Machine Learning Department
School of Computer Science
Carnegie Mellon University

VAEs and Variational Inference & In-context Learning

Matt Gormley & Pat Virtue

Lecture 9

Feb. 12, 2025

Reminders

- Quiz 2:
  - In-class: Mon, Feb 17
  - Lectures 5-8 + 9 (but not In-context Learning
- Homework 2: Generative Models of Images
  - Due: Sat, Feb 22 at 11:59pm

Q & A

Q: I just asked a question in OH and now my TA is crying quietly what did I do wrong?

A: You've just committed the worst of crimes: asking a question that was directly answered in a recitation.

The TA you asked spent hours carefully writing careful recitation notes and solutions, practicing their recitation, responding to criticism / changes from instructors, etc.

To increase OH efficiency, please review the HW recitation before asking HW questions in OHs.

VARIATIONAL AUTOENCODERS

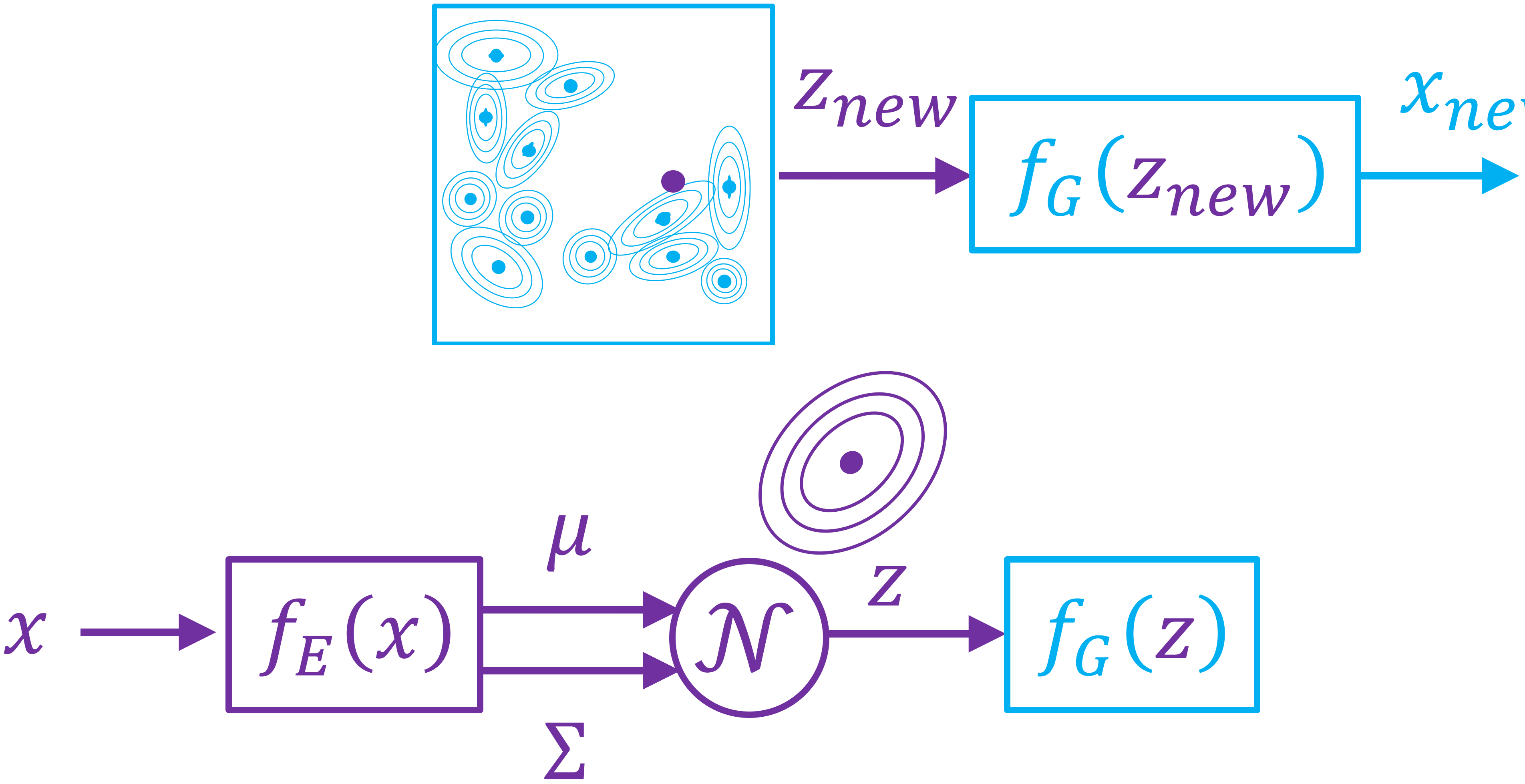
Variational Autoencoders

Last time
- Autoencoders
- Sampling from Autoencoders???
- Autoencoders → Variational Autoencoders

Today
- VAE Reparameterization trick
- VAE Objective
- KL Divergence diversion ☺
- Variational Inverence
  - ELBO (Evidence Lower BOund)

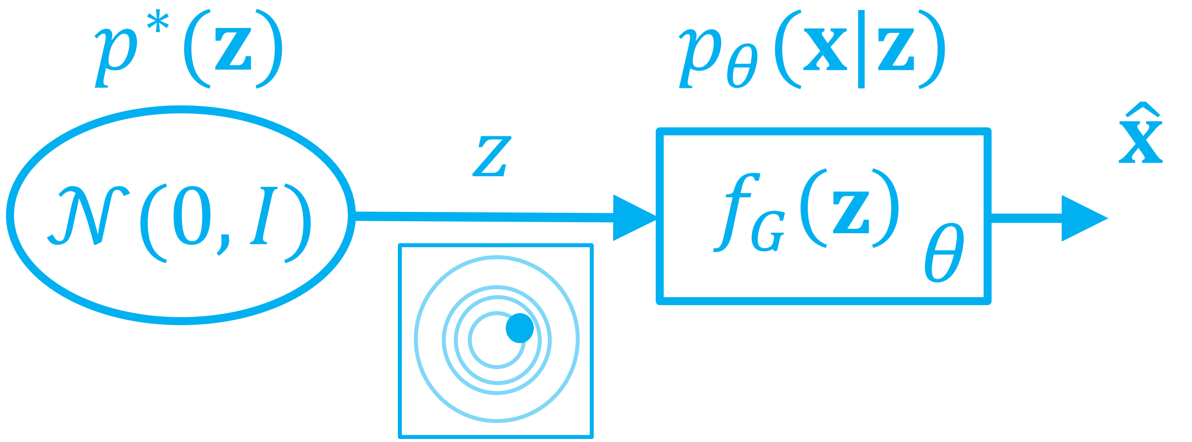
Sampling $p(x)$ from Autoencoder

How do we sample a new image?

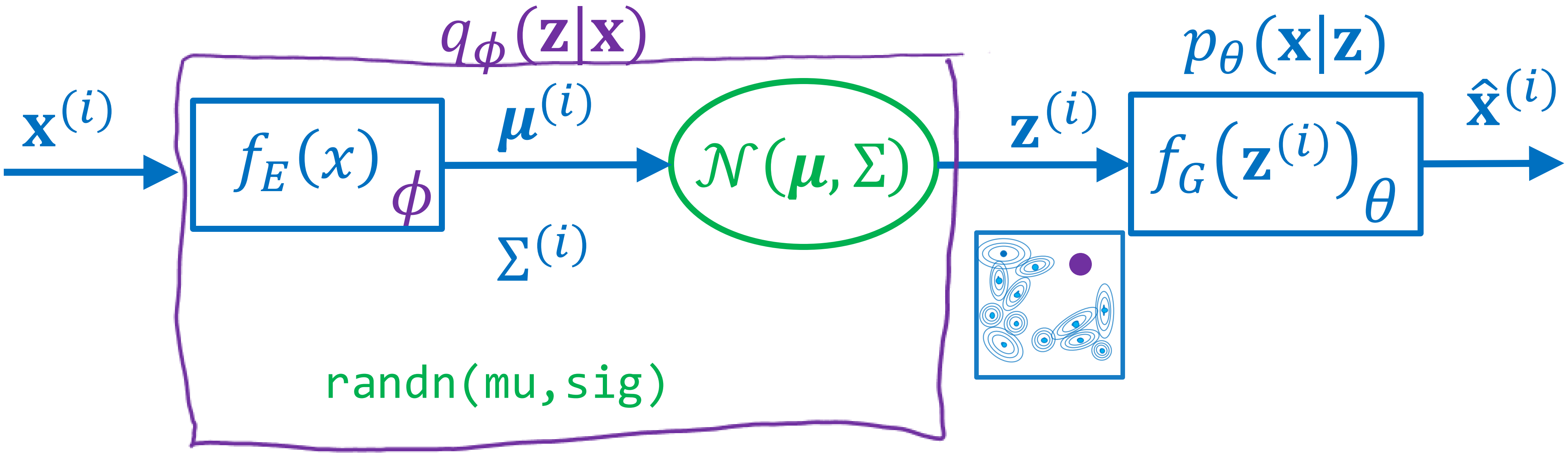


From Autoencoders to VAE

Want:



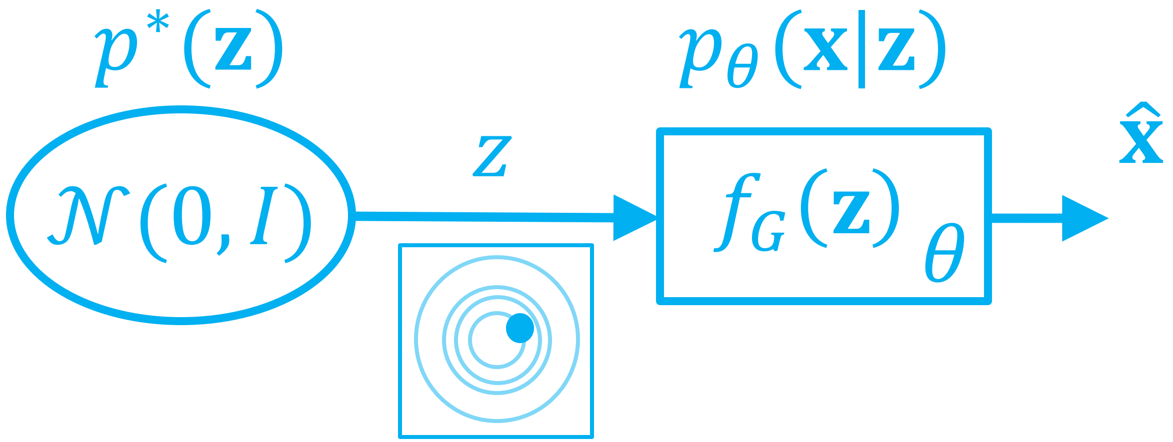
Autoencoder:



