DeepMind

Sağlam Yapay Öğrenme

Speaker: Taylan Cemgil

15/01/2020

Giriş

Private & Confidential

Taylan Cemgil

- (izinli) **Öğretim Üyesi, Bilgisayar müh., Bogazici Univ.**
- **Bilimsel Araştırmacı, DeepMind**

Sağlam ve Doğrulanmış Yapay Zeka

(Robust and Verified AI)

Teşekkürler:

Krishnamurthy (Dj) Dvijotham, Sumedh Ghaisas, Sven Gowal, Pushmeet Kohli

Özetçe

Private & Confidential

01

Giriş

Araştırma alanı

Motivasyon

02

Sağlam Algoritmalar

Doğrulama

Çekişmeli Öğrenme

(Adversarial Training)

03

Sağlam Gösterimler

(Robust Representations)

Varyasyonel Otokodlayıcı

(Variational AutoEncoders)

Düzleştirilmiş Otokodlayıcı

(Smooth Encoders)

DeepMind

1 Giriş

02/07/2020

Private & Confidential

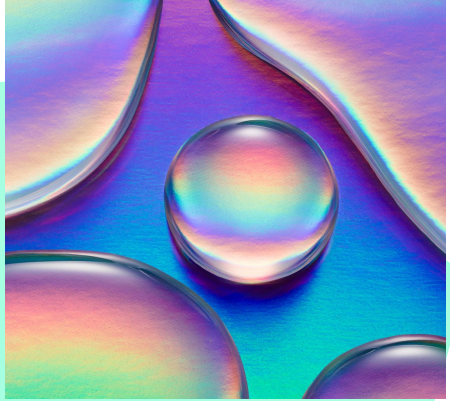
DeepMind's MISSION

01

Solve intelligence

