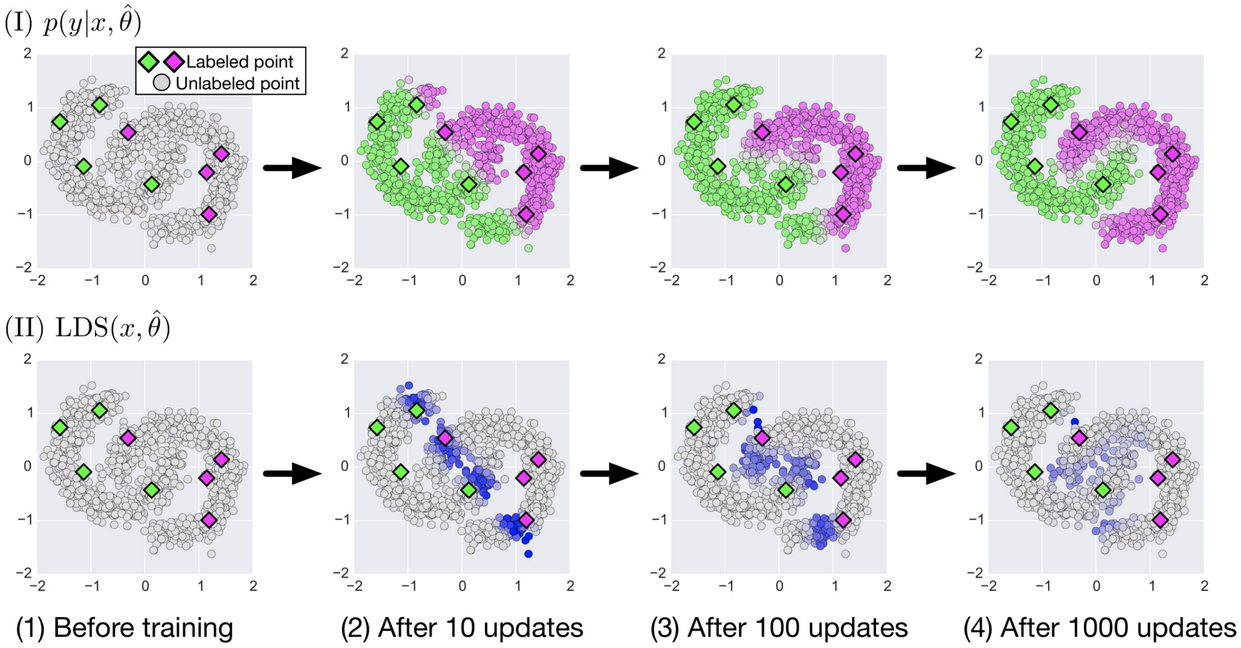




Add noise to regularize model prediction: VAT [Miyato et al., 2018]

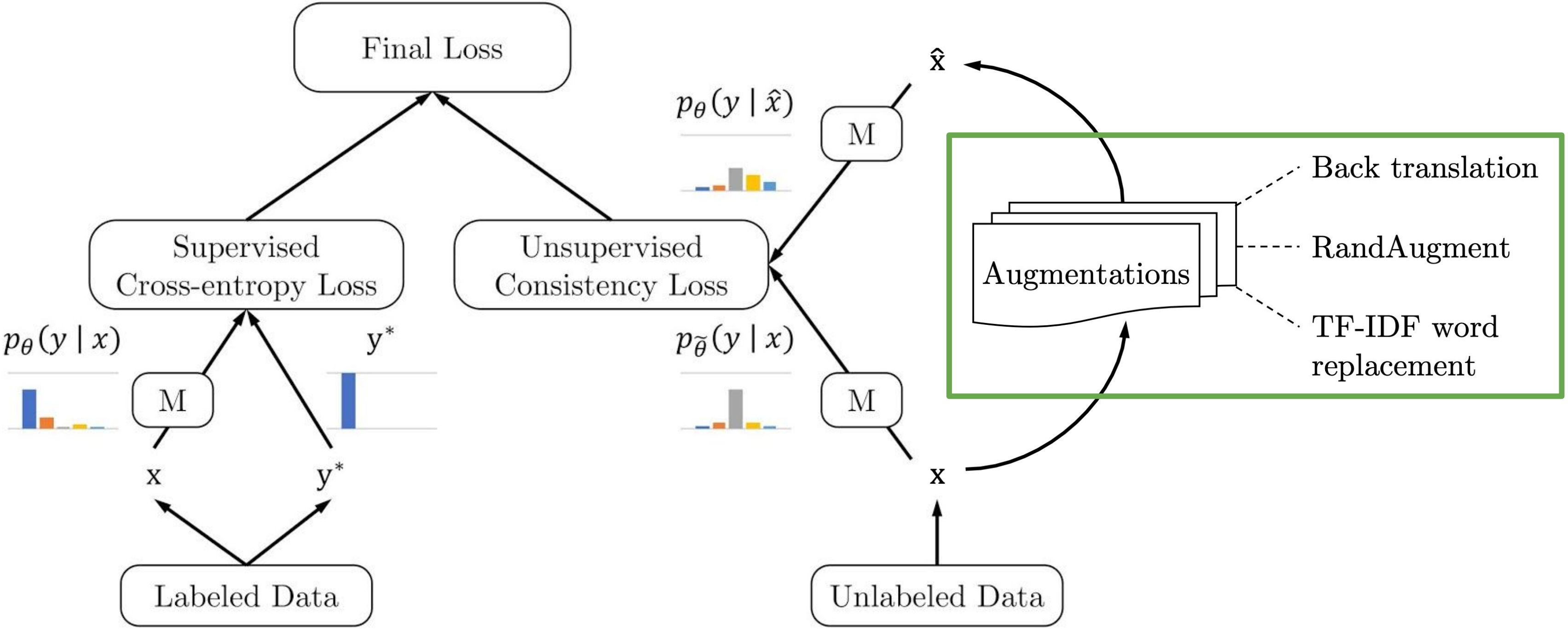


Label Propagation



Graph taken from VAT (Miyato et al. 2017)

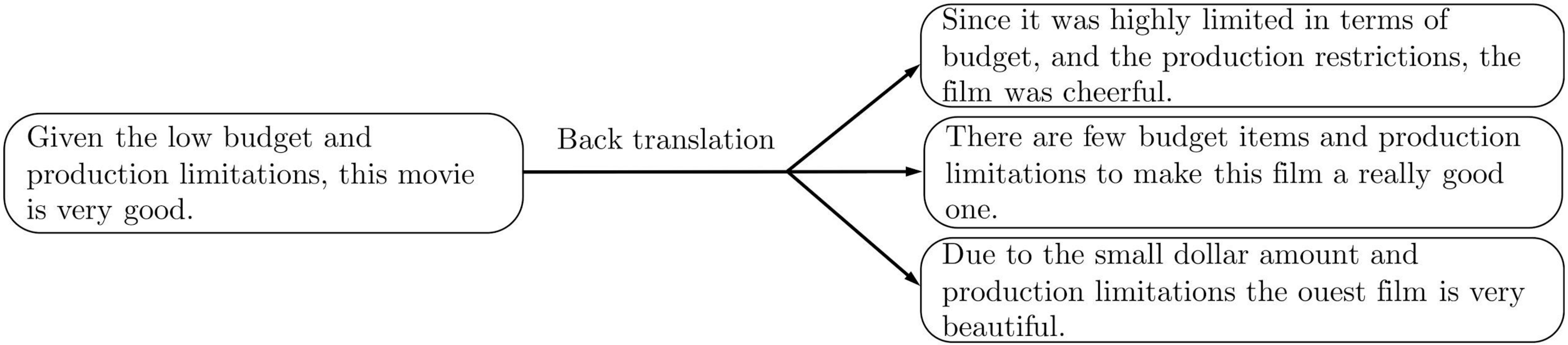
Unsupervised Data Augmentation (UDA)



UDA

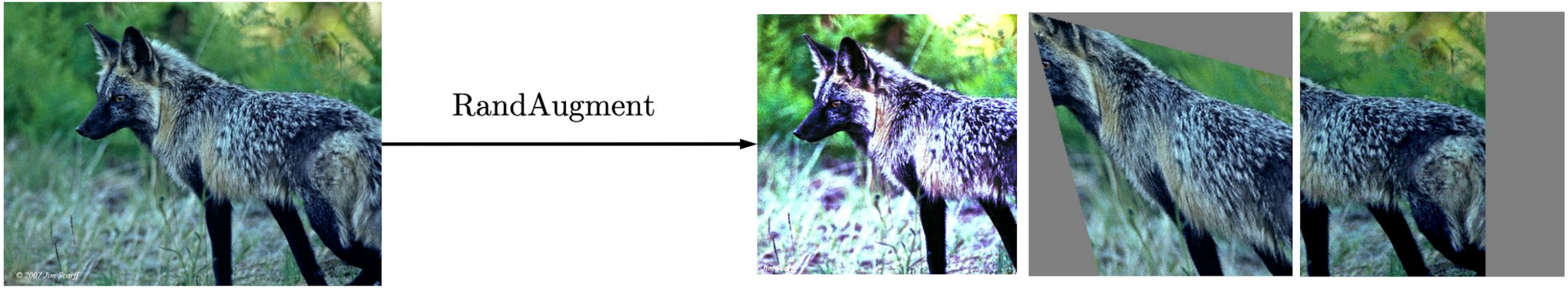
apply SOTA data augmentation to **unlabeled data**
to improve **semi-supervised learning**

Augmentation provides Diverse and Valid Perturbations



- Back translation for Text Classification:
  - English –> French –> English
  - Sampling: diverse (high-temperature) vs valid (low-temperature).
  - Used in QANet (Yu et al., 2018) for labeled data only.