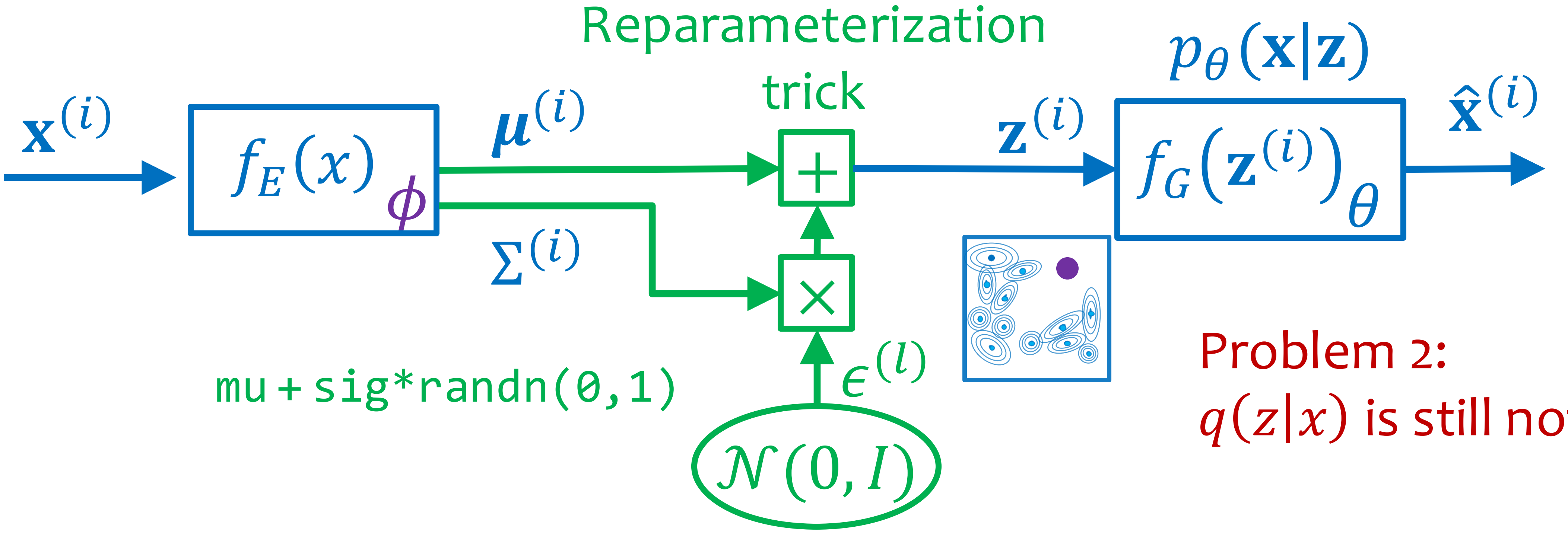
Problem 1: Random sampling function on computation path from ϕ to loss

From Autoencoders to VAE

Want:

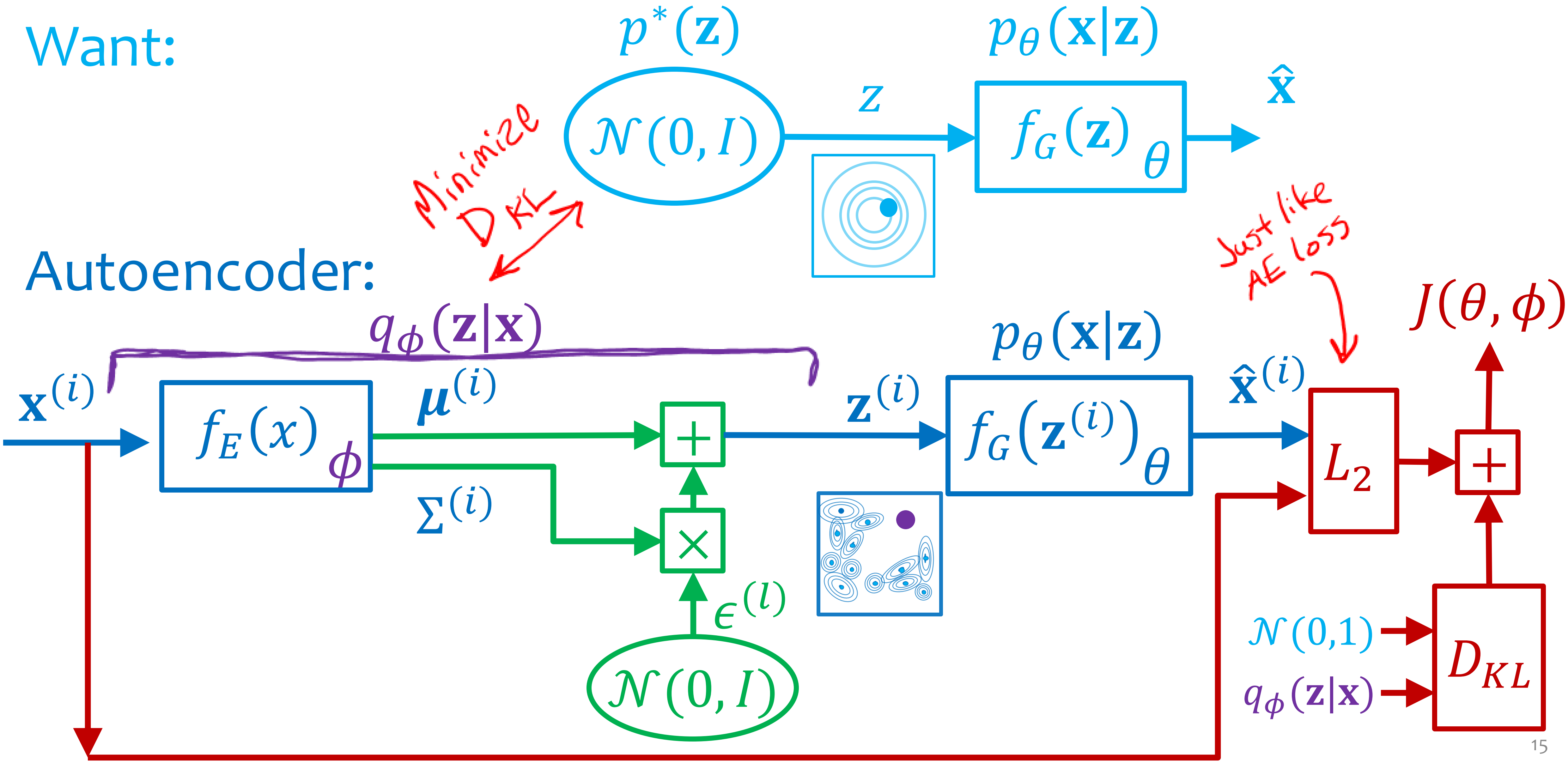


Autoencoder:



Problem 2:
$q(z|x)$ is still not $\mathcal{N}(0, I)$

From Autoencoders to VAE



VAE Objective

MLE:

$$\underset{\theta,\phi}{\operatorname{argmax}} \ \log p(x)$$

$$\log \int_z p_\theta(x \mid z)\, p(z)\, dz$$

Multiply by 1

$$\log \int_z p_\theta(x \mid z)\, p(z) \frac{q_\phi(z \mid x)}{q_\phi(z \mid x)} dz$$

$$\log \mathbb{E}_{z\sim q(z|x)}\left[p_\theta(x \mid z) \frac{p(z)}{q_\phi(z \mid x)}\right]$$

Jensen's inequality:

For convex $f(x)$:

$$f(a+b) \le f(a) + f(b)$$

$$-\log(a+b) \le -\log a - \log b$$

$$\log(a+b) \ge \log a + \log b$$

$$\mathbb{E}_{z\sim q(z|x)}\left[\log\left(p_\theta(x \mid z) \frac{p(z)}{q_\phi(z \mid x)}\right)\right]$$

Evidence Lower bound (ELBO)

$$\mathbb{E}_{z\sim q(z|x)}[\log p_\theta(x \mid z)] + \mathbb{E}_{z\sim q(z|x)}\left[\log \frac{p(z)}{q_\phi(z \mid x)}\right]$$

$$\mathbb{E}_{z\sim q(z|x)}[\log p_\theta(x \mid z)] - D_{KL}(q_\phi(z \mid x) \parallel p(z))$$

Regular Autoencoder Objective

MLE:

$$\underset{\theta,\phi}{\operatorname{argmin}} - \log p(\mathcal{D} \mid \theta, \phi)$$

Yeah, def cheating

$$- \sum_{i=1}^{N} \log p\left(\mathbf{x}^{(i)} \mid \mathbf{x}^{(i)}\right)$$

$$- \sum_{i=1}^{N} \sum_{j}^{M} \log p\left(x_j^{(i)} \mid \mathbf{x}^{(i)}\right)$$

$$- \sum_{i=1}^{N} \sum_{j}^{M} \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_j - \hat{x}_j)^2}{2\sigma^2}}$$

$$\sum_{i=1}^{N} \sum_{j}^{M} \left(x_j - \hat{x}_j\right)^2$$

$$\hat{\mathbf{x}} = h_{\theta,\phi}(\mathbf{x}) = f_G(\mathbf{z}; \theta)$$

$$\mathbf{z} = f_E(\mathbf{x}; \phi)$$

Regression: model each pixel as noisy samples of $h(\mathbf{x})$

$$p(\mathbf{x} \mid \mathbf{x}) = \prod_j p(x_j \mid \mathbf{x})$$

$$p(x_j \mid \mathbf{x}) \sim N(x_j, \sigma^2)$$

$$loss(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \left\|\mathbf{x} - f_G\left(f_E(\mathbf{x})\right)\right\|_2^2$$

VAE Objective

MLE:

$p^*(z)$

$$\underset{\theta,\phi}{\operatorname{argmin}} -\log p(\mathcal{D} \mid \theta, \phi)$$

$$-\sum_{i=1}^{N} \left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})} \left[\log p_\theta\left(\mathbf{x}^{(i)} \mid \mathbf{z}^{(i)}\right) \right] - D_{KL}\left(q_\phi\left(\mathbf{z}^{(i)} \mid \mathbf{x}^{(i)}\right) \parallel \mathcal{N}(0,1) \right) \right)$$

$$\sum_{i=1}^{N} \left(\frac{1}{L} \sum_{l=1}^{L} -\log p_\theta\left(\mathbf{x}^{(i)} \mid \mathbf{z}^{(i,l)}\right) \quad + D_{KL}\left(q_\phi\left(\mathbf{z}^{(i)} \mid \mathbf{x}^{(i)}\right) \parallel \mathcal{N}(0,1) \right) \right)$$