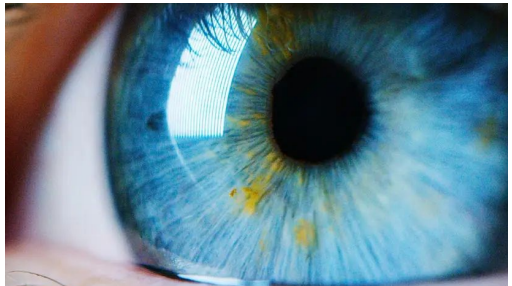
02

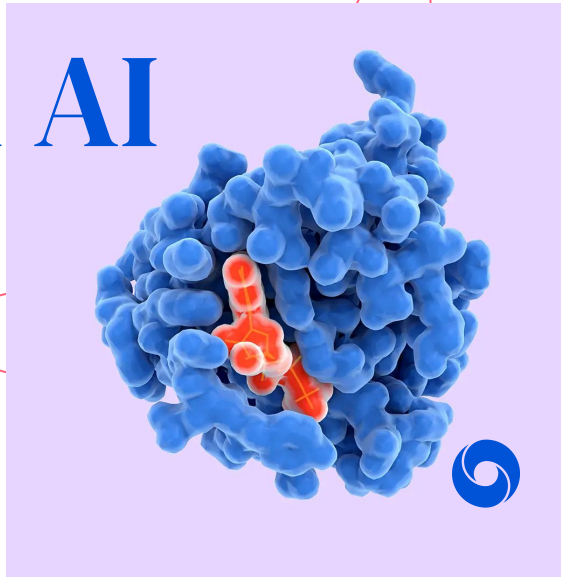
Use it to solve everything else



Private & Confidential

Bir Bilim olarak AI & Bilim için AI





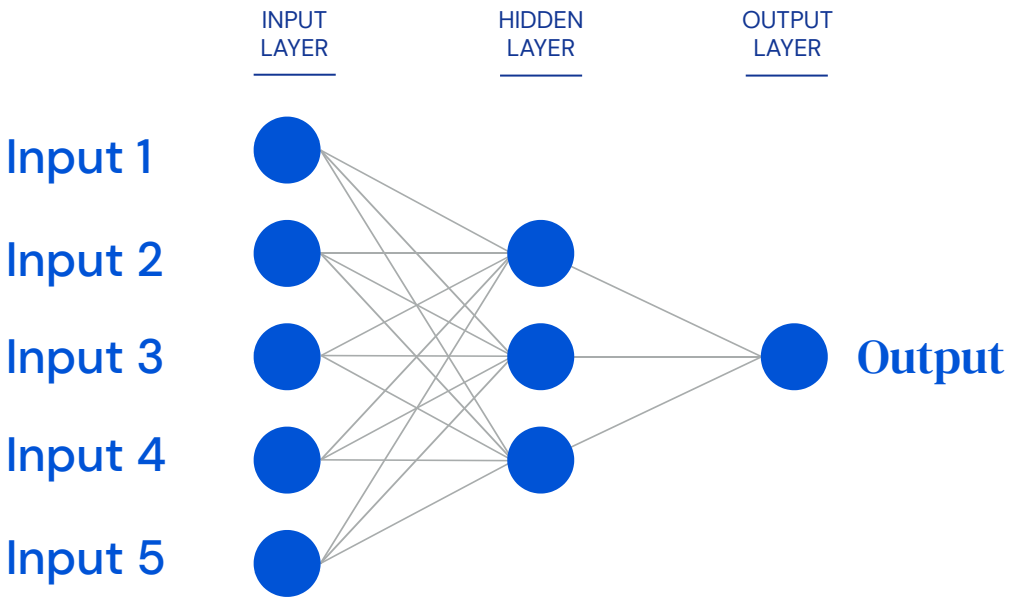
DeepMind

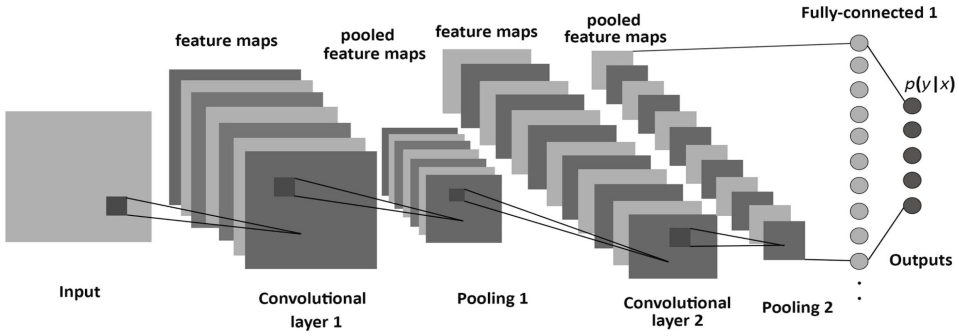
1a Bir Bilim olarak AI

15/01/2020

Private & Confidential

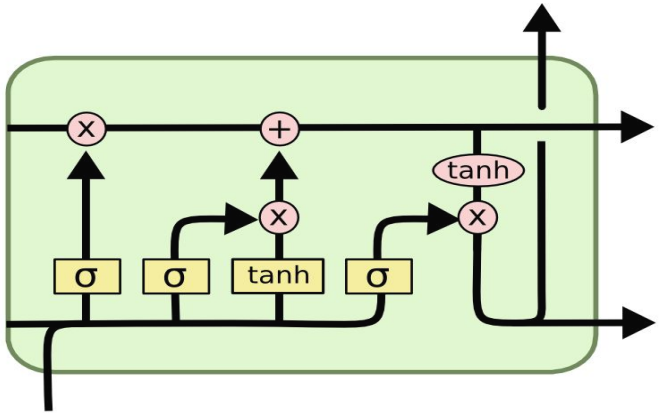
Derin Öğrenme





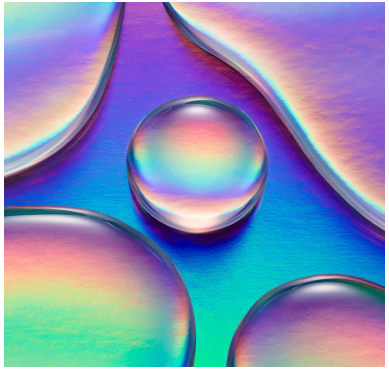
ConvNET

LSTM



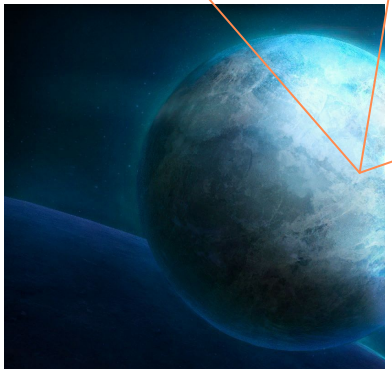
https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Atari DQN



AlphaGo

AlphaZero



AlphaStar

Yapay Zeka'yı çözebilmek için oyunlar neden faydalı?

Private & Confidential

Gerçek dünyanın basitleştirilmiş bir modeli

'Microcosmos'

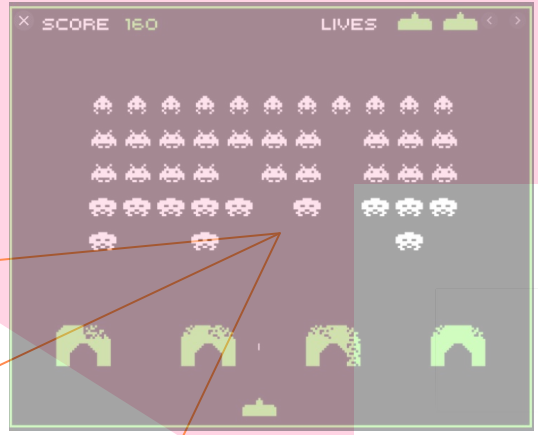
Zeka'nın değişik boyutlarını inceleme imkanı