Augmentation injects task-specific knowledge



- RandAugment (Cubuk et al., 2019) for Image Classification:
  - Example policies: (Rotate, 0.8, 2), (Brightness, 0.8, 4)

Google Research

Results

Ablation study on data augmentation

Augmentation (# Sup examples)	Sup (50k)	Semi-Sup (4k)
Crop & flip	5.36	16.17
Cutout	4.42	6.42
RandAugment	**4.23**	**5.29**

Table 1: Error rates on CIFAR-10.

Google Research

Ablation study on data augmentation

Augmentation (# Sup examples)	Sup (50k)	Semi-Sup (4k)
Crop & flip	5.36	16.17
Cutout	4.42	6.42
RandAugment	**4.23**	**5.29**

Table 1: Error rates on CIFAR-10.

Augmentation (# Sup examples)	Sup (650k)	Semi-sup (2.5k)
✗	38.36	50.80
Switchout	37.24	43.38
Back-translation	**36.71**	**41.35**

Table 2: Error rate on Yelp-5.

State-of-the-art augmentation is important!

SSL Benchmarks on CIFAR-10 and SVHN (Sep, 2019)

Method	**Model**	**# Param**	CIFAR-10 (4k)	SVHN (1k)
Π-Model (Laine & Aila, 2016)	Conv-Large	3.1M	12.36 ± 0.31	4.82 ± 0.17
Mean Teacher (Tarvainen & Valpola, 2017)	Conv-Large	3.1M	12.31 ± 0.28	3.95 ± 0.19
VAT + EntMin (Miyato et al., 2018)	Conv-Large	3.1M	10.55 ± 0.05	3.86 ± 0.11
SNTG (Luo et al., 2018)	Conv-Large	3.1M	10.93 ± 0.14	3.86 ± 0.27
VAdD (Park et al., 2018)	Conv-Large	3.1M	11.32 ± 0.11	4.16 ± 0.08
Fast-SWA (Athiwaratkun et al., 2018)	Conv-Large	3.1M	9.05	-
ICT (Verma et al., 2019)	Conv-Large	3.1M	7.29 ± 0.02	3.89 ± 0.04
Pseudo-Label (Lee, 2013)	WRN-28-2	1.5M	16.21 ± 0.11	7.62 ± 0.29
LGA + VAT (Jackson & Schulman, 2019)	WRN-28-2	1.5M	12.06 ± 0.19	6.58 ± 0.36
mixmixup (Hataya & Nakayama, 2019)	WRN-28-2	1.5M	10	-
ICT (Verma et al., 2019)	WRN-28-2	1.5M	7.66 ± 0.17	3.53 ± 0.07
MixMatch (Berthelot et al., 2019)	WRN-28-2	1.5M	6.24 ± 0.06	2.89 ± 0.06
Mean Teacher (Tarvainen & Valpola, 2017)	Shake-Shake	26M	6.28 ± 0.15	-
Fast-SWA (Athiwaratkun et al., 2018)	Shake-Shake	26M	5.0	-
MixMatch (Berthelot et al., 2019)	WRN	26M	4.95 ± 0.08	-

SSL Benchmarks on CIFAR-10 and SVHN (Sep, 2019)

Method	Model	# Param	CIFAR-10 (4k)	SVHN (1k)
Π-Model (Laine & Aila, 2016)	Conv-Large	3.1M	12.36 ± 0.31	4.82 ± 0.17
Mean Teacher (Tarvainen & Valpola, 2017)	Conv-Large	3.1M	12.31 ± 0.28	3.95 ± 0.19
VAT + EntMin (Miyato et al., 2018)	Conv-Large	3.1M	10.55 ± 0.05	3.86 ± 0.11
SNTG (Luo et al., 2018)	Conv-Large	3.1M	10.93 ± 0.14	3.86 ± 0.27
VAdD (Park et al., 2018)	Conv-Large	3.1M	11.32 ± 0.11	4.16 ± 0.08
Fast-SWA (Athiwaratkun et al., 2018)	Conv-Large	3.1M	9.05	-
ICT (Verma et al., 2019)	Conv-Large	3.1M	7.29 ± 0.02	3.89 ± 0.04
Pseudo-Label (Lee, 2013)	WRN-28-2	1.5M	16.21 ± 0.11	7.62 ± 0.29
LGA + VAT (Jackson & Schulman, 2019)	WRN-28-2	1.5M	12.06 ± 0.19	6.58 ± 0.36
mixmixup (Hataya & Nakayama, 2019)	WRN-28-2	1.5M	10	-
ICT (Verma et al., 2019)	WRN-28-2	1.5M	7.66 ± 0.17	3.53 ± 0.07
MixMatch (Berthelot et al., 2019)	WRN-28-2	1.5M	6.24 ± 0.06	2.89 ± 0.06
Mean Teacher (Tarvainen & Valpola, 2017)	Shake-Shake	26M	6.28 ± 0.15	-
Fast-SWA (Athiwaratkun et al., 2018)	Shake-Shake	26M	5.0	-
MixMatch (Berthelot et al., 2019)	WRN	26M	4.95 ± 0.08	-
UDA (RandAugment)	WRN-28-2	1.5M	5.29 ± 0.25	$\mathbf{2.55 \pm 0.09}$
UDA (RandAugment)	Shake-Shake	26M	3.7	-
UDA (RandAugment)	PyramidNet	26M	**2.7**	-

15% error reduction from previous SOTA (30% in Apr, 2019)

SSL Benchmarks on CIFAR-10 and SVHN (Sep, 2019)

Method	Model	# Param	CIFAR-10 (4k)	SVHN (1k)
Π-Model (Laine & Aila, 2016)	Conv-Large	3.1M	12.36 ± 0.31	4.82 ± 0.17
Mean Teacher (Tarvainen & Valpola, 2017)	Conv-Large	3.1M	12.31 ± 0.28	3.95 ± 0.19
VAT + EntMin (Miyato et al., 2018)	Conv-Large	3.1M	10.55 ± 0.05	3.86 ± 0.11
SNTG (Luo et al., 2018)	Conv-Large	3.1M	10.93 ± 0.14	3.86 ± 0.27
VAdD (Park et al., 2018)	Conv-Large	3.1M	11.32 ± 0.11	4.16 ± 0.08
Fast-SWA (Athiwaratkun et al., 2018)	Conv-Large	3.1M	9.05	-
ICT (Verma et al., 2019)	Conv-Large	3.1M	7.29 ± 0.02	3.89 ± 0.04
Pseudo-Label (Lee, 2013)	WRN-28-2	1.5M	16.21 ± 0.11	7.62 ± 0.29
LGA + VAT (Jackson & Schulman, 2019)	WRN-28-2	1.5M	12.06 ± 0.19	6.58 ± 0.36
mixmixup (Hataya & Nakayama, 2019)	WRN-28-2	1.5M	10	-
ICT (Verma et al., 2019)	WRN-28-2	1.5M	7.66 ± 0.17	3.53 ± 0.07
MixMatch (Berthelot et al., 2019)	WRN-28-2	1.5M	6.24 ± 0.06	2.89 ± 0.06
Mean Teacher (Tarvainen & Valpola, 2017)	Shake-Shake	26M	6.28 ± 0.15	-
Fast-SWA (Athiwaratkun et al., 2018)	Shake-Shake	26M	5.0	-
MixMatch (Berthelot et al., 2019)	WRN	26M	4.95 ± 0.08	-
UDA (RandAugment)	WRN-28-2	1.5M	5.29 ± 0.25	$\mathbf{2.55 \pm 0.09}$
UDA (RandAugment)	Shake-Shake	26M	3.7	-
UDA (RandAugment)	PyramidNet	26M	**2.7**	-

Further advancing the SOTA with larger networks

Works follow UDA in using strong augmentation!

Google Research

Algorithm	Artificial label augmentation	Prediction augmentation	Artificial label post-processing
TS [39]/Π-Model [36]	Weak	Weak	None
Temporal Ensembling [21]	Weak	Weak	None
Mean Teacher [43]	Weak	Weak	None
Virtual Adversarial Training [28]	None	Adversarial	None
UDA [45]	Weak	Strong	Sharpening
MixMatch [3]	Weak	Weak	Sharpening
ReMixMatch [2]	Weak	Strong	Sharpening
FixMatch	Weak	Strong	Pseudo-labeling

FixMatch (Sohn et al, 2020) & ReMixMatch (Berthelot et al., 2019) use **strong augmentation** (RandAugment, CTAugment)

(Table taken from FixMatch paper)

Summary

- Data augmentation is an effective perturbation for SSL.
- UDA significantly improves for both language and vision.
- UDA combines well with transfer learning, e.g., BERT.

Paper: https://arxiv.org/abs/1904.12848
Code: https://github.com/google-research/uda

So far, success has only been in low-data regime!

Small labeled Data
(CIFAR, SVHN)

State-of-the-art

FixMatch, ReMixMatch
UDA, MixMatch, S4L,
ICT, VAT, etc.

Large **labeled** Data
(ImageNet)

No state-of-the art results

Self-training with Noisy Student improves ImageNet classification




Qizhe Xie Thang Luong Eduard Hovy Quoc Le

Paper: https://arxiv.org/abs/1911.04252
Code: https://github.com/google-research/noisystudent

What is NoisyStudent?

4 simple steps:

What is NoisyStudent?