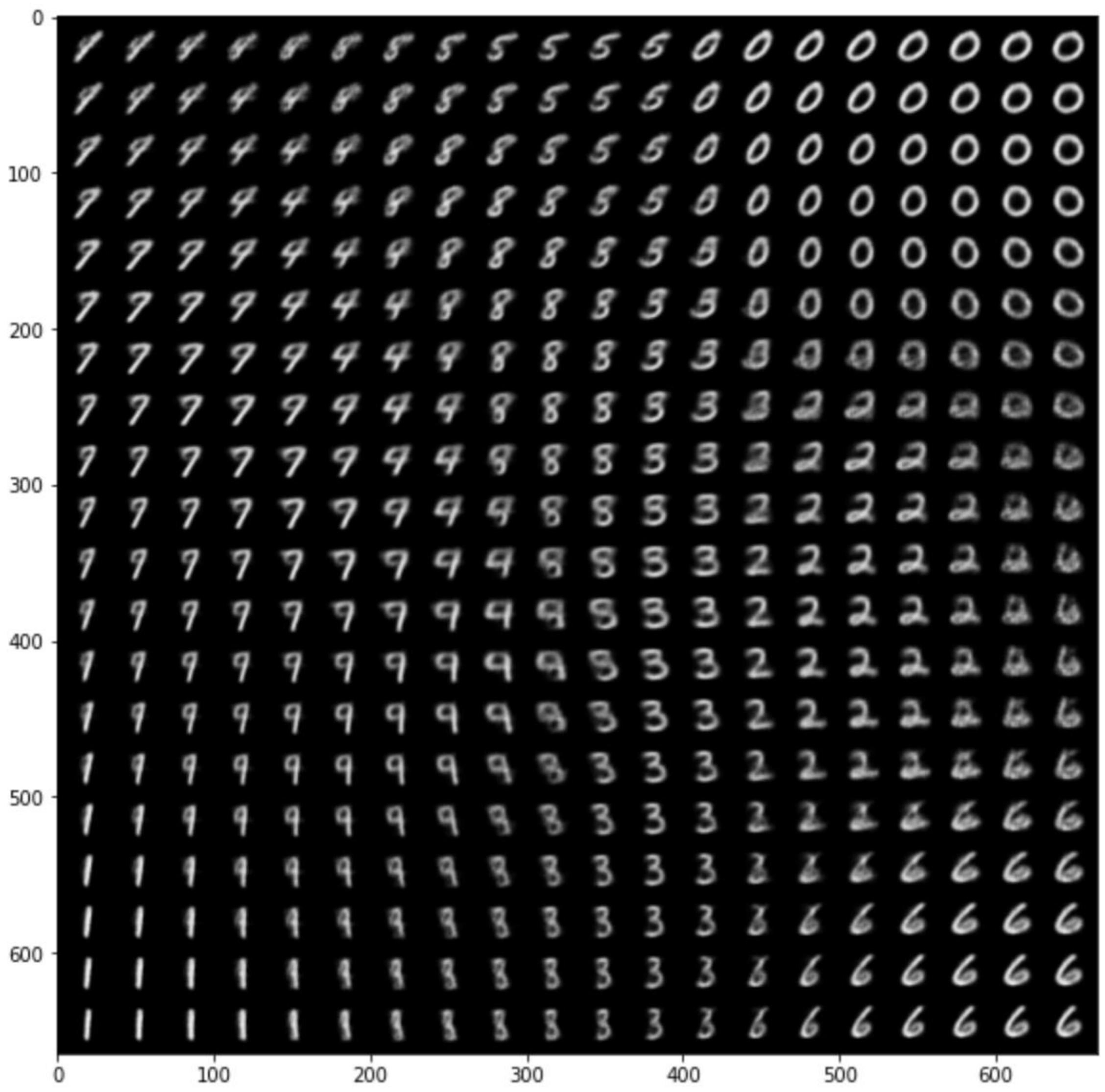
↓ see AE obj.

$$\sum_{i=1}^{N} \left(\frac{1}{L} \sum_{l=1}^{L} \left\| \mathbf{x}^{(i)} - f_G\left(\mathbf{z}^{(i,l)}; \theta\right) \right\|_2^2 \quad + D_{KL}\left(q_\phi\left(\mathbf{z}^{(i)} \mid \mathbf{x}^{(i)}\right) \parallel \mathcal{N}(0,1) \right) \right)$$

where we sample $\mathbf{z}^{(i,l)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(i)}, \Sigma^{(i)}\right)$ and $\boldsymbol{\mu}^{(i)}, \Sigma^{(i)} = f_E\left(\mathbf{x}^{(i)}; \phi\right)$

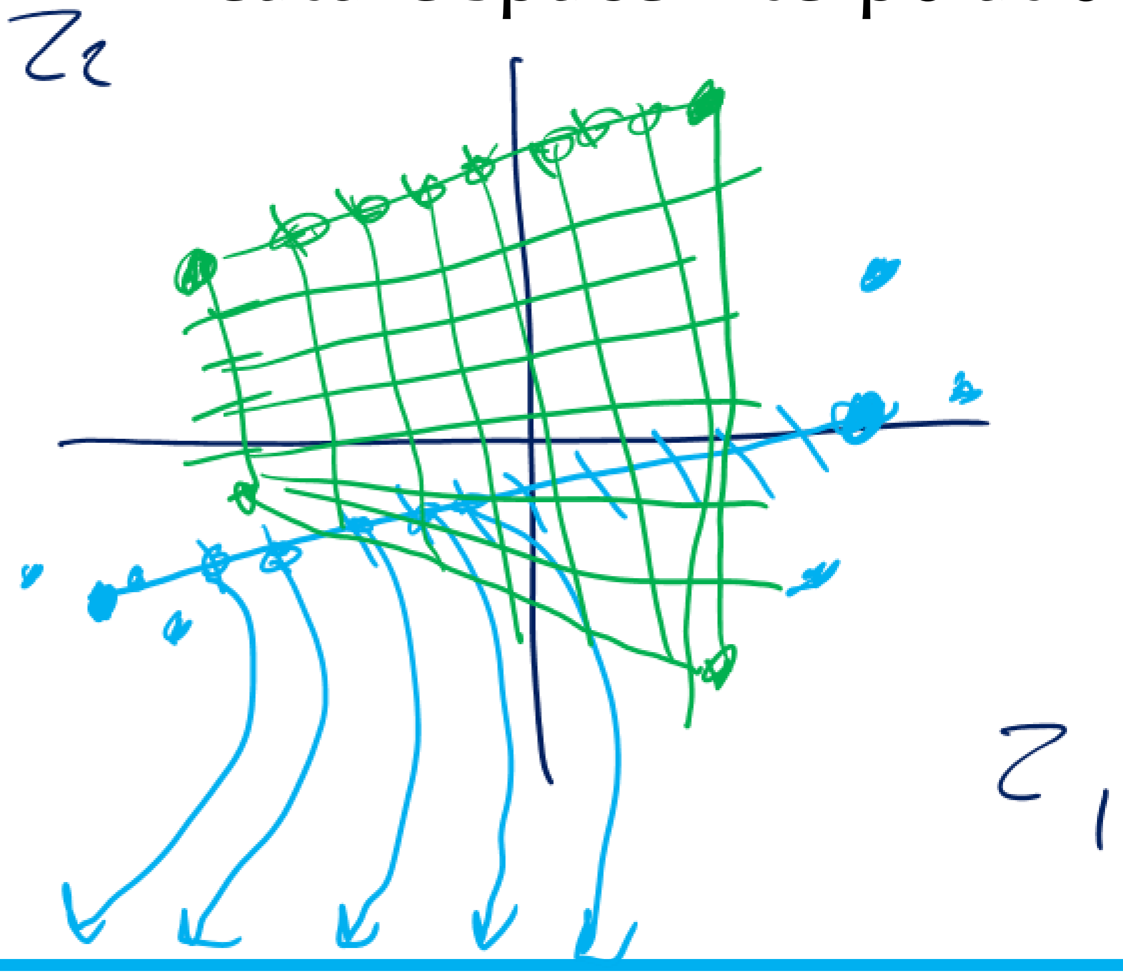
VAE Demo

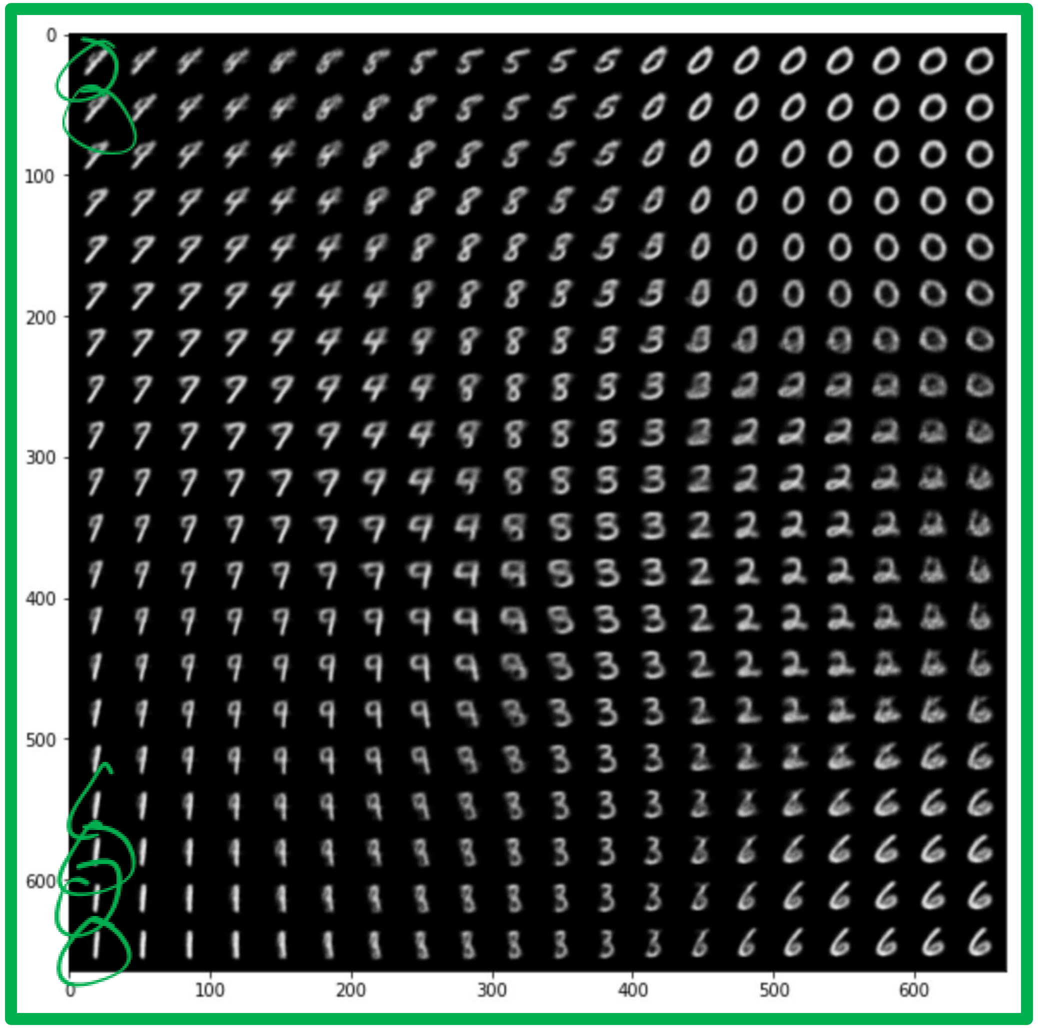
- Zhuoyue Lyu, Safinah Ali, and Cynthia Breazeal. EAAI 2022.
- https://colab.research.google.com/gist/ZhuoyueLyu/5046225a9ae3675cf633c1df5f63be06/digits-interpolation-notebook-eaai.ipynb



VAE Demo

- Feature space interpolation





Recall…

KL DIVERGENCE

KL Divergence

Recall...