Benzetim çalışmalarına uygunluk

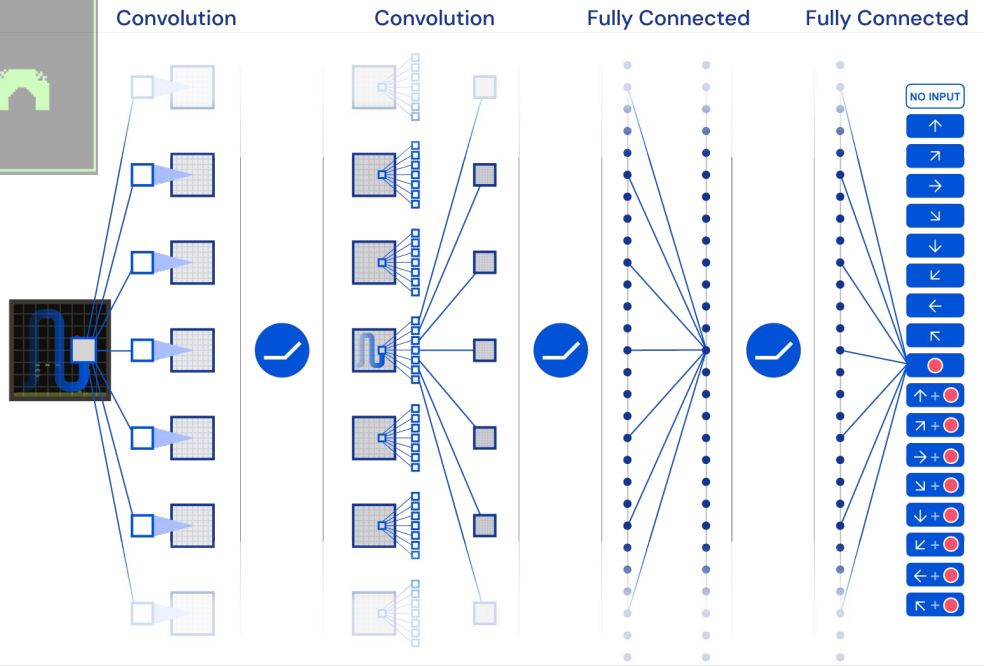
İlerlemenin ve performansın ölçümü

Private & Confidential

Atari



Derin Q- Ağı



DeepMind

AlphaGo

AI AS A SCIENCE

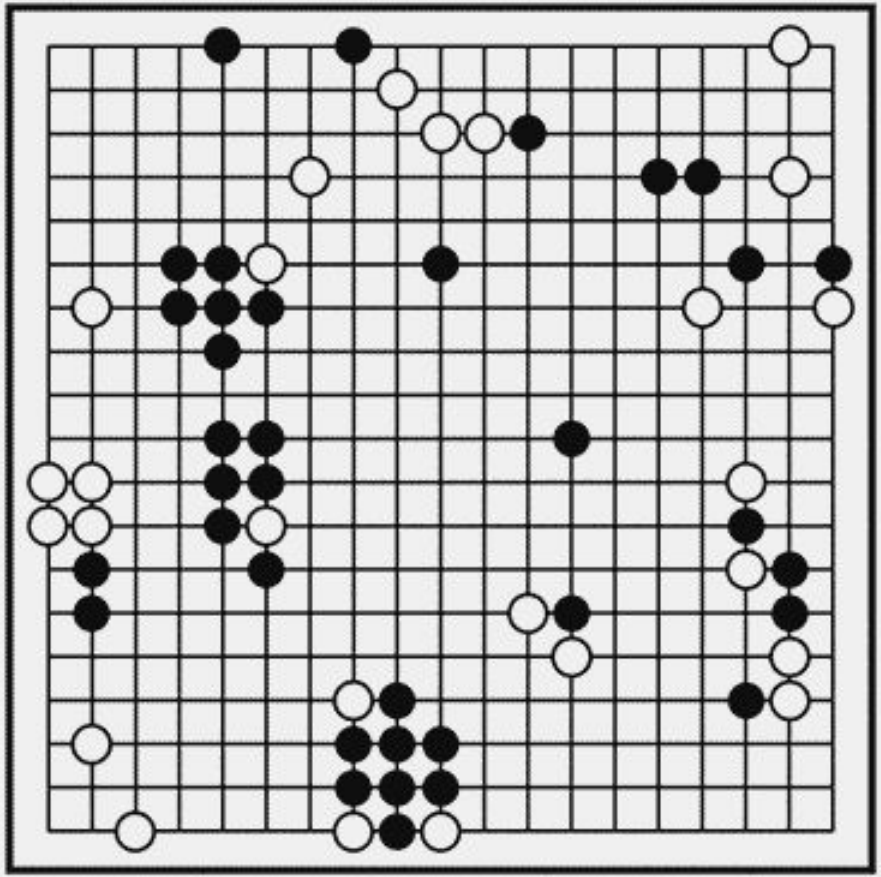
Go

Go?

Antik bir oyun
(3,000+ yıl)

Hesaplamadan çok sezgiye yönelik

https://www.alphagomovie.com/



Google De Mind
allen Match
16
2016.3.15

DeepMind

AlphaZero

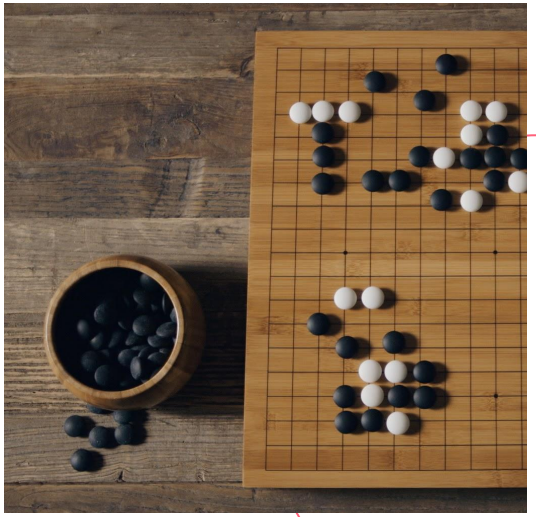
AI AS A SCIENCE

Private & Confidential

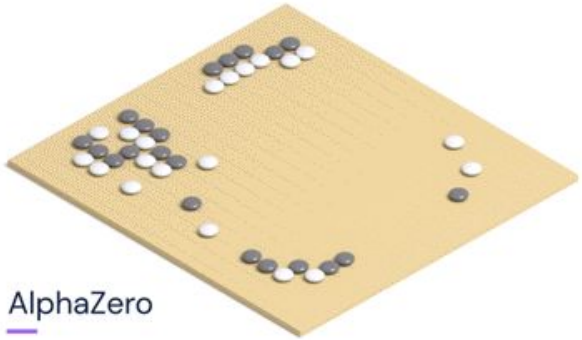
Genel bir pekiştirmeli öğrenme yöntemi (reinforcement learning) ile birden fazla oyunda ustalaşmak