Google Research

4 simple steps:

1. Train a classifier on the labeled (L) data (teacher)

steel arch bridge canoe

Labeled data

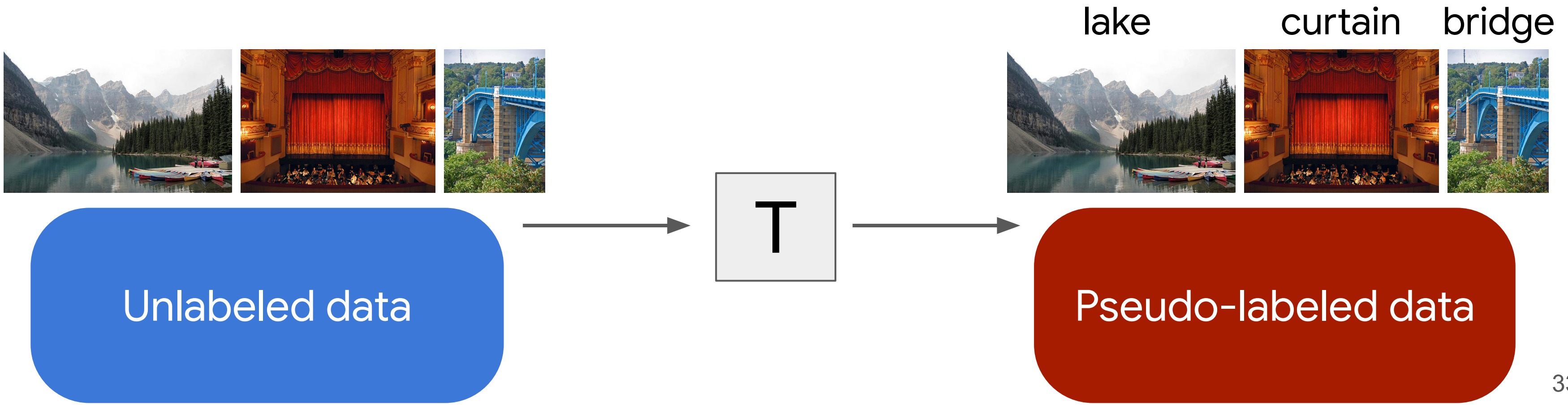
T

Google Research

What is NoisyStudent?

4 simple steps:

1. Train a classifier on the labeled (L) data (teacher)
2. Infer labels on a much larger unlabeled dataset ⟶ P

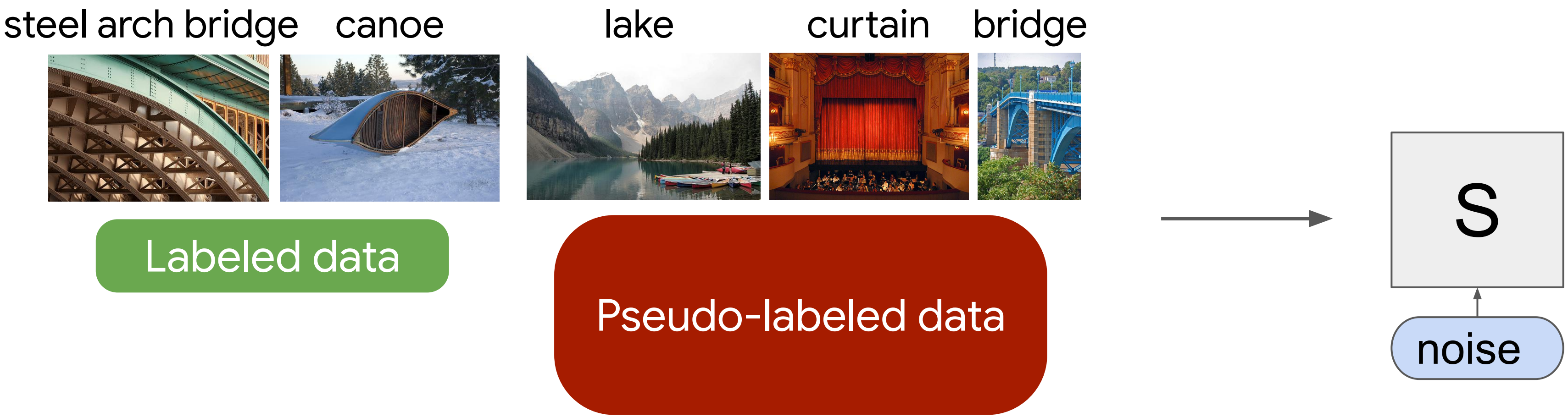


Google Research

What is NoisyStudent?

4 simple steps:

1. Train a classifier on the labeled (L) data (teacher)
2. Infer labels on a much larger unlabeled dataset ⟶ P
3. Train a larger classifier on **L + P**, adding noise (noisy student)



Google Research

What is NoisyStudent?

4 simple steps:

1. Train a classifier on the labeled (L) data (teacher)
2. Infer labels on a much larger unlabeled dataset ⟶ P
3. Train a larger classifier on **L + P**, adding noise (noisy student)
    a. Data Augmentation
    b. Dropout
    c. Stochastic Depth

Google Research

What is NoisyStudent?

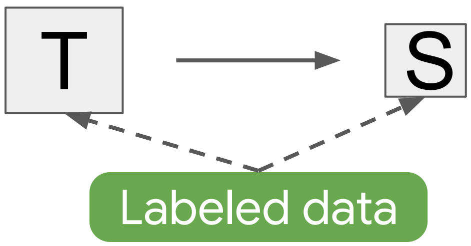
4 simple steps:

1. Train a classifier on the labeled (L) data (teacher)
2. Infer labels on a much larger unlabeled dataset ⟶ P
3. Train a larger classifier on **L + P**, adding noise (noisy student)
    a. Data Augmentation
    b. Dropout
    c. Stochastic Depth
4. Go to step 2, with student as teacher

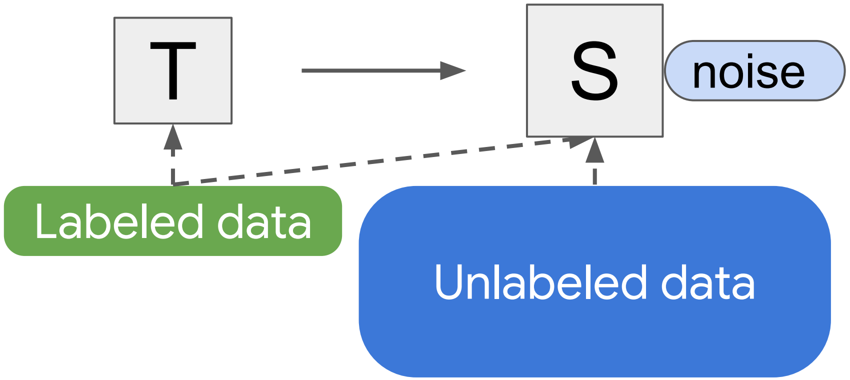
Google Research

NoisyStudent vs. Distillation

- Distillation focuses on speed rather than quality
    - no student noise, no unlabeled data, smaller student

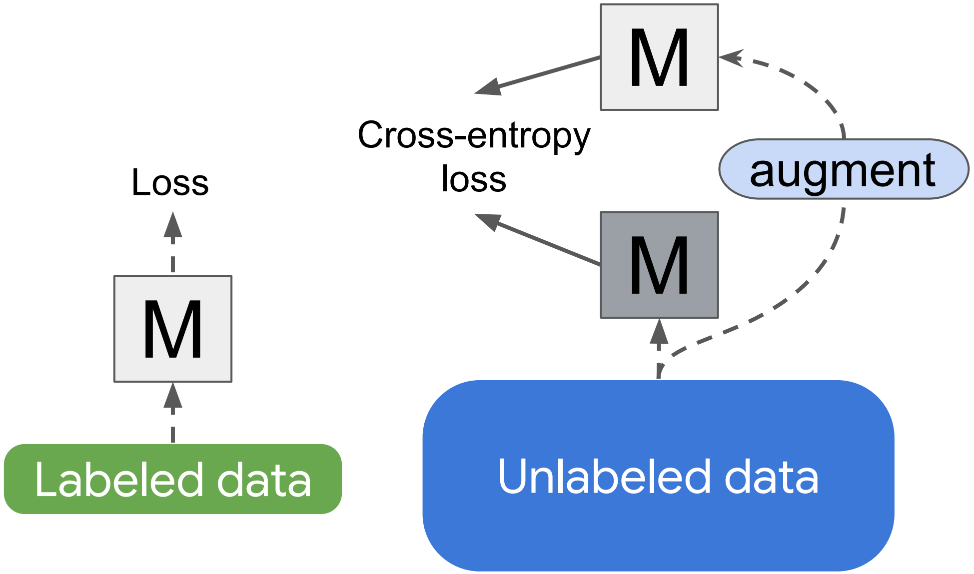


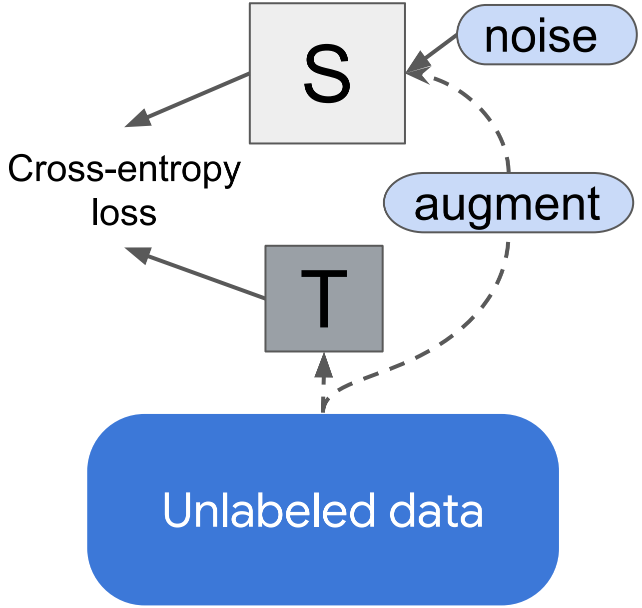
Distillation



NoisyStudent

Consistency Training vs. Self-Training





Consistency training (UDA, FixMatch)