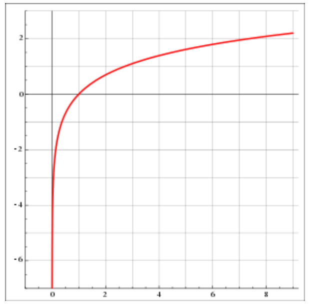
- *Definition*: for two distributions q(x) and p(x) over x ∈ $\mathcal{X}$, the **KL Divergence** is:

$$\text{KL}(q||p) = E_{q(x)}\left[\log \frac{q(x)}{p(x)}\right] = \begin{cases} \sum_x q(x) \log \frac{q(x)}{p(x)} \\ \int_x q(x) \log \frac{q(x)}{p(x)} dx \end{cases}$$

- *Properties*:
  – KL(q || p) measures the **proximity** of two distributions q and p
  – KL is **not** symmetric: KL(q || p) ≠ KL(p || q)
  – KL is minimized when q(x) = p(x) for all x ∈ $\mathcal{X}$

$$\mathrm{KL}(q||p) = E_{q(x)}\left[\log \frac{q(x)}{p(x)}\right]$$

KL Divergence



Understanding the Behavior of KL as an objective function

Example 1: Keeping all else constant, consider the effect of a particular x' on KL(q || p)

x'	q(x')	p(x')	q(x') log(q(x')/p(x'))	effect on KL(q || p)
1	0.9	0.9	0	no increase
2	0.9	0.1	1.97	big increase
3	0.1	0.9	-0.21	little decrease
4	0.1	0.1	0	no increase

KL **does** insist on good approximations for values that have **high** probability in q

KL **does not** insist on good approximations for values that have **low** probability in q

Example 2: Which q distribution minimizes KL(q || p)?

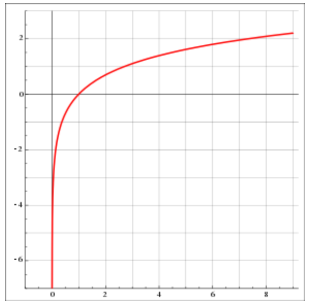
$$\mathbf{p} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix} \quad \mathbf{q}^{(1)} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \quad \mathbf{q}^{(2)} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix} \quad \mathbf{q}^{(3)} = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}$$

Q: If we're minimizing KL, why not return $q^{(3)}$?

A: Because it's not a distribution!

$\text{KL}(q||p) = E_{q(x)}\left[\log \frac{q(x)}{p(x)}\right]$

KL Divergence

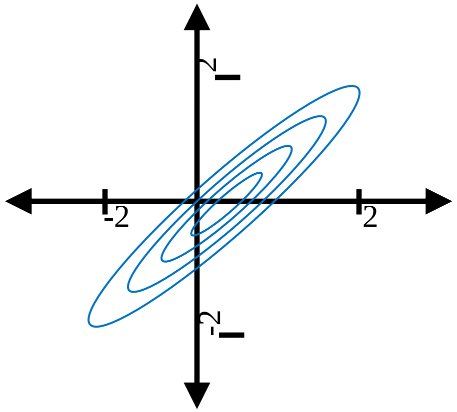


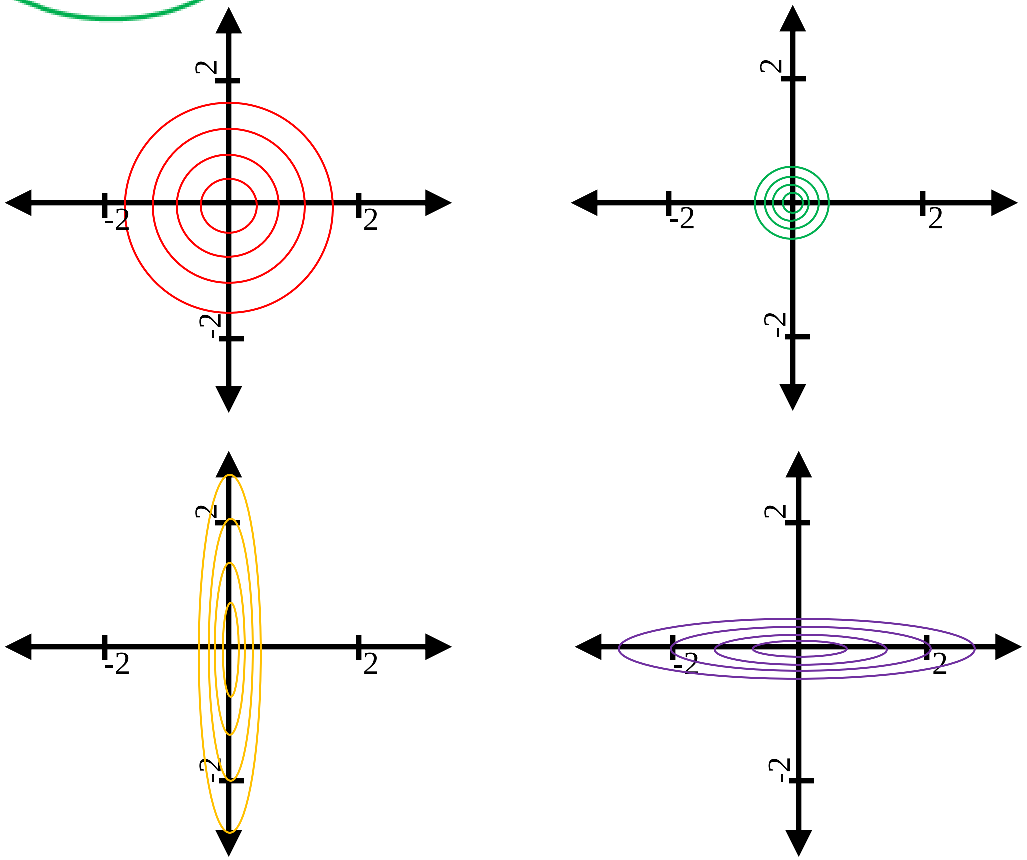
Understanding the Behavior of KL as an objective function

Example 3: Which q distribution minimizes KL(q || p)?

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu} = [0,0]^T, \boldsymbol{\Sigma})$$

$$q(x_1, x_2) = \mathcal{N}_1(x_1 \mid \mu_1, \sigma_1^2)\mathcal{N}_2(x_2 \mid \mu_2, \sigma_2^2)$$





Optimizing KL Divergence

- *Question*: How do we minimize KL?

$$\hat{\theta} = \underset{\theta \in \Theta}{\text{argmin}}\, \text{KL}(q_\theta(\mathbf{z}) \,\|\, p_\alpha(\mathbf{z} \mid \mathbf{x}))$$

Want our encoding to be the same as true

Pardon the change of var.
$\theta \to \alpha$
$\phi \to \theta$

Also for VAE we used $q(z|x)$ as our $q(z)$ (equivalent)

- *Answer #1*: Oh no! We can't even compute this KL.

Why we can't compute KL…

$$p(z|x) = \frac{p(x,z)}{p(x)}$$

$$\begin{aligned}
\text{KL}(q(\mathbf{z}) \,\|\, p(\mathbf{z} \mid \mathbf{x})) &= E_{q(\mathbf{z})}\left[\log\left(\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})}\right)\right] \\
&= E_{q(\mathbf{z})}[\log q(\mathbf{z})] - E_{q(\mathbf{z})}[\log p(\mathbf{z} \mid \mathbf{x})] \\
&= E_{q(\mathbf{z})}[\log q(\mathbf{z})] - E_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z})] + E_{q(\mathbf{z})}[\log p(\mathbf{x})] \\
&= E_{q(\mathbf{z})}[\log q(\mathbf{z})] - E_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z})] + \log p(\mathbf{x})
\end{aligned}$$

not dep. on z

we have the same problem with an intractable data likelihood p(x)

we assumed this is intractable to compute!

Optimizing KL Divergence

- *Question*: How do we minimize KL?

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \, \text{KL}(q_\theta(\mathbf{z}) \parallel p_\alpha(\mathbf{z} \mid \mathbf{x}))$$

- *Answer #2*: We don't need to compute this KL
  We can instead maximize the ELBO (i.e. **E**vidence **L**ower **BO**und)

$$\text{ELBO}(q_\theta) = E_{q_\theta(\mathbf{z})}[\log p_\alpha(\mathbf{x}, \mathbf{z})] - E_{q_\theta(\mathbf{z})}[\log q_\theta(\mathbf{z})]$$

Here is why…

$$\begin{aligned}
\theta &= \underset{\theta}{\operatorname{argmin}} \, \text{KL}(q_\theta(\mathbf{z}) \parallel p_\alpha(\mathbf{z} \mid \mathbf{x})) \\
&= \underset{\theta}{\operatorname{argmin}} \, E_{q_\theta(\mathbf{z})}[\log q_\theta(\mathbf{z})] - E_{q_\theta(\mathbf{z})}[\log p_\alpha(\mathbf{x}, \mathbf{z})] + \log p_\alpha(\mathbf{x}) \\
&= \underset{\theta}{\operatorname{argmin}} \, E_{q_\theta(\mathbf{z})}[\log q_\theta(\mathbf{z})] - E_{q_\theta(\mathbf{z})}[\log p_\alpha(\mathbf{x}, \mathbf{z})] \\
&= \underset{\theta}{\operatorname{argmax}} \, \text{ELBO}(q_\theta)
\end{aligned}$$

dropping the intractable term gives the ELBO

ELBO as Objective Function

What does maximizing ELBO(q_θ) accomplish?

$$\text{ELBO}(q_\theta) = E_{q_\theta(\mathbf{z})}\left[\log p_\alpha(\mathbf{x}, \mathbf{z})\right] - E_{q_\theta(\mathbf{z})}\left[\log q_\theta(\mathbf{z})\right]$$

1. The first expectation is high if q_θ puts probability mass on the same values of **z** that p_α puts probability mass

2. The second term is the entropy of q_θ and the entropy will be high if q_θ spreads its probability mass evenly

ELBO's relation to log p(x)

Theorem:

Proof #1:

Proof #2:

Key Takeaway:

ELBO's relation to log p(x)

Theorem:

for any q, $\log p(x) \geq ELBO(q)$

i.e. $ELBO(q)$ is a lower bound on $\log p(x)$

Proof #1:

Recall Jensen's Inequality: $f(E[x]) \geq E[f(x)]$, for concave f

$\log p(x) = \log \int_z p(x,z)\, dz$ (marginal)

$= \log \int_z p(x,z) \frac{q(z)}{q(z)}\, dz$ (mult. by 1)

$= \log E_{q(z)}[p(x,z)/q(z)]$ (def. of expectation)

$\geq E_{q(z)}[\log(\frac{p(x,z)}{q(z)})]$ (by Jensen's Ineq.)