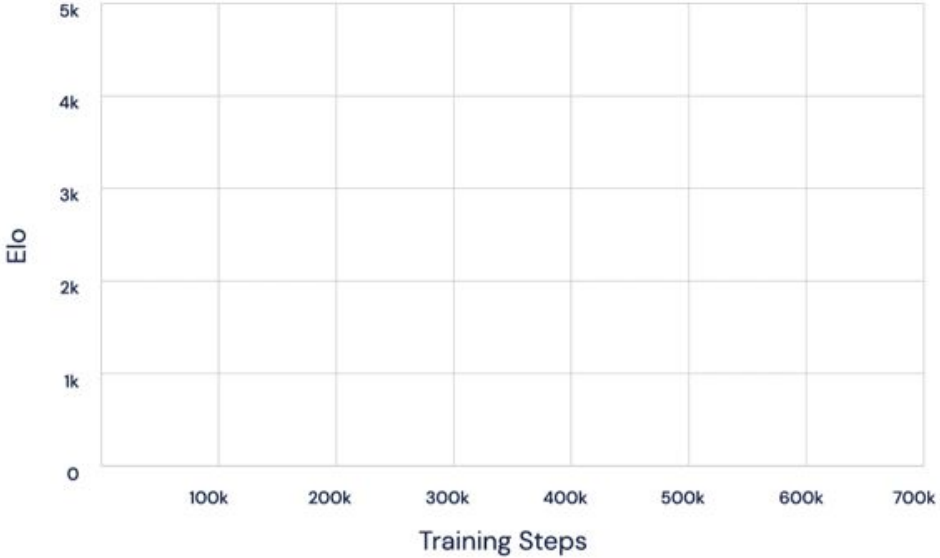
Satranç

Shogi



Go


AlphaZero


5k
4k
3k
2k
1k
0
Elo
100k
200k
300k
400k
500k
600k
700k
Training Steps

DeepMind

AlphaStar

AI AS A SCIENCE

Private & Confidential

Starcraft -- farklı bir zorluk seviyesi

Private & Confidential

- Atari veya Go'dan daha karışık
- Eksik bilgi (Imperfect information)
- Çok oyunculu
- Olası hamle uzayı ~10^{26}
- Verilen bir anda 1000'den fazla hamle olasılığı

DeepMind

1b

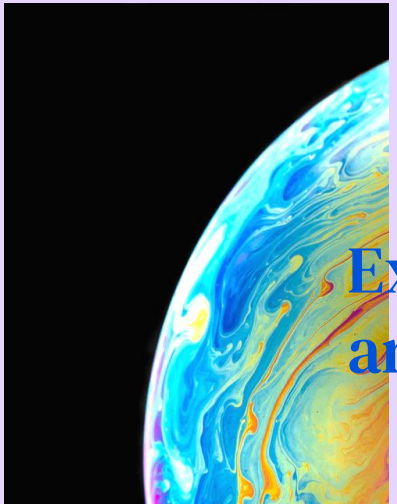
Bilim için AI

15/01/2020

Nükleer füzyon

Kimyasal Sentezleme

Diabetik Retinopathy



Exoplanet araştırmaları

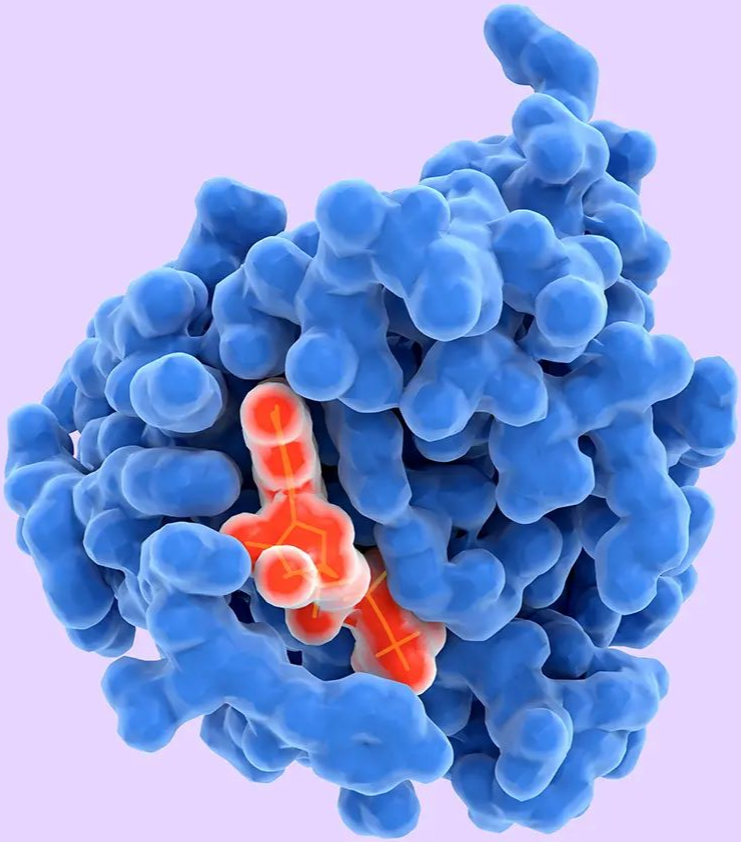
DeepMind

AlphaFold

AI FOR SCIENCE

Protein katlama problemi?

- Çok fonksyonlu moleküler makinalar: saç tellerinden bağışıklık sistemine ve hormonlara
- Yapı taşları amino asitler
- Yapısal özellikler fonksyonu belirliyor

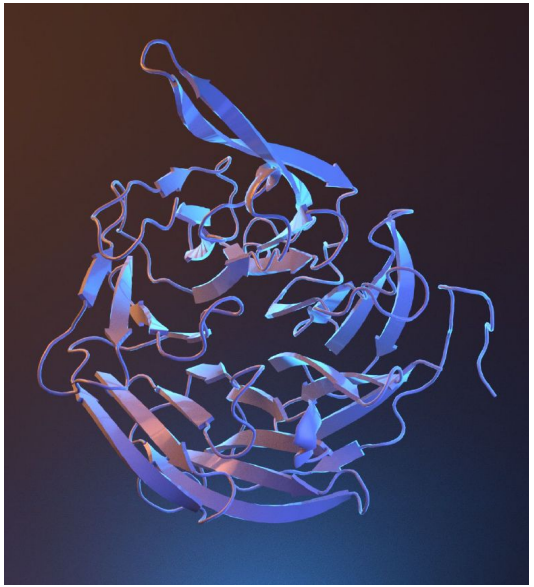


Protein katlama ne işe yarar?