Single model M jointly trained from scratch

Works great with small labeled data

Self-Training (NoisyStudent)

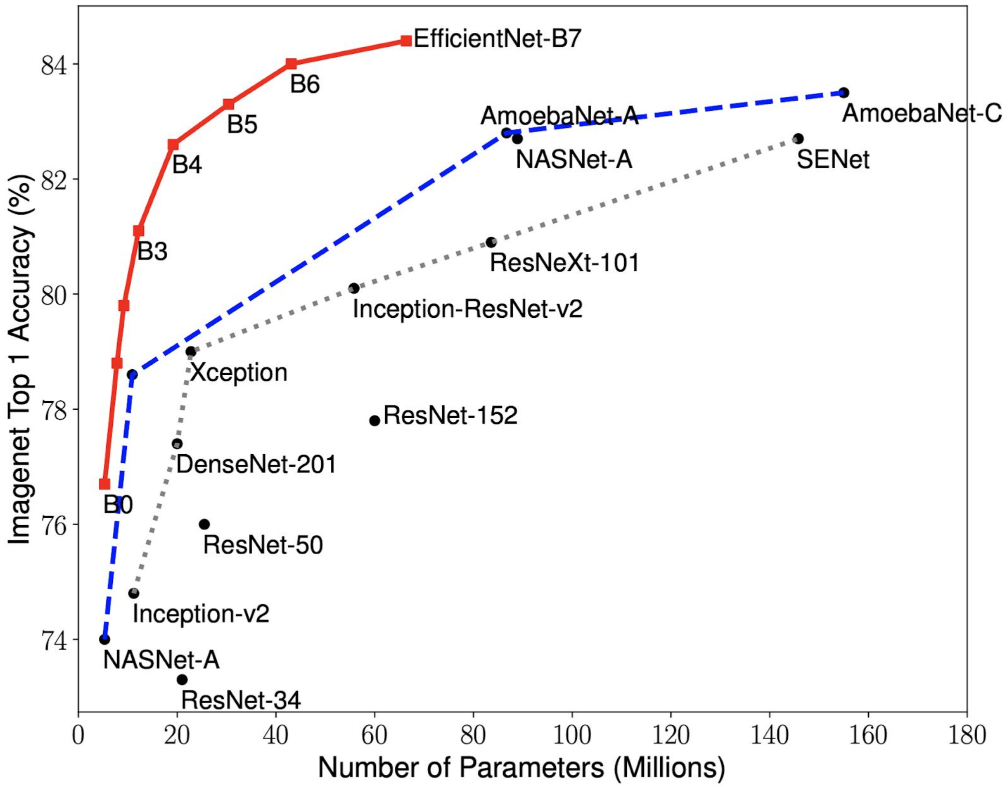
Requires a converged teacher T

Works great with large labeled data

Experiments

Settings

- Architecture: EfficientNets.



Settings

- Architecture: EfficientNets (Tan & Le, 2019).
- Labeled dataset: ImageNet (1.3M images).
- Unlabeled dataset: JFT (300M unlabeled images).
    - Pseudo-labels: soft pseudo-labels (continuous).
- Iterative training: B7->L2->L2->L2

ImageNet Results

Method	# Param	Extra Data	Top-1 Acc.	Top-5 Acc.
GPipe	557M	-	84.3%	97.0%
EfficientNet-B7	66M	-	85.0%	97.2%
EfficientNet-L2	480M	-	85.5%	97.5%
ResNeXt-101 WSL	829M	3.5B instagram images labeled with tags	85.4%	97.6%
FixRes ResNeXt-101 WSL	829M	3.5B instagram images labeled with tags	86.4%	98.0%
Noisy Student (EfficientNet-L2)	480M	300M unlabeled images	**88.4%**	**98.7%**

- SOTA: 2% improvement of top-1 accuracy.

ImageNet Results

Method	# Param	Extra Data	Top-1 Acc.	Top-5 Acc.
GPipe	557M	-	84.3%	97.0%
EfficientNet-B7	66M	-	85.0%	97.2%
EfficientNet-L2	480M	-	85.5%	97.5%
ResNeXt-101 WSL	829M	3.5B instagram images labeled with tags	85.4%	97.6%
FixRes ResNeXt-101 WSL	829M	3.5B instagram images labeled with tags	86.4%	98.0%
Noisy Student (EfficientNet-L2)	480M	300M unlabeled images	**88.4%**	**98.7%**

- SOTA: 2% improvement of top-1 accuracy.
- One order of magnitude less unlabeled data.

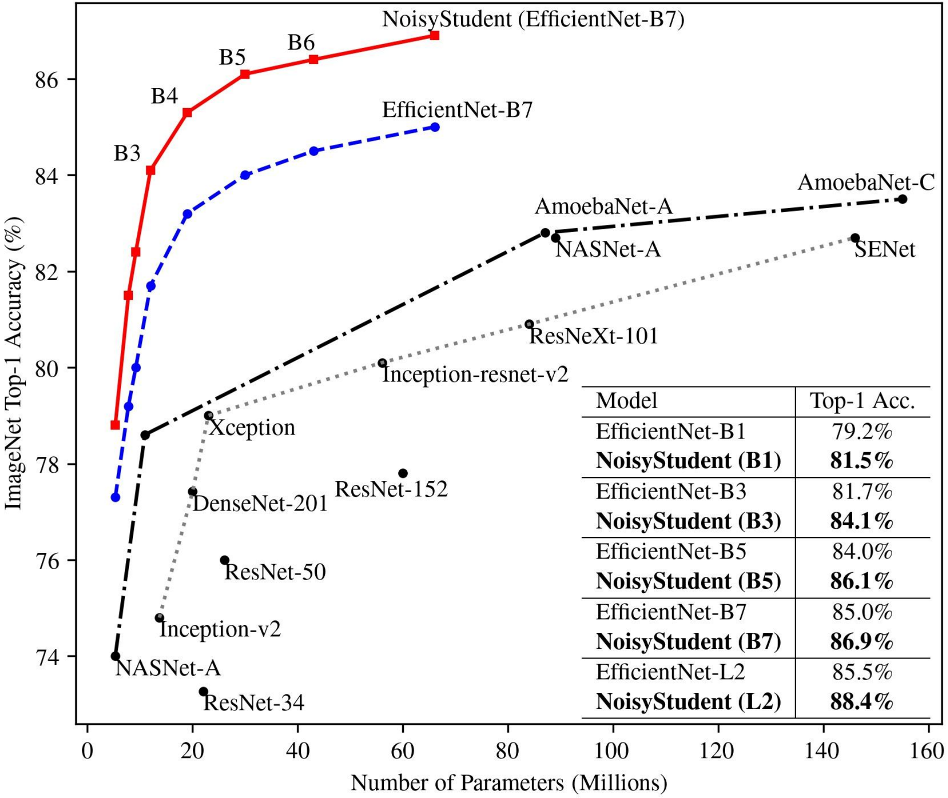
ImageNet Results

Method	# Param	Extra Data	Top-1 Acc.	Top-5 Acc.
GPipe	557M	-	84.3%	97.0%
EfficientNet-B7	66M	-	85.0%	97.2%
EfficientNet-L2	480M	-	85.5%	97.5%
ResNeXt-101 WSL	829M	3.5B instagram images labeled with tags	85.4%	97.6%
FixRes ResNeXt-101 WSL	829M	3.5B instagram images labeled with tags	86.4%	98.0%
Noisy Student (EfficientNet-L2)	480M	300M unlabeled images	**88.4%**	**98.7%**

- SOTA: 2% improvement of top-1 accuracy.
- One order of magnitude less unlabeled data.
- Twice as small in the number of parameters.

Google Research

Improvements across model sizes

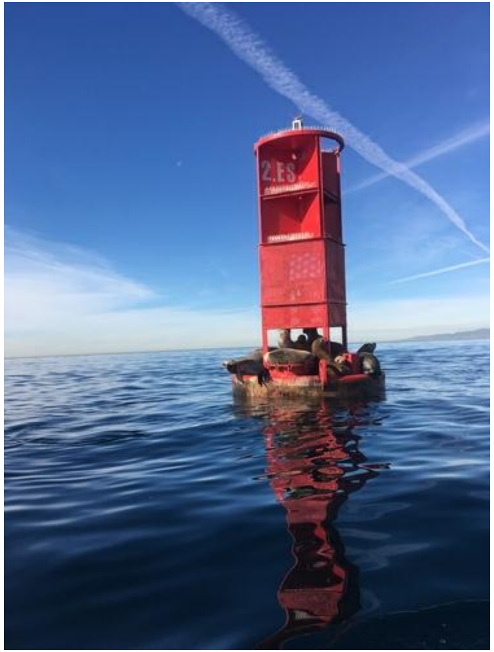


Model	Top-1 Acc.
EfficientNet-B1	79.2%
NoisyStudent (B1)	**81.5%**
EfficientNet-B3	81.7%
NoisyStudent (B3)	**84.1%**
EfficientNet-B5	84.0%
NoisyStudent (B5)	**86.1%**
EfficientNet-B7	85.0%
NoisyStudent (B7)	**86.9%**
EfficientNet-L2	85.5%
NoisyStudent (L2)	**88.4%**

Surprising Gains on Robustness Benchmarks

ImageNet-A: difficult images SOTA models failed.