$= E_{q(z)}[\log p(x,z)] - E_{q(z)}[\log q(z)] = ELBO(q)$

$\Rightarrow \log p(x) \geq ELBO(q)$

Proof #2:

① $\log p(x) = KL(q\|p) + ELBO(q)$

② $KL(q\|p) \geq 0$ (without proof)

③ $\Rightarrow \log p(x) \geq ELBO(q)$

Key Takeaway:

minimizing KL is the same as finding a tight ELBO(q) lower bound for $\log p(x)$

IN-CONTEXT LEARNING

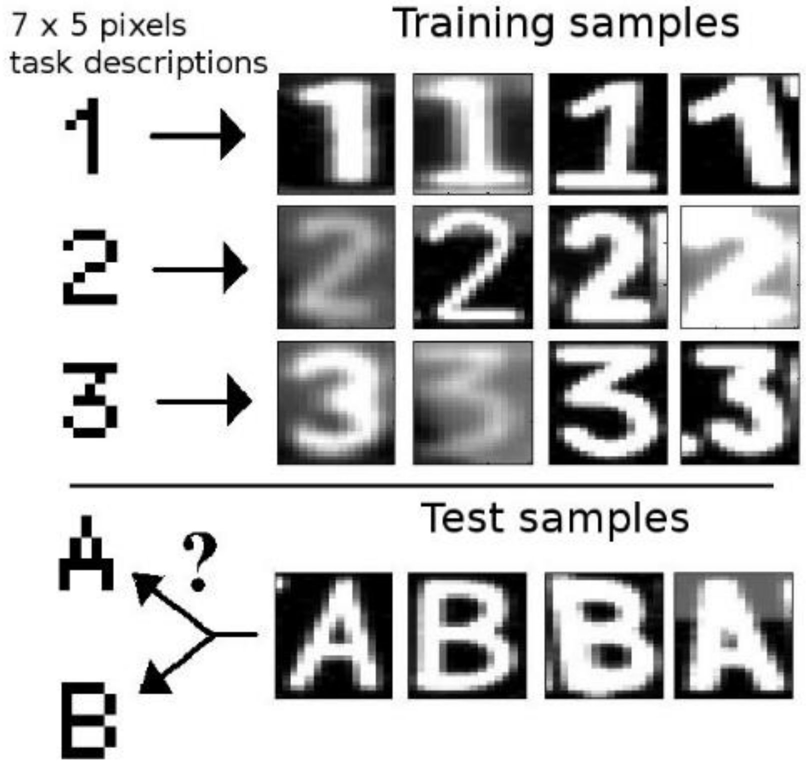
ZERO-SHOT AND FEW-SHOT LEARNING

Zero-shot Learning

- ***Definition***: in **zero-shot learning** we assume that training data does not contain any examples of the labels that appear in the test data

$z \rightarrow$	1	2	3	4	5
$d(z) \rightarrow$	1	2	3	A	B
x_t	y_t^1	y_t^2	y_t^3	y_t^4	y_t^5
1	1	0	0	-	-
2	0	1	0	-	-
3	0	0	1	-	-
A	-	-	-	1	0
B	-	-	-	0	1

training data (y^1–y^3) test data (y^4–y^5)



Question:
How can we hope to learn in a setting where have no labeled examples?

Answer:

Figure from https://cdn.aaai.org/AAAI/2008/AAAI08-103.pdf

Few-shot Learning

- ***Definition***: in **few-shot learning** we assume that training data contains a handful (maybe two, three, or four) examples of each label

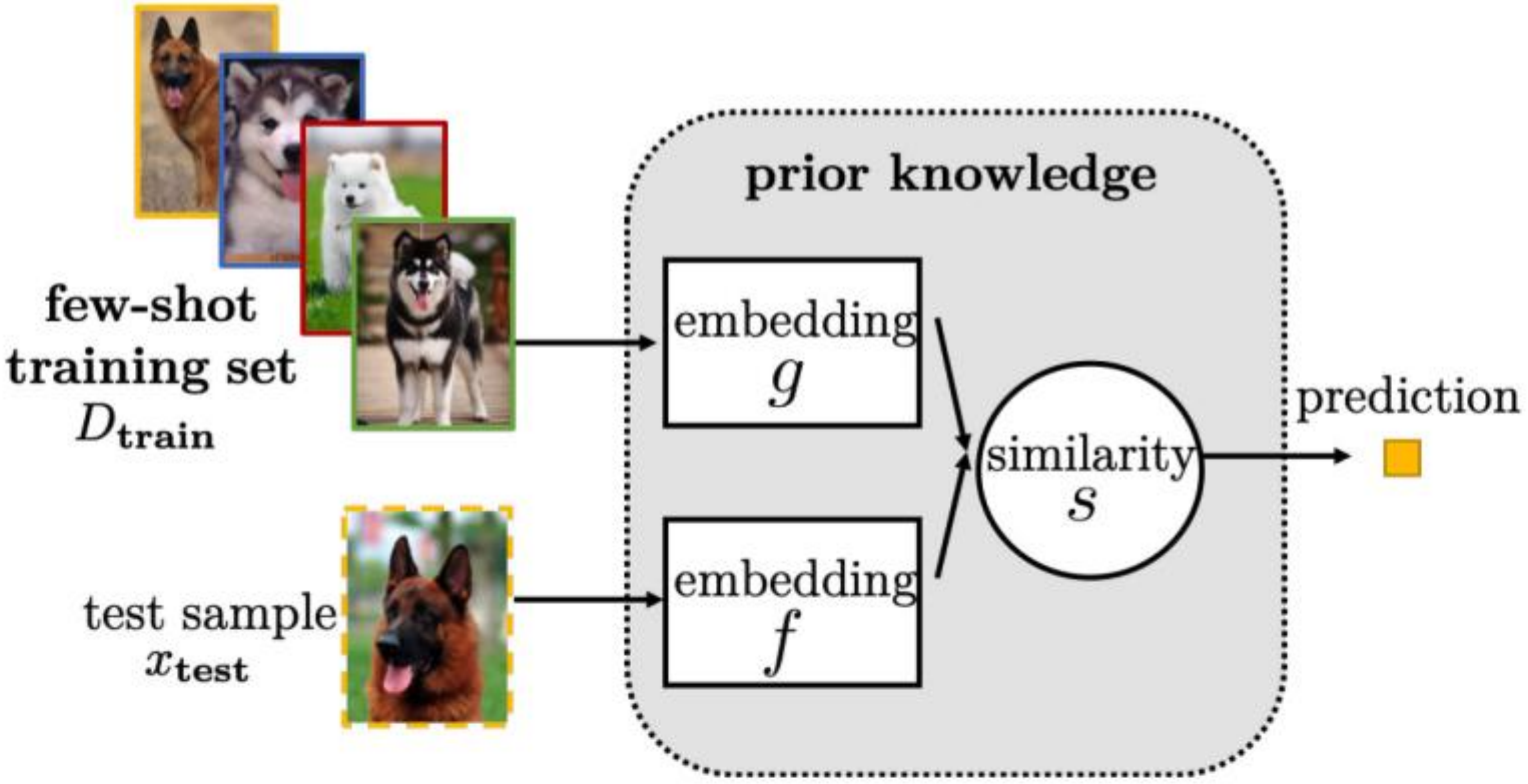


Figure from https://dl.acm.org/doi/10.1145/3386252

Few-shot Learning

- ***Definition***: in **few-shot learning** we assume that training data contains a handful (maybe two, three, or four) examples of each label

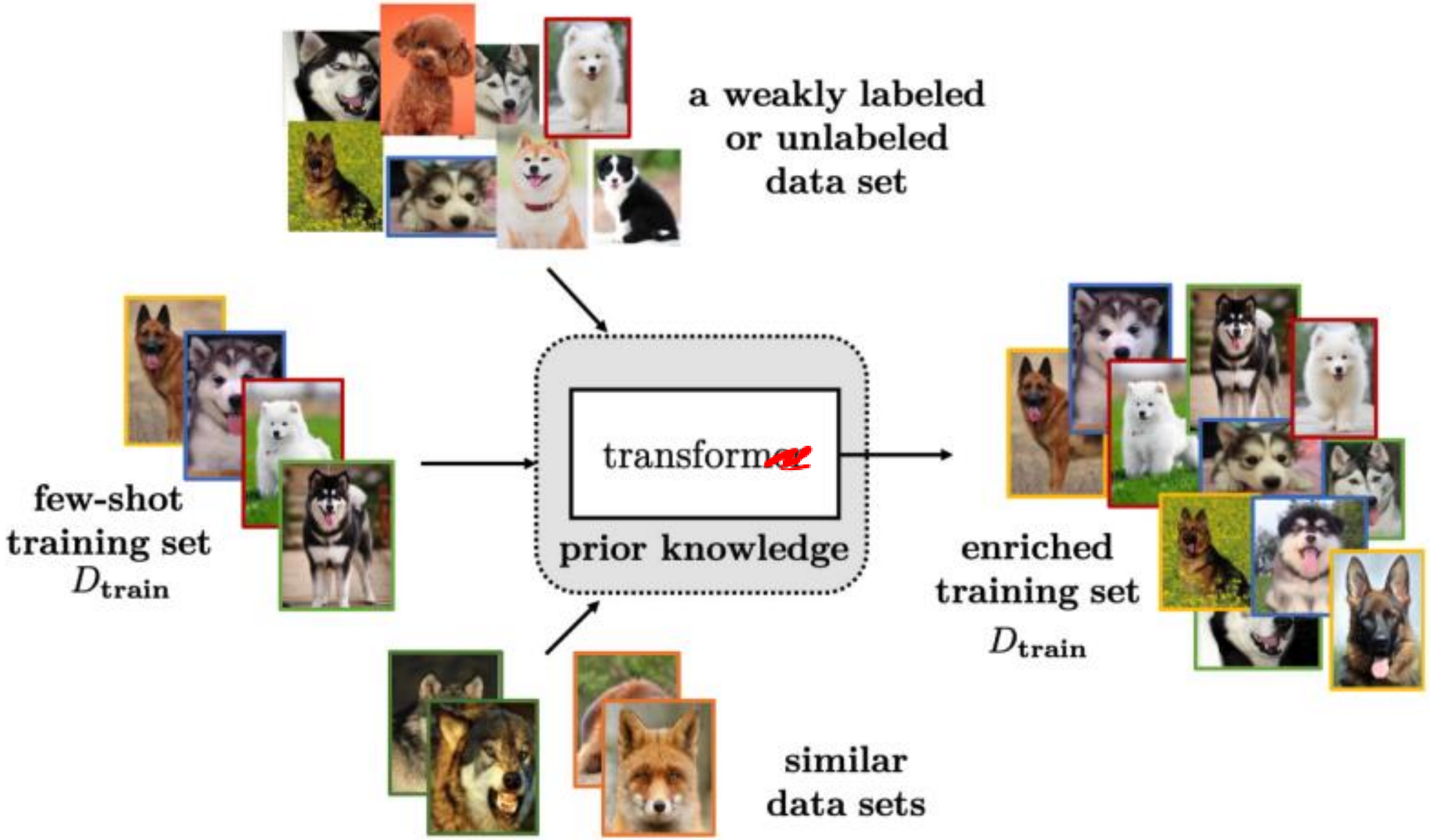


Figure from https://dl.acm.org/doi/10.1145/3386252

PROMPTING

Prompting

- Language models are trained to maximize the likelihood of sequences in the training data
- Most condition on the previous tokens to generate the next tokens

Definition: An *autoregressive language model* defines a probability distribution over sequences $\mathbf{x}_{1:T}$ of the form:

$$p(\mathbf{x}_{1:T}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

- Key idea behind prompting: provide a prefix string to the prompt such that its likely completion is the answer you want text completion

Prompting Examples

GPT-3

Prompt (i.e. context)

```
The City
BY C. P. CAVAFY
TRANSLATED BY EDMUND KEELEY
[Poem text omitted]

SOME TREES
John Ashbery
[Poem text omitted]

Shadows on the Way
Wallace Stevens
```

Output (i.e. generated poem)

```
I must have shadows on the way
If I am to walk I must have
Each step taken slowly and alone
To have it ready made

And I must think in lines of grey
To have dim thoughts to be my guide
Must look on blue and green
And never let my eye forget
That color is my friend
And purple must surround me too

The yellow of the sun is no more
Intrusive than the bluish snow
That falls on all of us.  I must have
Grey thoughts and blue thoughts walk with me
If I am to go away at all.
```

Conditional Sampling:

To draw a sample, feed in a prompt (i.e. context) and then sample from the model conditioned on that prompt

Examples from GPT-3 paper

translation

Prompting Examples

Llama-2-70B

Prompt

El último dueño de esta propiedad había sido un hombre soltero, que alcanzó una muy avanzada edad, y que durante gran parte de su existencia tuvo en su hermana una fiel compañera y ama de casa.