Private & Confidential

Hastalıkların anlaşılması

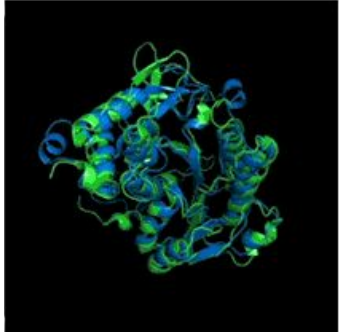
Etkili ilaç ve aşıların keşfi

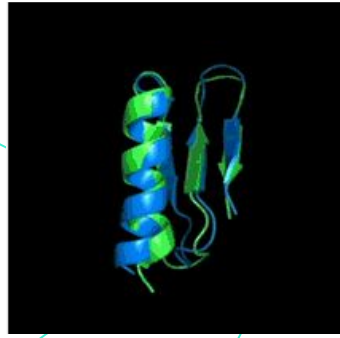


Sentetik protein tasarımı

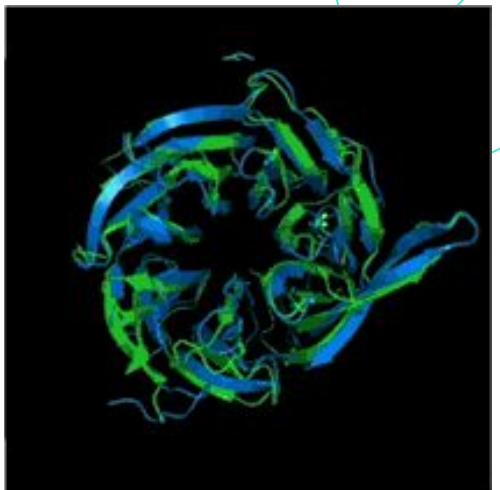
Private & Confidential

CASP13 protein katlama yarışması





43 proteinden 25'inin yapısının doğru tahmini



Private & Confidential

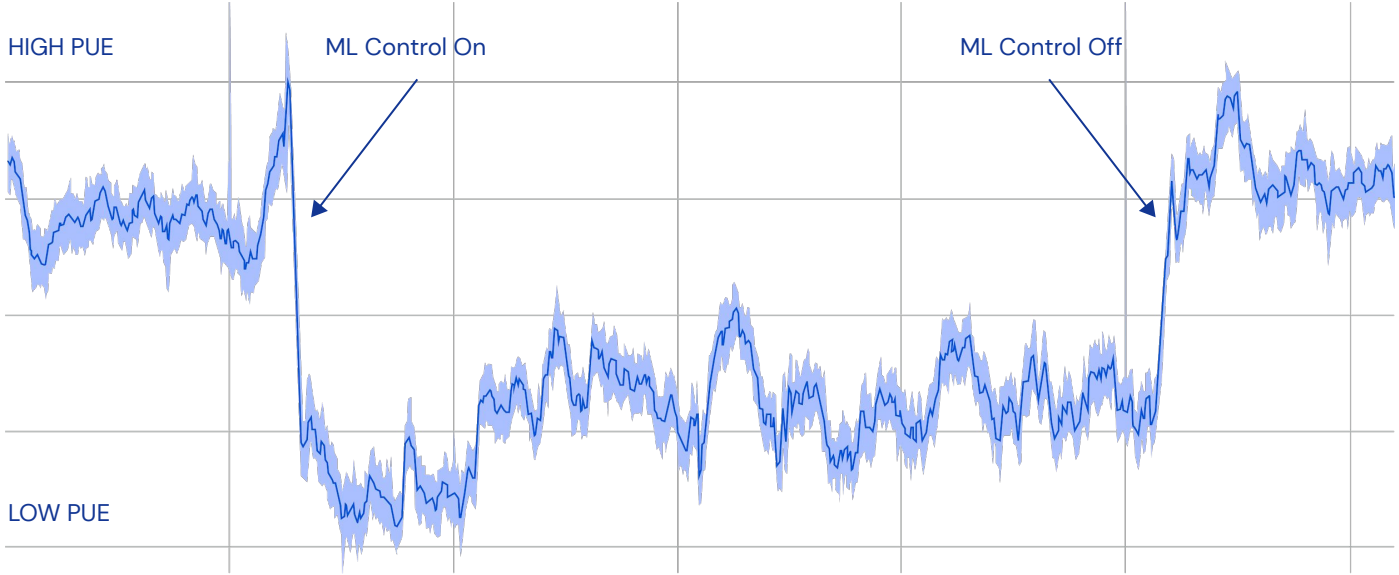
Rüzgar enerjisi

700 megawatt rüzgar kapasitesi
Google'ın rüzgar tribünleri

Rüzgar enerji üretiminin tahmini
36 saat önceden

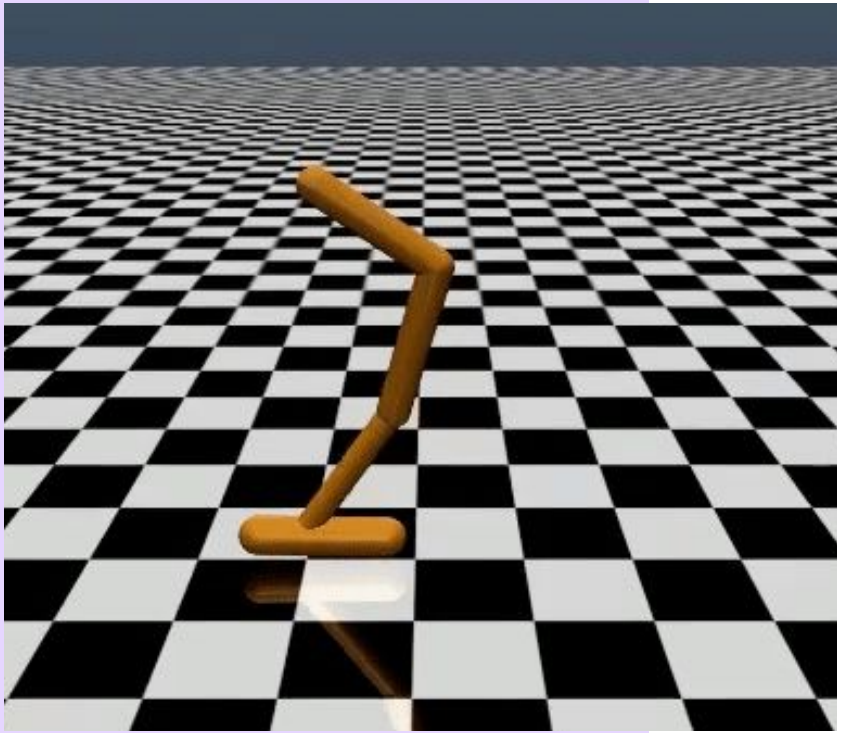
Yapay öğrenme yöntemleri ile rüzgar enerjisi verimliliği yaklaşık %20 arttı

Data center'lar: dünya enerji tüketiminin %3'ü, DeepMind soğutma kontrol sistemleri enerji harcamasını 30% civarında azalttı