ImageNet-A



Sea Lion (NoisyStudent)

Lighthouse (Baseline)

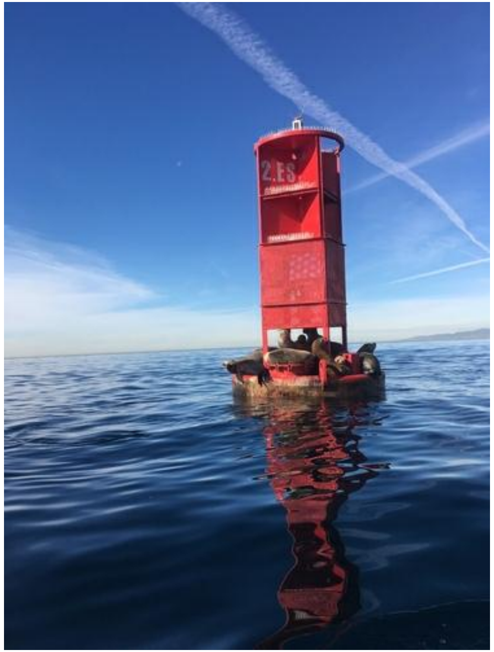
Surprising Gains on Robustness Benchmarks

	ImageNet-A top-1 acc. ↑	ImageNet-C mCE ↓	ImageNet-P mFR ↓
Prev. SOTA	61.0%	45.7	27.8
Ours	**83.7%**	**28.3**	**12.2**

ImageNet-A: difficult images SOTA models failed.

ImageNet-C & P: corrupted and perturbed images (blurring, fogging, rotation and scaling).

ImageNet-A



Sea Lion (NoisyStudent) | Lighthouse (Baseline)

ImageNet-C

Parking Meter (NoisyStudent) — Vacuum (Baseline)

Swing (NoisyStudent) — Mosquito Net (Baseline)

ImageNet-P


plate rack refrigerator
racing car car wheel
plate rack medicine chest
racing car fire engine
plate rack medicine chest
racing car car wheel

The Importance of Noise in Self-training

Model / Unlabeled Set Size	1.3M	130M
EfficientNet-B5	83.3%	84.0%
Noisy Student (B5)	**83.9%**	**84.9%**
student w/o Aug	83.6%	84.6%
student w/o Aug, SD, Dropout	83.2%	84.3%
teacher w. Aug, SD, Dropout	83.7%	84.4%

- Standard data augmentation is used when we use 1.3M unlabeled images.
- RandAugment is used when we use 130M unlabeled images.

Summary

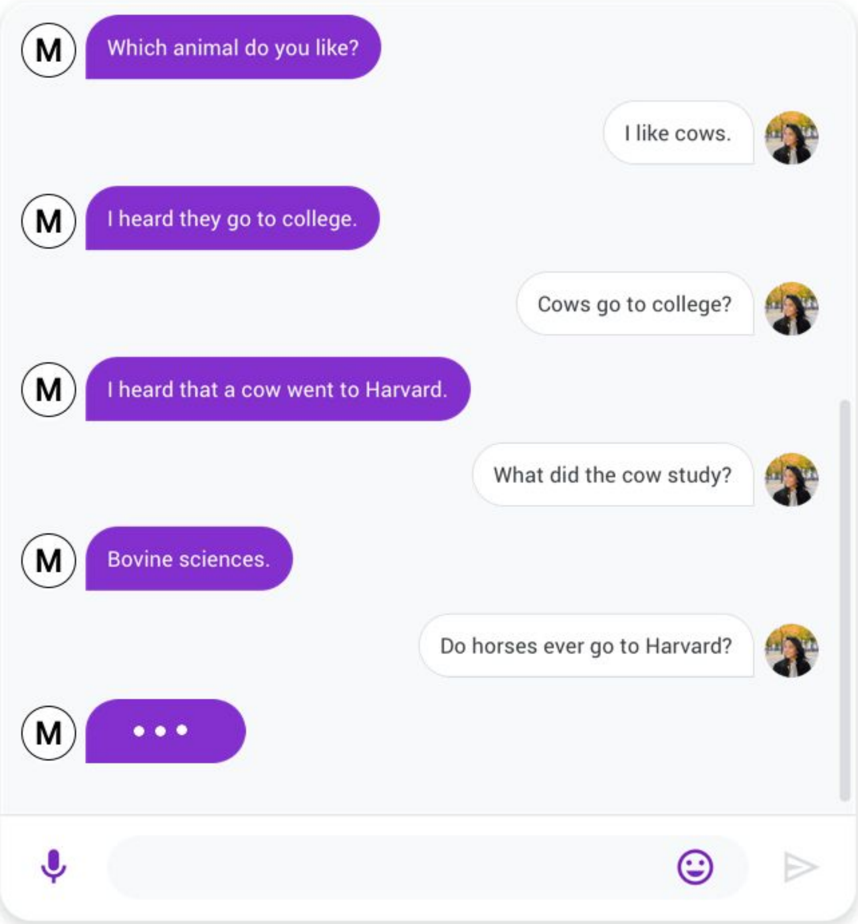
- Semi-supervised learning works at all scale!
- Possible to use unlabeled images to advance ImageNet SOTA
- Robustness gains for free.

Paper: https://arxiv.org/abs/1911.04252
Code: https://github.com/google-research/noisystudent

Google Research

Let's switch gear!

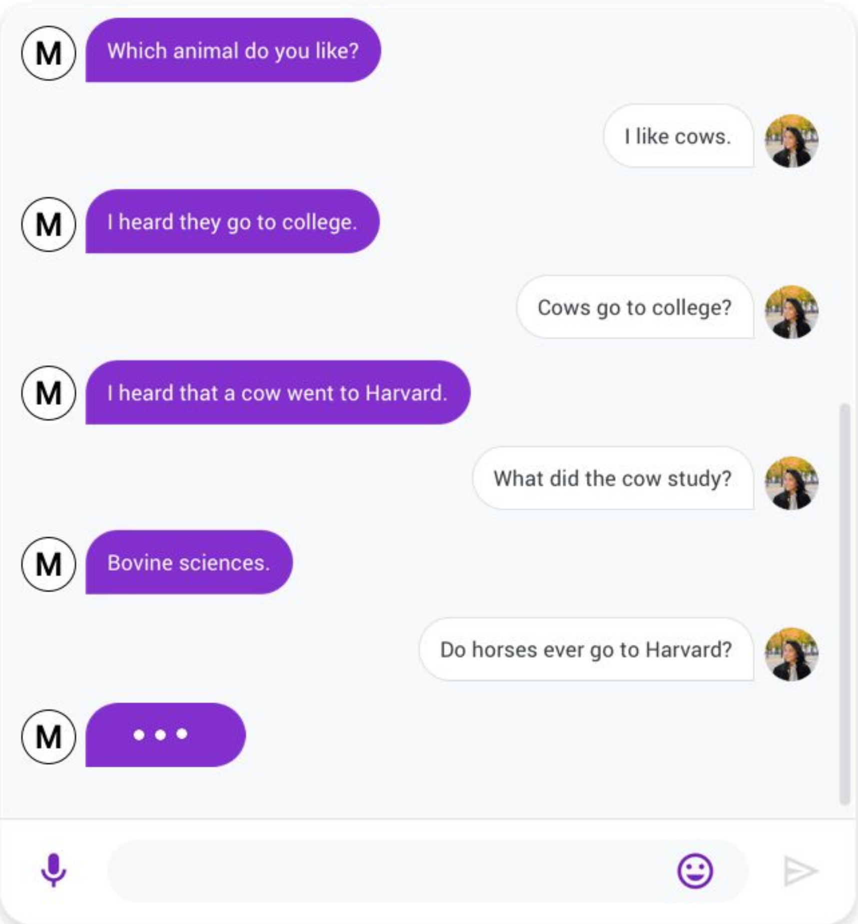
How many jokes do you see?



Google Research

Horses go to Hayvard!

And one more joke after that ...

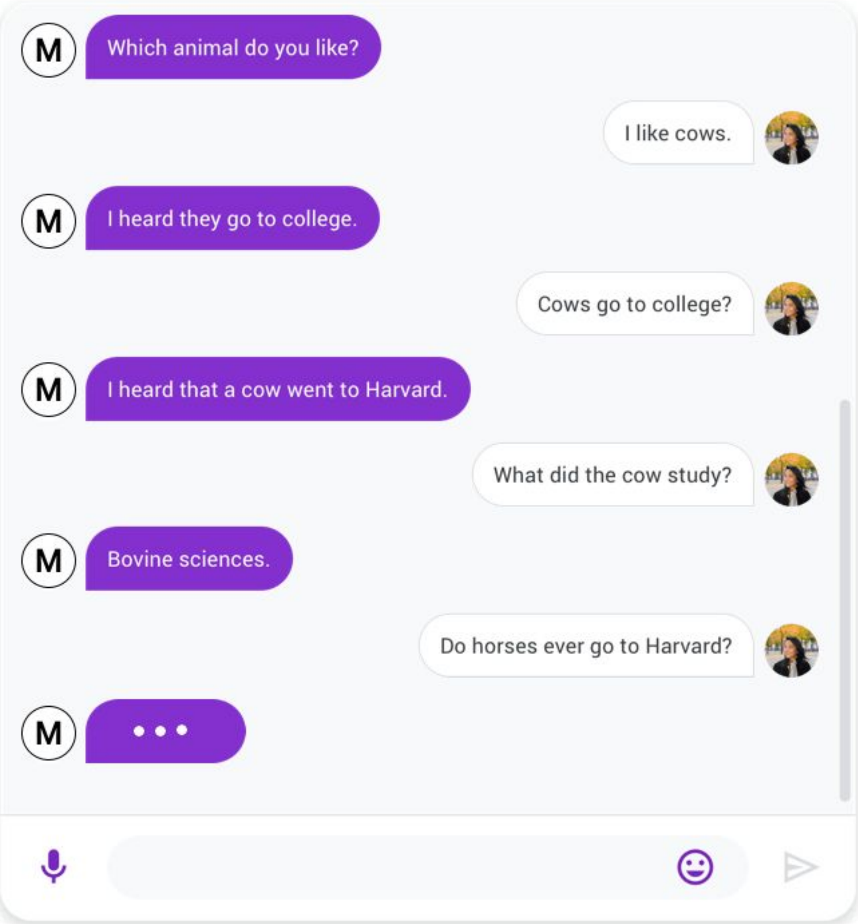


Horses go to Hayvard!