English translation:

Conditional Sampling:
To draw a sample, feed in a prompt (i.e. context) and then sample from the model conditioned on that prompt

Model Output

The last owner of this property was a bachelor man who lived to a very advanced age, and who had his sister as a faithful companion and housekeeper for most of his life.

English source from Sense & Sensibility

The late owner of this estate was a single man, who lived to a very advanced age, and who for many years of his life, had a constant companion and housekeeper in his sister.

Text from https://huggingface.co/datasets/opus_books/viewer/en-es question answering (entailment)

Prompting Examples

Llama-2-70B

Prompt

Rhodochiton is a genus of flowering plants within the family Plantaginaceae, native to southern Mexico and neighbouring Guatemala. They climb by means of twining leaf stalks. One of the three species, "Rhodochiton atrosanguineus", the purple bell vine, is grown as an ornamental plant. All three species are sometimes included in "Lophospermum".

Question: You can find the purple bell vine in more than one country. True, False, or Neither?

Model Output

Answer: True.

Explanation: The purple bell vine (Rhodochiton atrosanguineus) is native to southern Mexico and neighbouring Guatemala, which means it can be found in more than one country.

Correct Answer

True.

Text from https://huggingface.co/datasets/facebook/anli summarization

Prompting Examples

Llama-2-70B

Prompt

Story: Jason listened to the weather and heard it was goingto be sunny. He thought the kids might like to goswimming. He gathered up the swimsuits, towelsand sunscreen. Jason and the kids got into the truckand drove to the beach. They spent the next 2 hoursplaying and splashing in the surf.

One-sentence Summary:

Model Output

Jason took the kids swimming at the beach after hearing the weather forecast, gathering necessary items and driving there.

Story text from https://arxiv.org/pdf/1910.00998.pdf

What else can a large LM (attempt to) do?

Using the idea of **prompts**, we can apply LMs to a **variety of different problems** in natural language processing.

In the **zero-shot setting**, we simply feed the context to the model and observe how it completes the sequence. (i.e. there is no additional training)

Answer fact-based questions:

Context →	Organisms require energy in order to do what?
Correct Answer →	mature and develop.
Incorrect Answer →	rest soundly.
Incorrect Answer →	absorb light.
Incorrect Answer →	take in nutrients.

Complete sentences logically:

Context →	My body cast a shadow over the grass because
Correct Answer →	the sun was rising.
Incorrect Answer →	the grass was cut.

Complete analogies:

Context →	lull is to trust as
Correct Answer →	cajole is to compliance
Incorrect Answer →	balk is to fortitude
Incorrect Answer →	betray is to loyalty
Incorrect Answer →	hinder is to destination
Incorrect Answer →	soothe is to passion

Reading comprehension:

Context →	anli 1: anli 1: Fulton James MacGregor MSP is a Scottish politician who is a Scottish National Party (SNP) Member of Scottish Parliament for the constituency of Coatbridge and Chryston. MacGregor is currently Parliamentary Liaison Officer to Shona Robison, Cabinet Secretary for Health & Sport. He also serves on the Justice and Education & Skills committees in the Scottish Parliament. Question: Fulton James MacGregor is a Scottish politican who is a Liaison officer to Shona Robison who he swears is his best friend. True, False, or Neither?
Correct Answer →	Neither
Incorrect Answer →	True
Incorrect Answer →	False

Examples from GPT-3

Zero-shot LLMs

- GPT-2 (1.5B parameters) for unsupervised prediction on various tasks
- GPT-2 models p(output | input, task)
  - translation: *(translate to french, english text, french text)*
  - reading comprehension: *(answer the question, document, question, answer)*
- Why does this work?

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**
In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”
“I hate the word ‘**perfume**,”’ Burr says. ‘It’s somewhat better in French: ‘**parfum**.’
If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.
If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?
“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

Figures from Radford et al. (2019)

Zero-shot LLMs

- GPT-2 (1.5B parameters) for unsupervised prediction on various tasks
- GPT-2 models p(output | input, task)
  - translation: *(translate to french, english text, french text)*
  - reading comprehension: *(answer the question, document, question, answer)*
- Why does this work?

Language Models are Unsupervised Multitask Learners

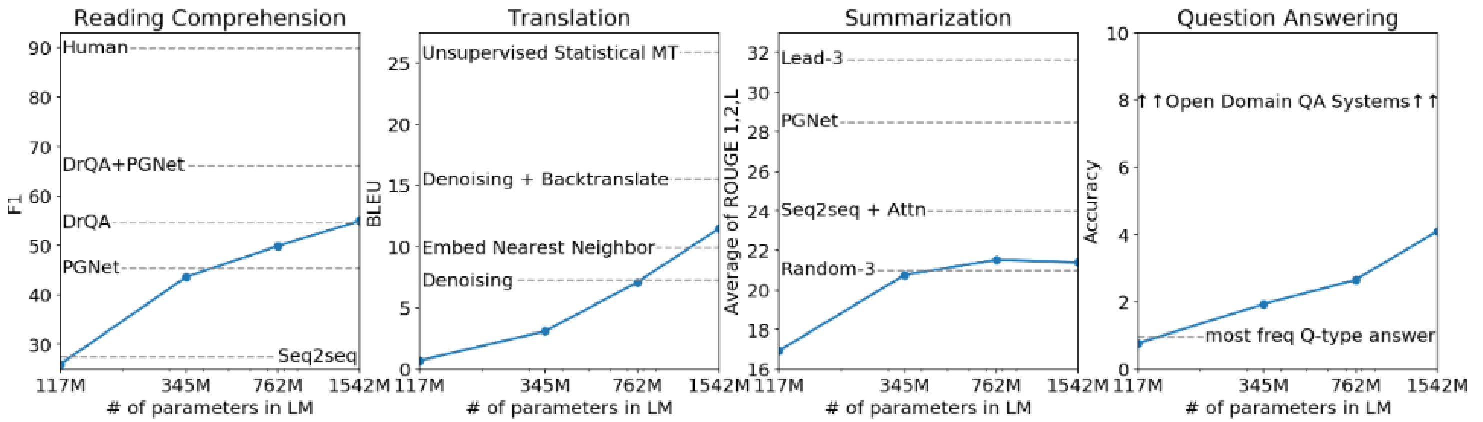


Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural Questions (Kwiatkowski et al., 2019). Section 3 contains detailed descriptions of each result.

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Figures from Radford et al. (2019)

Prompting for Instruction Fine-tuned Models

- Models like ChatGPT, Llama-2 Chat, etc. have been fine-tuned as chat assistants
- These (often) were trained with specific prompt templates that segment the prompt into different parts: (1) system (2) assistant (3) user

Llama-2 Chat

sys:	[INST] <<SYS>> You are a helpful AI assistant… <</SYS>> [/INST]
user:	[INST] Organisms require energy in order to do what? [/INST]
asst:	mature and develop

Alpaca

sys:	### Instruction: You are a helpful AI assistant…
user:	### Instruction: Organisms require energy in order to do what?
asst:	### Response: mature and develop

Prompt formats from https://github.com/BerriAI/litellm/blob/main/litellm/llms/prompt_templates/factory.py

Prompting for Instruction Fine-tuned Models

	Llama-2-70B	Llama-2-7B Chat
Prompt	Story: Jason listened to the weather and heard it was goingto be sunny. He thought the kids might like to goswimming. He gathered up the swimsuits, towelsand sunscreen. Jason and the kids got into the truckand drove to the beach. They spent the next 2 hoursplaying and splashing in the surf. One-sentence Summary:	Story: Jason listened to the weather and heard it was goingto be sunny. He thought the kids might like to goswimming. He gathered up the swimsuits, towelsand sunscreen. Jason and the kids got into the truckand drove to the beach. They spent the next 2 hoursplaying and splashing in the surf. One-sentence Summary:
Model Output	Jason took the kids swimming at the beach after hearing the weather forecast, gathering necessary items and driving there.	Sure! Here is a one-sentence summary of the story: Jason and the kids went on a fun-filled day at the beach, playing and splashing in the surf after gathering all the necessary items, including swimsuits, towels, and sunscreen.

Story text from https://arxiv.org/pdf/1910.00998.pdf

Prompting for Instruction Fine-tuned Models

Prompt

Model Output

Llama-2-70B

Story: Jason listened to the weather
and heard it was goingto be
thought the kids might like
goswimming. He gathered u
swimsuits, towelsand sunsc
Jason and the kids got into
truckand drove to the beac
spent the next 2 hoursplayi
splashing in the surf.

One-sentence Summary:

Jason took the kids swimming at the beach after hearing the weather forecast, gathering necessary items and driving there.

Llama-2-7B Chat

Story: Jason listened to the weather
t was goingto be sunny. He
kids might like to
g. He gathered up the
towelsand sunscreen.
he kids got into the
ove to the beach. They
ext 2 hoursplaying and
the surf.

One-sentence Summary:

Sure! Here is a one-sentence summary of the story:
Jason and the kids went on a fun-filled day at the beach, playing and splashing in the surf after gathering all the necessary items, including swimsuits, towels, and sunscreen.

Lllama-2-7B Chat was *instruction fine-tuned* (more on this later) and so its responses look quite different from the Llama-2 models that were not

Story text from https://arxiv.org/pdf/1910.00998.pdf

IN-CONTEXT LEARNING

Few-shot Learning with LLMs

Suppose you have…

- a dataset $D = \{(x_i, y_i)\}_{i=1}^N$ and N is rather small (i.e. few-shot setting)
- a very large (billions of parameters) pre-trained language model

There are two ways to "learn"

This section!