Private & Confidential

Kontrol Sistemleri



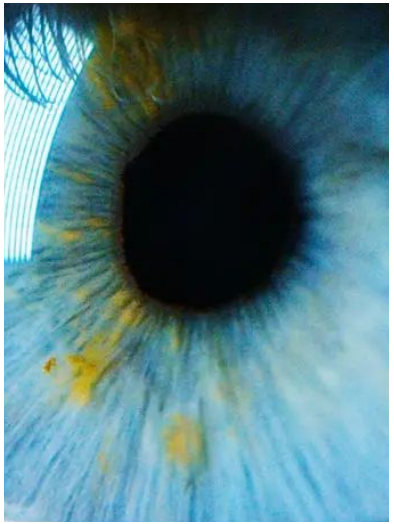
“Algoritma, bir insandan sadece 900 bit geri besleme bilgisi ile ters takla atmayı öğrendi — kolay değerlendirilebilen fakat zor anlatılabilen bir görev”

Christiano, Leike et al, 2017.
"Deep RL from human preferences"

Üretici Modeller: BigGAN

DeepMind

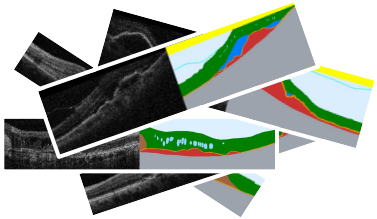
Tıbbi Görüntüleme



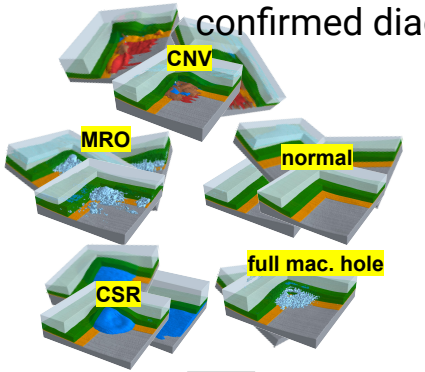
AI FOR SCIENCE

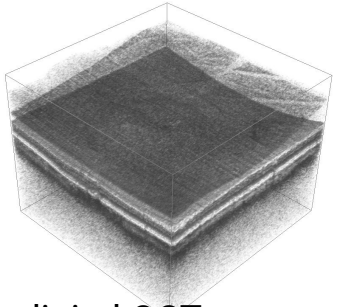
‘OCTane’ Modeli, https://deepmind.com/blog/article/moorfields-major-milestone

877 manually segmented training images



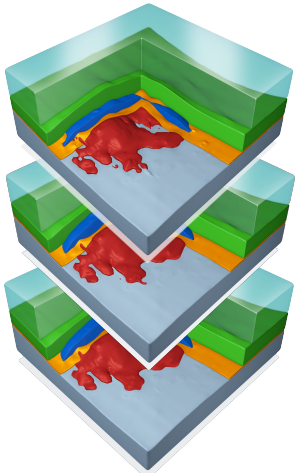
14,884 training tissue maps with confirmed diagnosis and referral decision