noun

noun: **steer**; plural noun: **steers**

a male domestic bovine animal that has been



Google Research

Meena – “Towards a Human-like Open-Domain Chatbot”

Daniel De Freitas Adiwardana

Thang Luong

Quoc Le

David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu

Paper: https://arxiv.org/abs/2001.09977
Blog: https://twitter.com/GoogleAI/status/1222230622355087360
Conversation samples:
https://github.com/google-research/google-research/tree/master/meena

By the way, we never said this …

MIT Technology Review

Artificial Intelligence Jan 30

Google says its new chatbot Meena is the best in the world

End-to-end neural conversational model

User: *How are things?*

Meena: They're good. How about you?

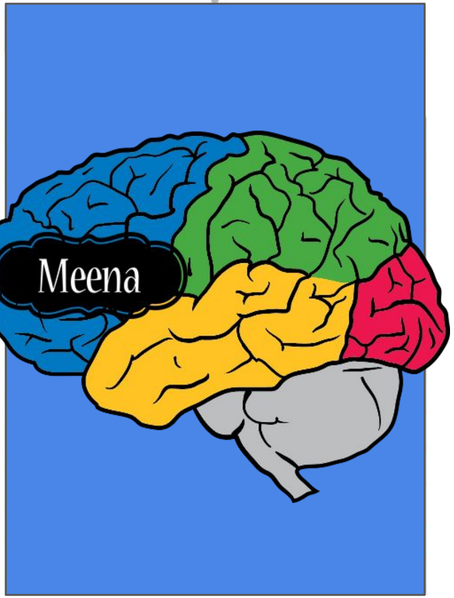
User: *I'm trying to decide what to watch*

Meena: What are some of your favorite shows?

User: *I like the Good Place*

Meena: I haven't watched that. My favorite shows are Supernatural, and Star Trek: TNG.

User: *What is TNG?*



7 turns of context

End-to-end neural conversational model

Google Research

User: *How are things?*

Meena: They're good. How about you?

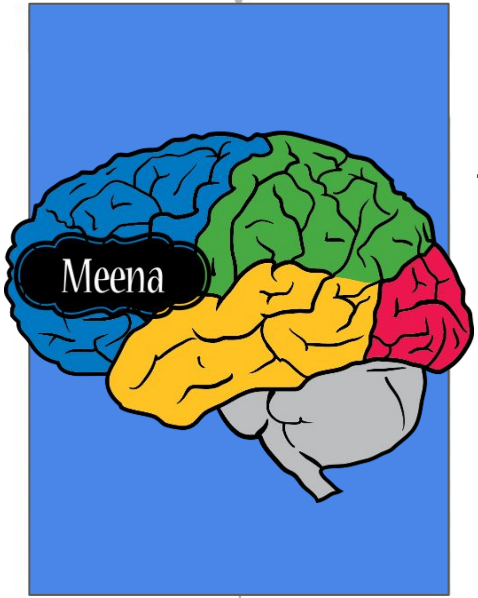
User: *I'm trying to decide what to watch*

Meena: What are some of your favorite shows?

User: *I like the Good Place*

Meena: I haven't watched that. My favorite shows are Supernatural, and Star Trek: TNG.

User: *What is TNG?*



7 turns of context

Trained to predict the next word.

Meena: The

End-to-end neural conversational model

Google Research

User: *How are things?*

Meena: They're good. How about you?

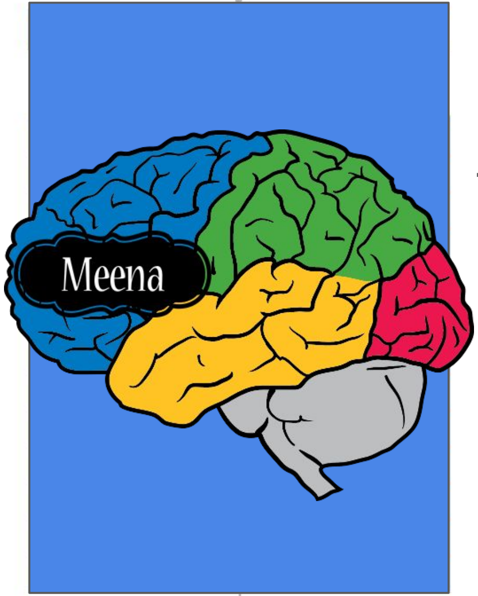
User: *I'm trying to decide what to watch*

Meena: What are some of your favorite shows?

User: *I like the Good Place*

Meena: I haven't watched that. My favorite shows are Supernatural, and Star Trek: TNG.

User: *What is TNG?*



7 turns of context

Trained to predict the next word.

Meena: The Next

End-to-end neural conversational model

User: *How are things?*

Meena: They're good. How about you?

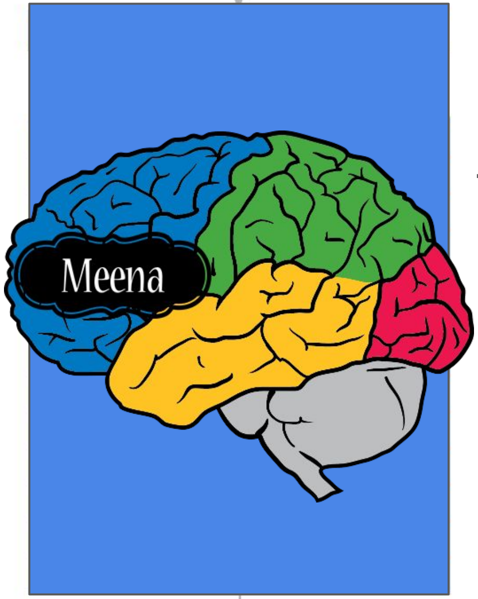
User: *I'm trying to decide what to watch*

Meena: What are some of your favorite shows?

User: *I like the Good Place*

Meena: I haven't watched that. My favorite shows are Supernatural, and Star Trek: TNG.

User: *What is TNG?*



7 turns of context

Trained to predict the next word.

Meena: The Next Generation

End-to-end neural conversational model

User: *How are things?*

Meena: They're good. How about you?

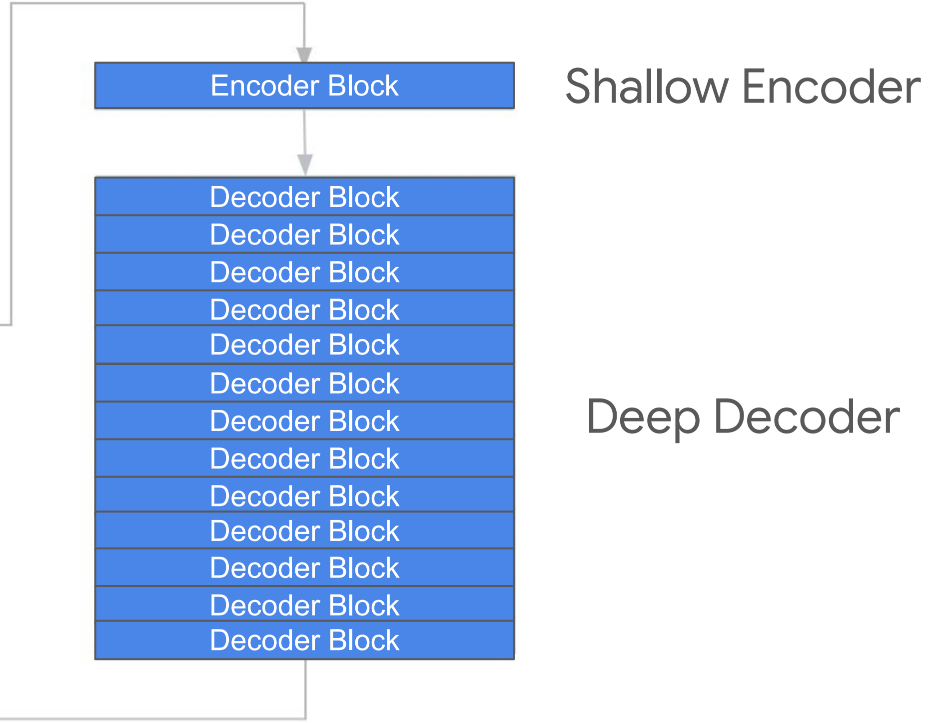
User: *I'm trying to decide what to watch*

Meena: What are some of your favorite shows?

User: *I like the Good Place*

Meena: I haven't watched that. My favorite shows are Supernatural, and Star Trek: TNG.

User: *What is TNG?*



Meena: The Next Generation

Sequence-to-sequence with attention

End-to-end neural conversational model

User: *How are things?*

Meena: They're good. How about you?