Option A: Supervised fine-tuning

- **Definition:** fine-tune the LLM on the training data using…
  - a standard supervised objective
  - backpropagation to compute gradients
  - your favorite optimizer (e.g. Adam)
- **Pro:** fits into the standard ML recipe
- **Pro:** still works if N is large
- **Con:** backpropagation requires ~3x the memory and computation time as the forward computation
- **Con:** you might not have access to the model weights at all (e.g. because the model is proprietary)

Option B: In-context learning

- **Definition:**
  1. feed training examples to the LLM as a prompt
  2. allow the LLM to infer patterns in the training examples during inference (i.e. decoding)
  3. take the output of the LLM following the prompt as its prediction
- **Con:** the prompt may be very long and Transformer LMs require $O(N^2)$ time/space where N = length of context
- **Pro:** no backpropagation required and only one pass through the training data
- **Pro:** does not require model weights, only API access

Few-shot In-context Learning

- Few-shot learning can be done via in-context learning
- Typically, a task description is presented first
- Then a sequence of input/output pairs from a training dataset are presented in sequence

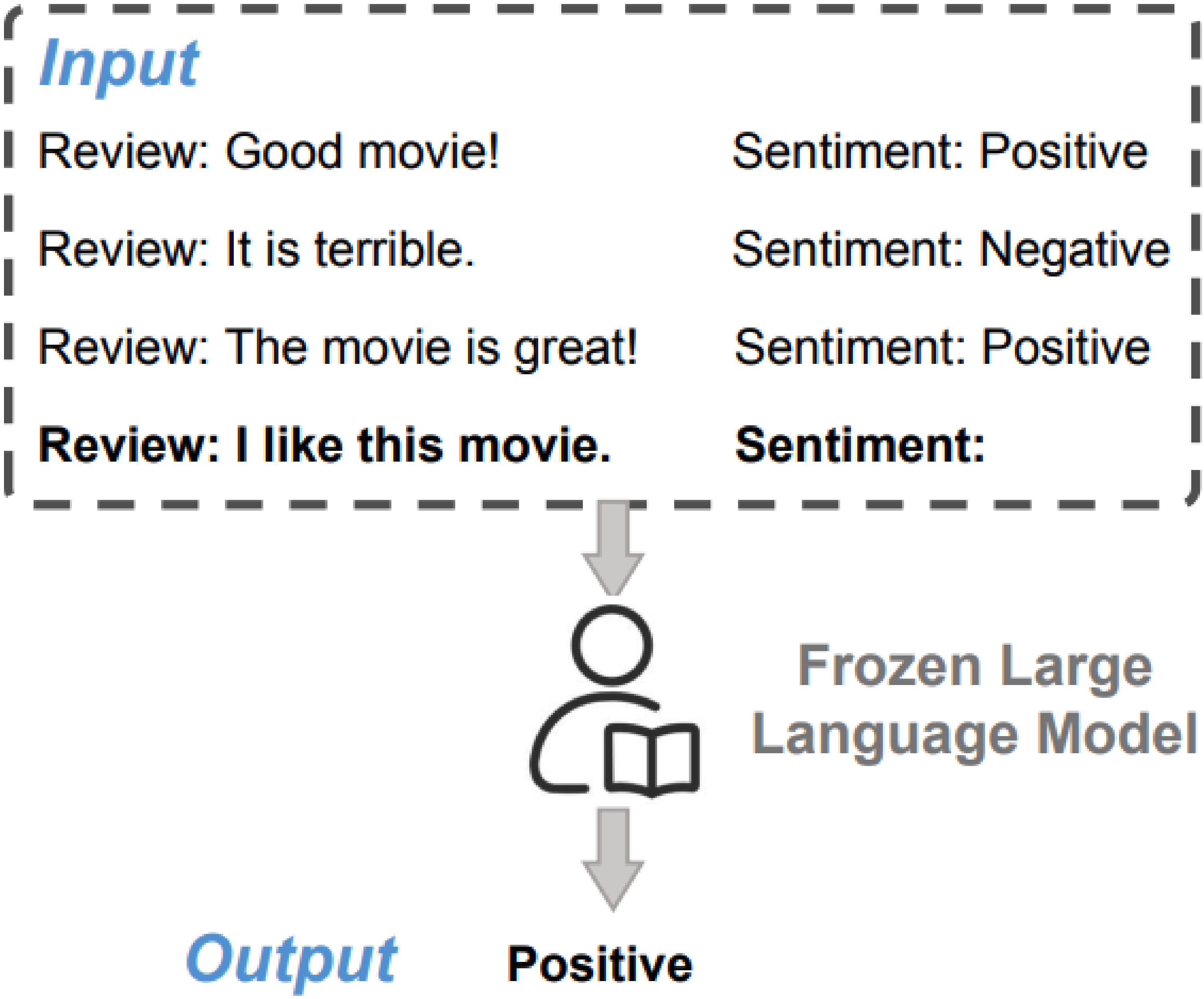


Figure from https://arxiv.org/pdf/2310.09881.pdf

Few-shot In-context Learning

- Few-shot learning can be done via in-context learning
- Typically, a task description is presented first
- Then a sequence of input/output pairs from a training dataset are presented in sequence

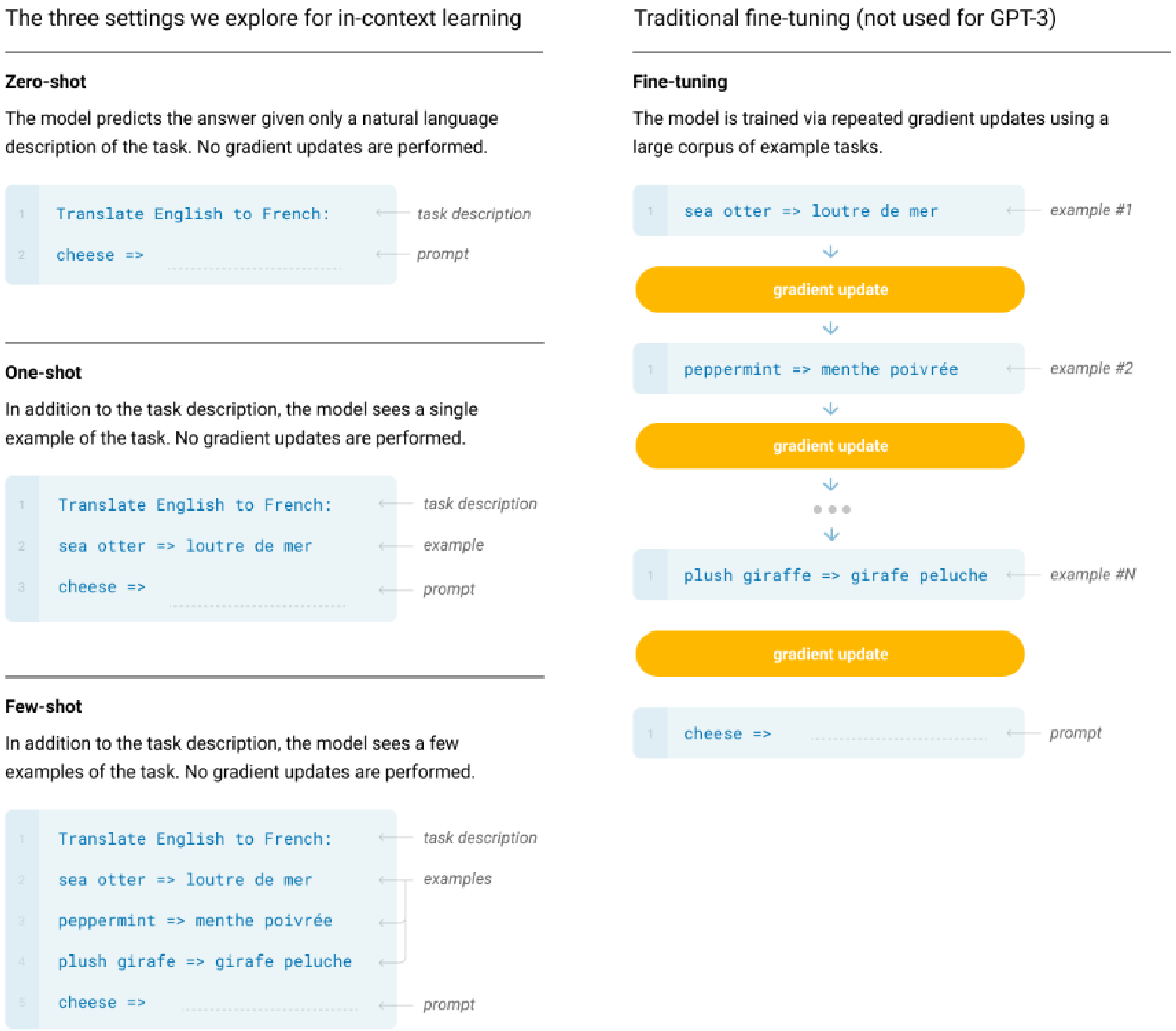


Figure from http://arxiv.org/abs/2005.14165

Few-shot In-context Learning

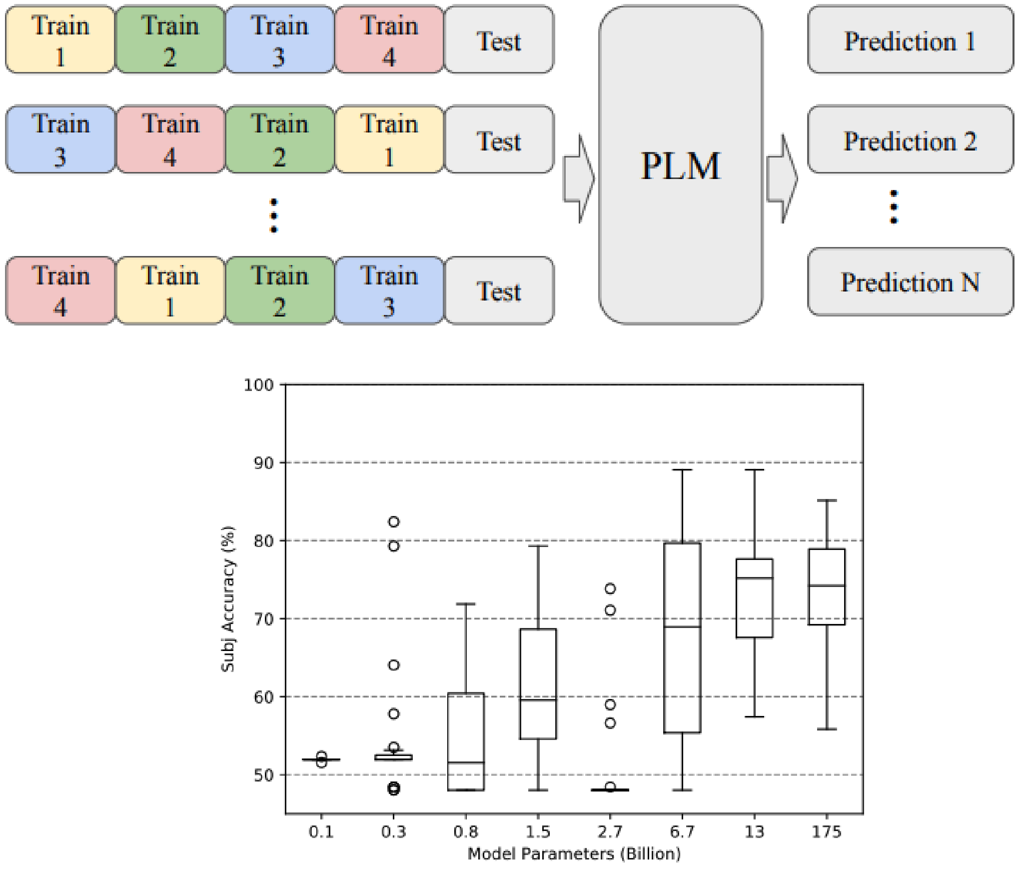


Figure 1: Four-shot performance for 24 different sample orders across different sizes of GPT-family models (GPT-2 and GPT-3) for the SST-2 and Subj datasets.