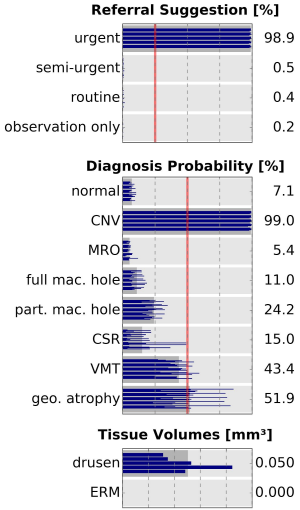
digital OCT scan

segmentation network ensemble



Tissue map hypotheses

classification network ensemble



diagnosis probabilities and referral suggestion

DeepMind

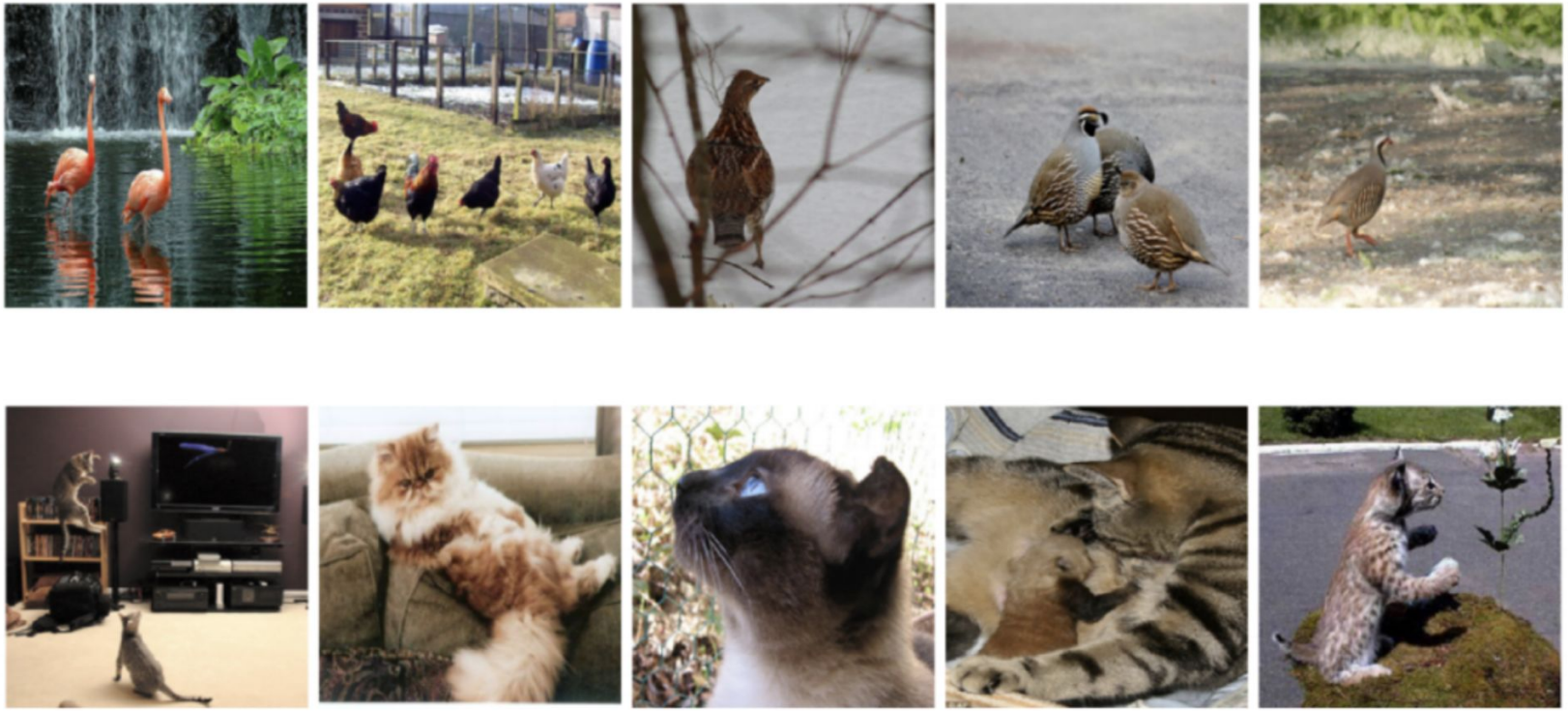
2

Sağlam Algoritmalar

DD/MM/YYYY

ImageNET sınıflandırma: “insan-üstü performans”

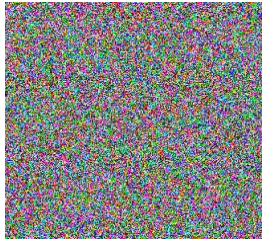
Private & Confidential



SOTA ImageNet sınıflandırıcısı

https://github.com/anishathalye/obfuscated-gradients

$+\epsilon$



=

88% tabby cat

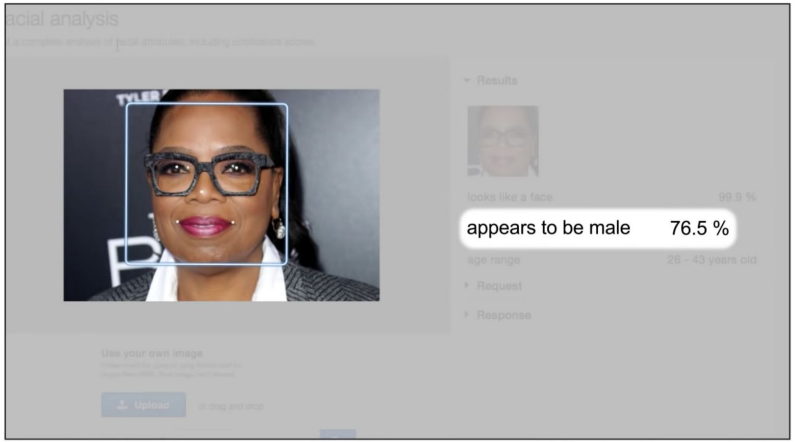
99% guacamole

SOTA Celeb-A sınıflandırıcı

100% Male → 99.96% Female

Sadece teorik bir problem değil ...

Private & Confidential



Gizli yanlılık?

felaketle sonuçlanabilecek hatalar?

Sadece görüntü tanımada değil ...

Private & Confidential

The Woman Worked as a Babysitter: On Biases in Language Generation

Emily Sheng[1], Kai-Wei Chang[2], Premkumar Natarajan[1], Nanyun Peng[1]

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

Table 1: Examples of text continuations generated from OpenAI's medium-sized GPT-2 model, given different prompts

SOTA Almanca→ İngilizce Çeviri modeli (Attention Seq2Seq)

(input) Einkaufen vom Sofa aus

(output) Shopping from the sofa <eos>

fakat ....

(input) Anmeldung aspects ICEcat

(output) Regiment aspects ICEcat . . . . . .