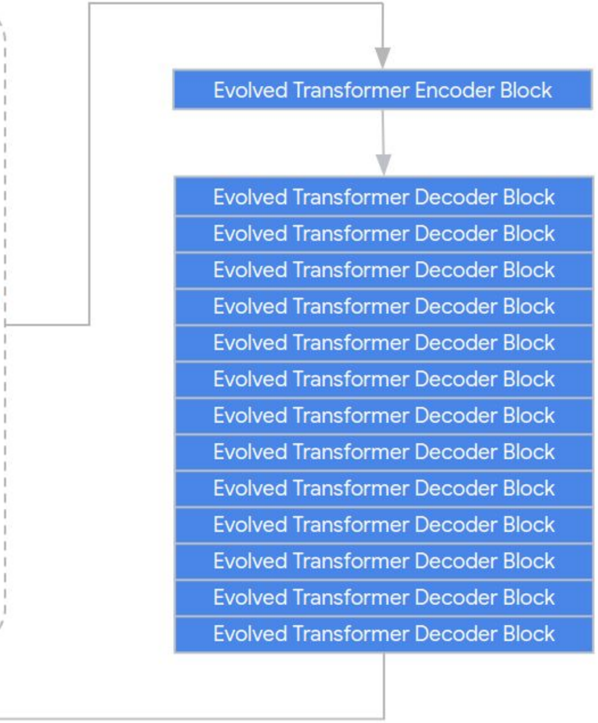
User: *I'm trying to decide what to watch*

Meena: What are some of your favorite shows?

User: *I like the Good Place*

Meena: I haven't watched that. My favorite shows are Supernatural, and Star Trek: TNG.

User: *What is TNG?*



Meena: The Next Generation

[Evolved Transformer] at the core

found by Neural Architecture Search

Better perplexity

The largest conversational model

Meena Scale	2.6 Bn Parameters
GPT2 Scale	1.5 Bn Parameters

1.7x model capacity

The largest conversational model

Meena Scale — 2.6 Bn Parameters — 341 GB conversation text

GPT2 Scale — 1.5 Bn Parameters — 40 GB Internet text

1.7x model capacity

8.5x more data

The Meena Dataset

A curated version of public social media conversations

- **867M** (context, response) pairs or **61B** tokens

Filtered content:

- offensive, repetitive, too-long/short, non-textual.

Evaluation Methodology

Sensibleness & Specificity Average (SSA)

- Our proposed human evaluation metric

<u>Context</u>

Human: Do you know how to swim?

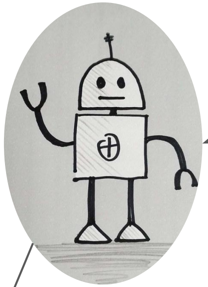
Chatbot: *yes*

Human: What's your favorite stroke?

Google Research

Sensibleness & Specificity Average (SSA)

- Our proposed human evaluation metric



<u>Context</u>

Human: Do you know how to swim?

Chatbot: *yes*

Human: What's your favorite stroke?

<u>Response</u>

Butterfly stroke

Sensible: 1

Specific: 1

Google Research

Sensibleness & Specificity Average (SSA)

- Our proposed human evaluation metric

Context

Human: Do you know how to swim?

Chatbot: *yes*

Human: What's your favorite stroke?

Response

Butterfly stroke

Sensible: 1

Specific: 1

Response

I don't know

Sensible: 1

Specific: 0

Sensibleness & Specificity Average (SSA)

- Our proposed human evaluation metric

<u>Context</u>
Human: Do you know how to swim?
Chatbot: *yes*
Human: What's your favorite stroke?

<u>Response</u>
Butterfly stroke
Sensible: 1
Specific: 1

<u>Response</u>
I don't know
Sensible: 1
Specific: 0

<u>Response</u>
The one that shines
Sensible: 0
Specific: 0 (default)

Sensibleness & Specificity Average (SSA)

- Our proposed human evaluation metric
- Each response rated by 5 crowdworkers
    - majority voting to see if a response is sensible / specific

Google Research

Sensibleness & Specificity Average (SSA)

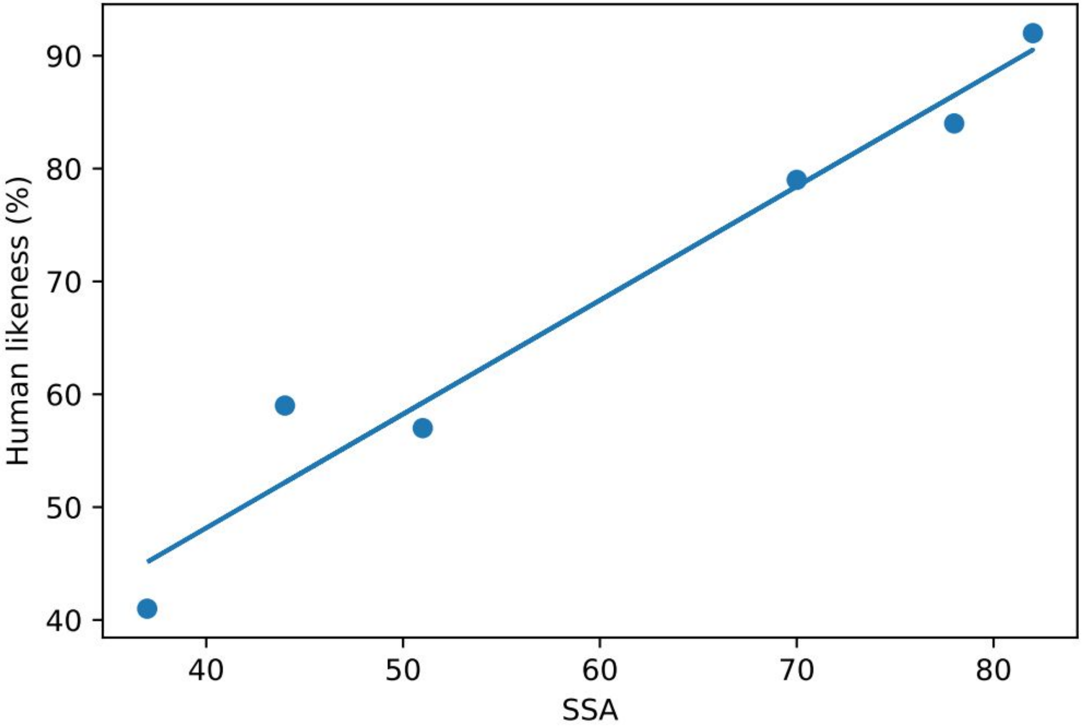
- Our proposed human evaluation metric
- Each response rated by 5 crowdworkers
  - majority voting to see if a response is sensible / specific

Sensibleness: % of responses that are sensible.

Specificity: % of responses that are specific.

SSA = (Sensibleness + Specificity) / 2

Sanity check: SSA correlates with human likeness



Results

Existing chatbots and models

Mitsuku

5-time winner of Turing Test style Loebner Prize

cleverbot

~Oldest bot, 150M conversations

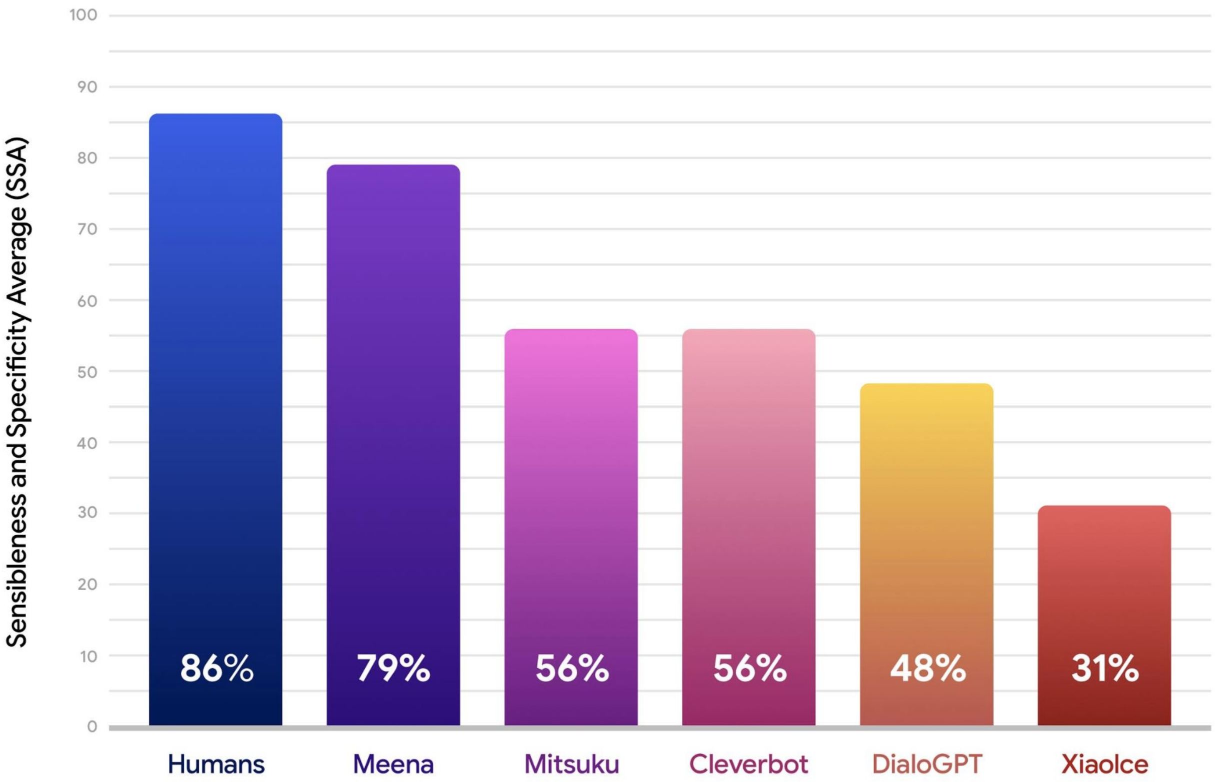
XiaoIce

From Microsoft (660M users)