In-context learning can be sensitive to…

1. **the order the training examples are presented**
2. the balance of labels (e.g. positive vs. negative)
3. the number of unique labels covered

Figure from http://arxiv.org/abs/2104.08786

Few-shot In-context Learning

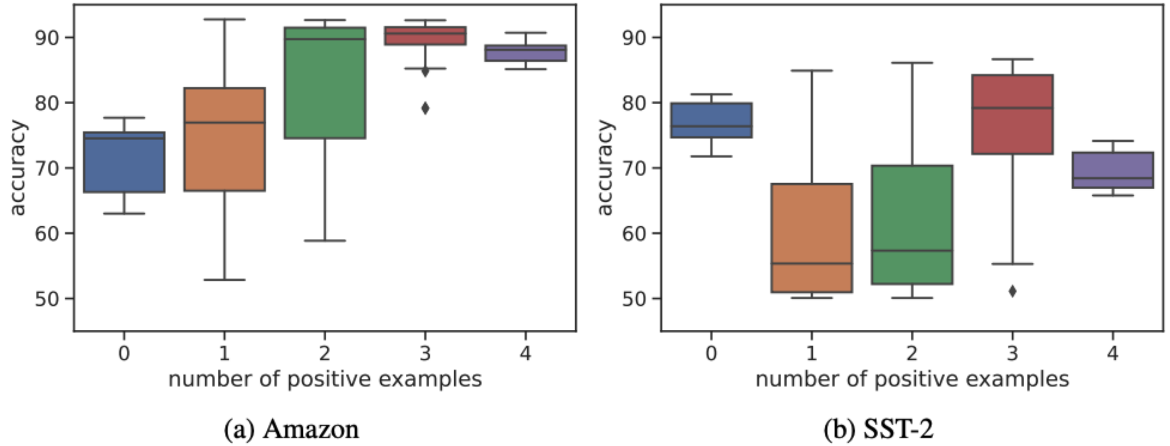


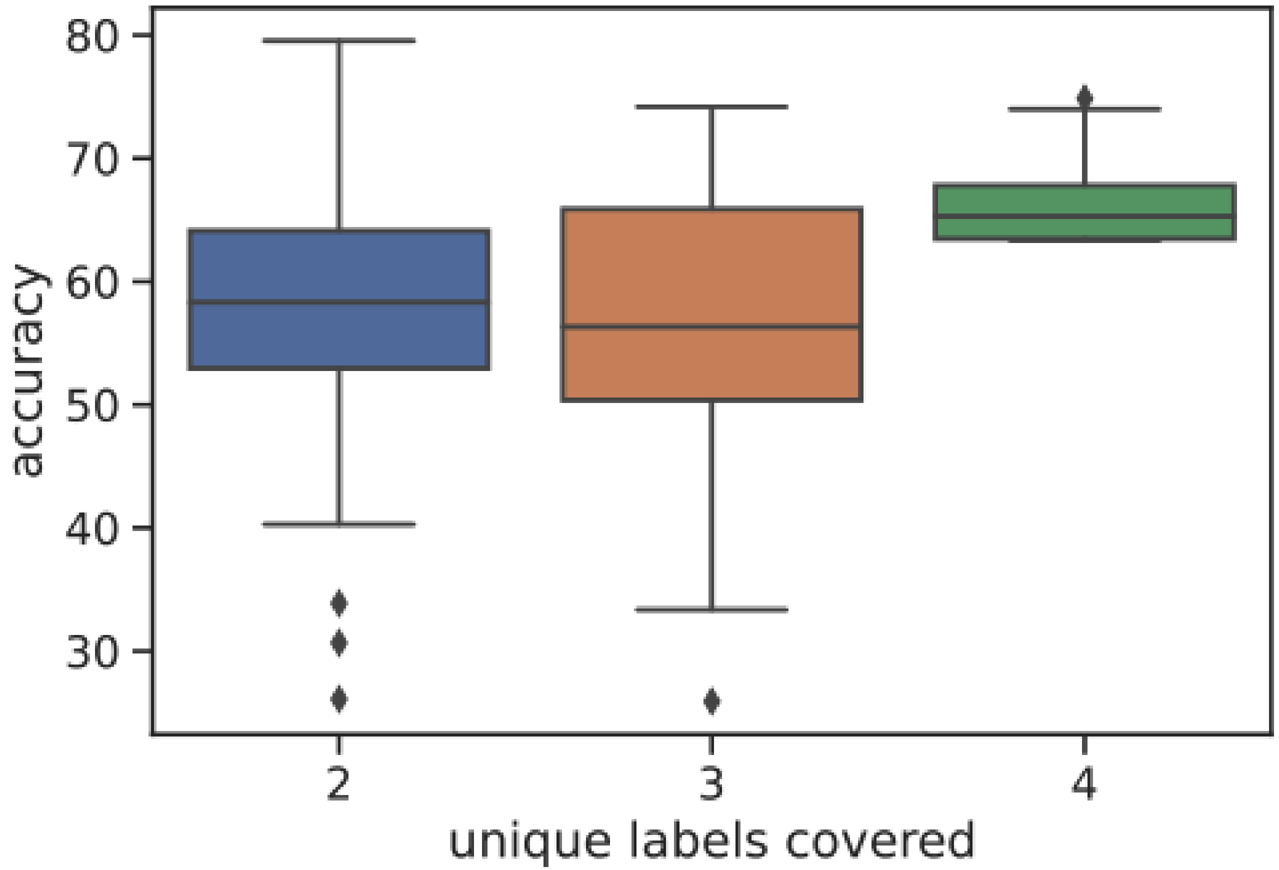
Figure 3: Accuracies of Amazon and SST-2 with varying **label balance** (number of positive examples in demonstration), across 100 total random samples of 4 demonstration examples.

In-context learning can be sensitive to…

1. the order the training examples are presented
2. **the balance of labels (e.g. positive vs. negative)**
3. the number of unique labels covered

Figure from https://aclanthology.org/2022.emnlp-main.622.pdf

Few-shot In-context Learning



In-context learning can be sensitive to…

1. the order the training examples are presented
2. the balance of labels (e.g. positive vs. negative)
3. **the number of unique labels covered**

Figure from https://aclanthology.org/2022.emnlp-main.622.pdf

Few-shot In-context Learning

You would expect these to be important…

A. whether or not the training examples have the true label (as opposed to a random one)

B. having more in-context training examples

…but it's not always the case

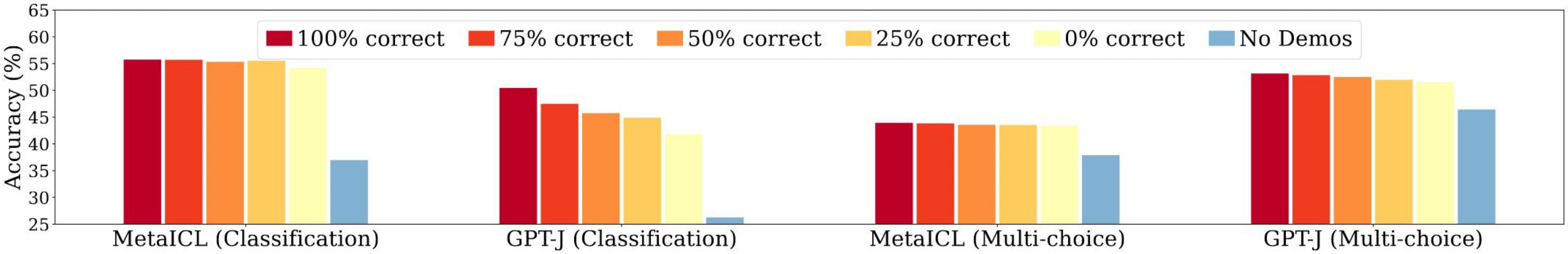


Figure from http://arxiv.org/abs/2202.12837

Few-shot In-context Learning

You would expect these to be important…

A. whether or not the training examples have the true label (as opposed to a random one)

B. having more in-context training examples

…but it's not always the case

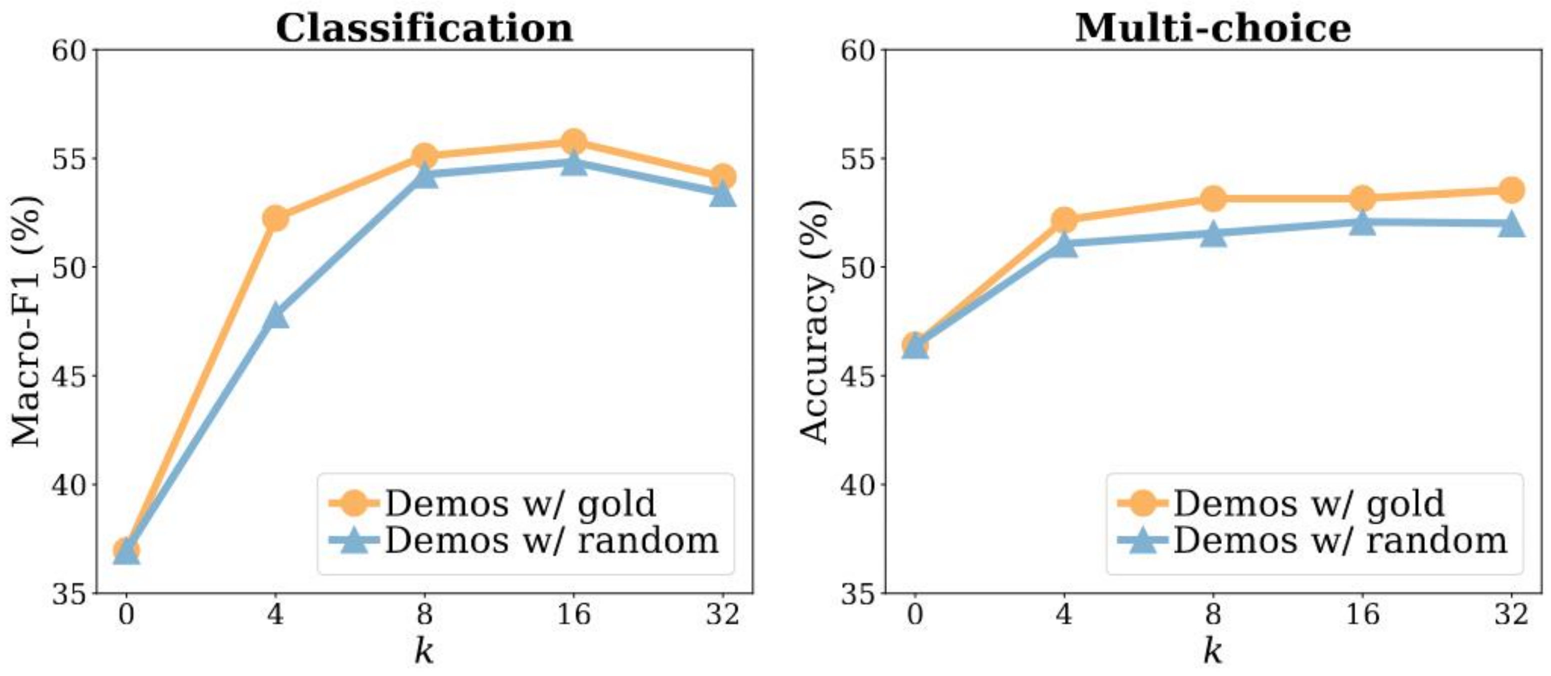


Figure from http://arxiv.org/abs/2202.12837