GPT-2

Many chatbots, e.g., Microsoft DialoGPT

Evaluation of Free-form Chat



14-28 turns / conversation

100 conversations / chatbot

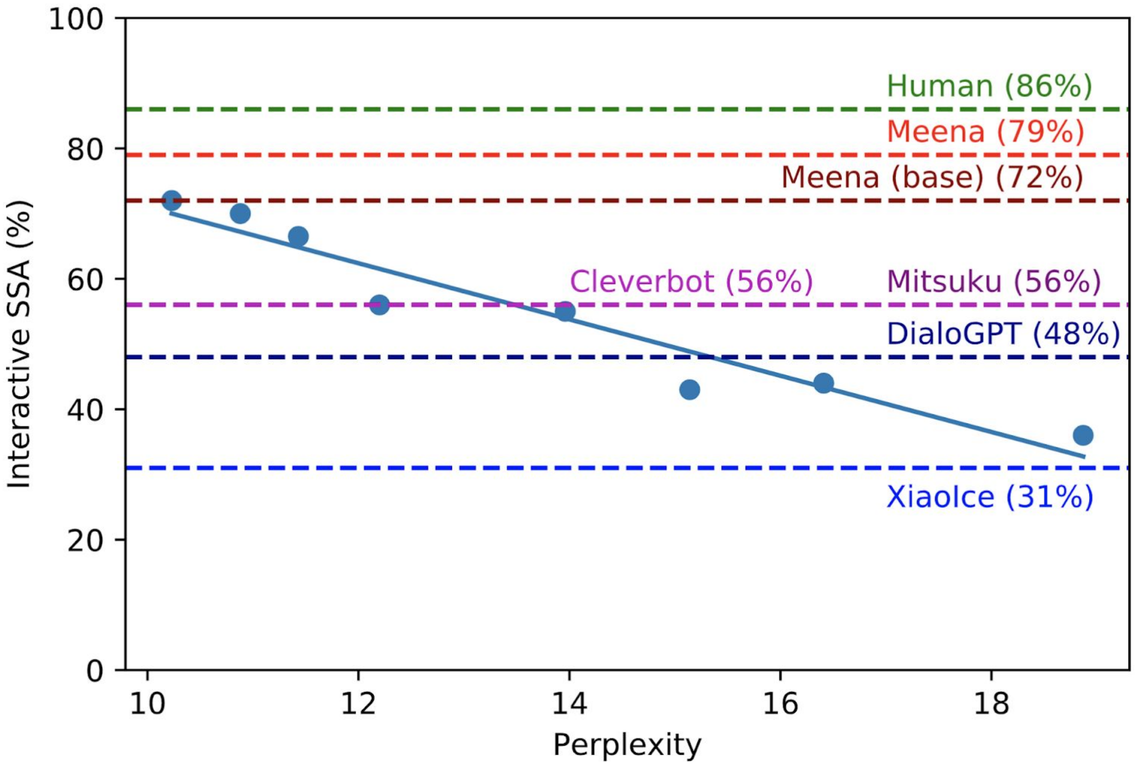
Sample Responses

Input	Meena	Cleverbot	DialoGPT
A: What is your favorite island in the world?	**B:** That's a tough one. I'd have to say Hawaii.	**B:** Daniel I like frites.	**B:** Probably the one on the north island

Sample Responses

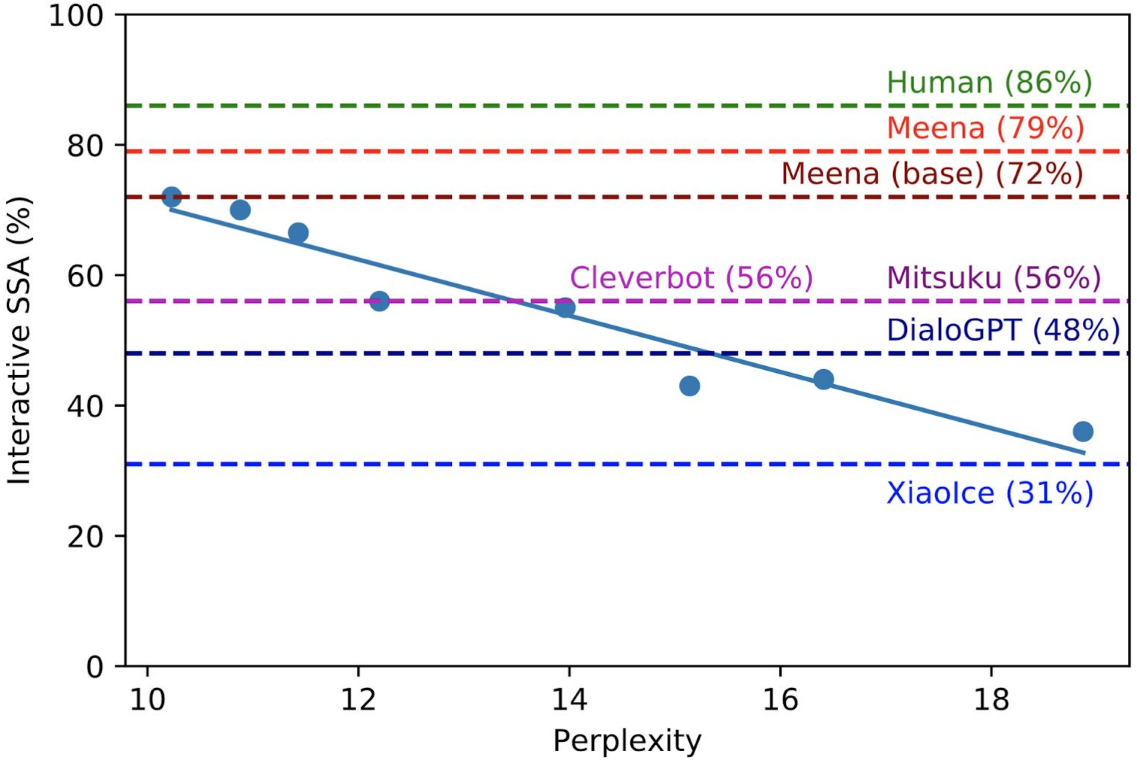
Input	Meena	Cleverbot	DialoGPT
A: What is your favorite island in the world? **B**: I would say Fiji	**A:** Fiji, yes! I've been there once. It's the most beautiful place I've ever been. It definitely makes me want to move there. I hope you get to visit again soon!	**A:** I would have guessed 666.	**A:** Fiji is awesome! But I would say Fiji and Oceania are more popular at the moment.

Perplexity strongly correlates with SSA!



Took us 2 year to "verify" our belief!

Perplexity strongly correlates with SSA!



Pure e2e Meena (base) scores 72%

Took us 2 year to "verify" our belief!

Google Research

Summary

- We're getting closer to a human-like chatbot!
- Perplexity might be "almost" all we need.
- Other aspects: factuality, personality, empathy, & fairness.

Paper: https://arxiv.org/abs/2001.09977
Blog: https://twitter.com/GoogleAI/status/1222230622355087360
Conversation samples:
https://github.com/google-research/google-research/tree/master/meena

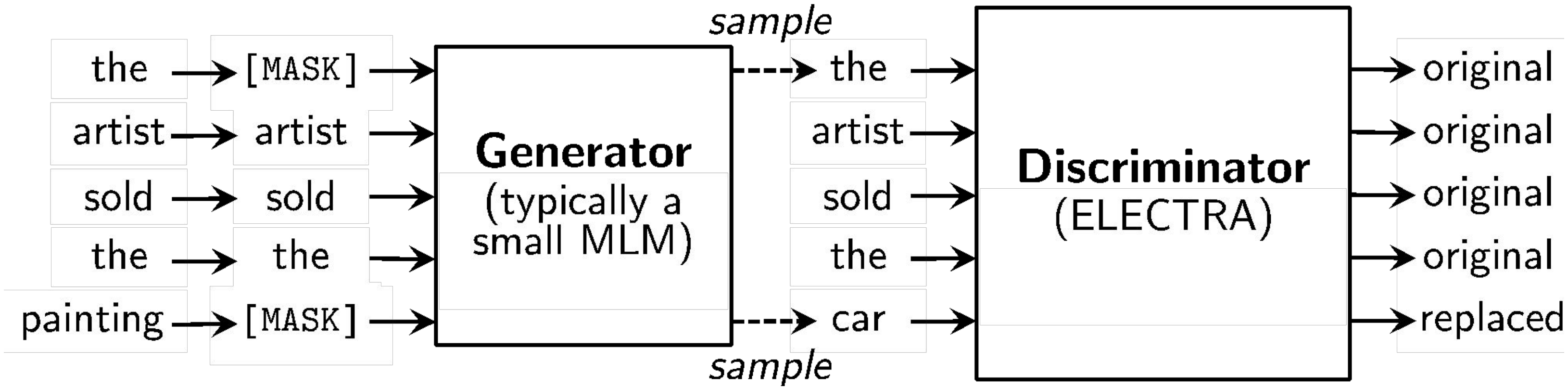
Thank you!

Google Research

ELECTRA: Pre-training Text Encoders as Discriminators rather than Generators

Kevin Clark, Thang Luong, Quoc Le, Chris Manning

ICLR'2020, https://openreview.net/forum?id=r1xMH1BtvB



- Trained in a few days on a single GPU, better accuracy than GPT (30x compute).
- Trained at scale, SOTA results on the SQuAD question answering benchmark.