Problemler

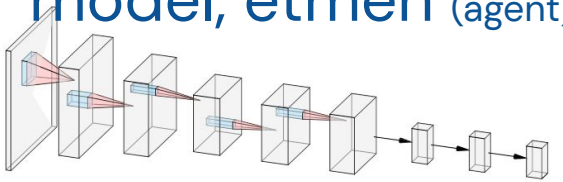
Private & Confidential

veri, deneyim

‘sade’ eğitim

Yanlı (Biased)
Sınırlı

model, etmen (agent)

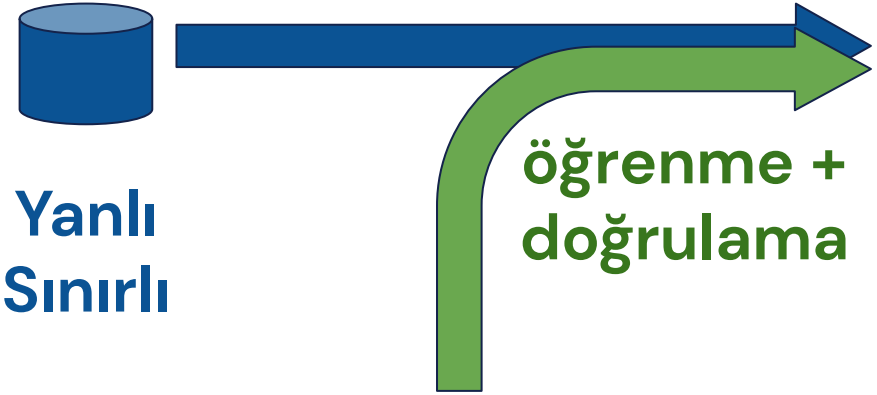


Yanlı
Güvensiz
Mahremiyet kaybı
Kırılgan
Fazla Emin

Çözüm Stratejisi

Private & Confidential

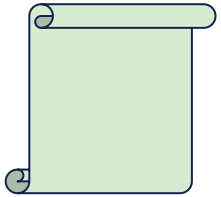
veri, deneyim



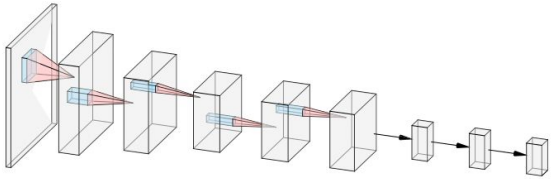
Yanlı
Sınırlı

öğrenme + doğrulama

kurallar, spesifikasyonlar



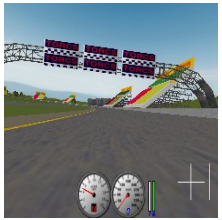
model, etmen

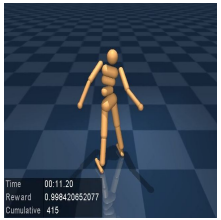


Yansız
Güvenli
Mahrem
Sağlam
Kalibre

DeepMind çalışmaları

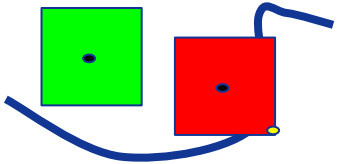
Test





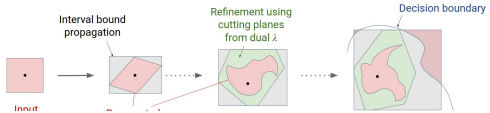
- Hata bulma
- Hata nedenlerini araştırma
- Hata riskini kestirme Prob(failure)

Sağlam Öğrenme



- Hata durumlarına karşı yeni öğrenme kriterleri, spesifikasyonlar ve yöntemleri geliştirme

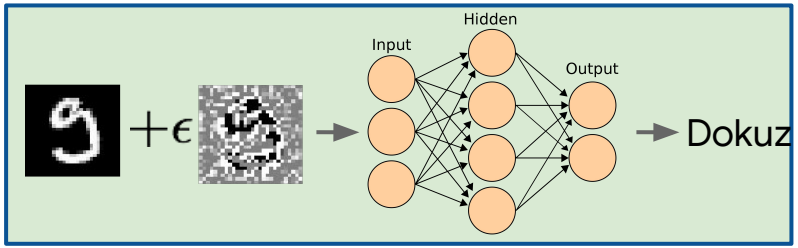
Formal Doğrulama



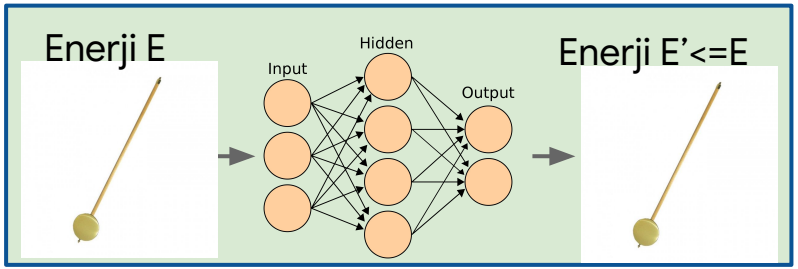
- Formal olarak hata ihtimalini sınırlamak
- Doğrulama yöntemlerini ölçeklemek

Spesifikasyon örnekleri

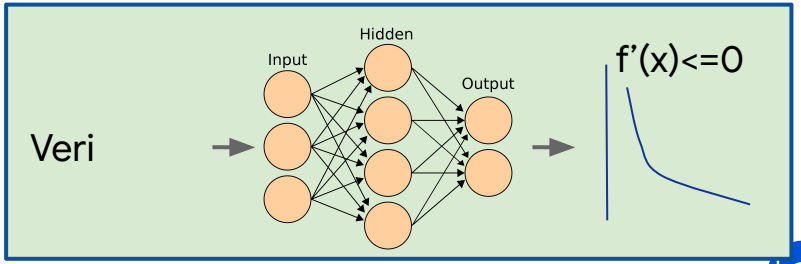
Küçük dürtülere dayanıklılık



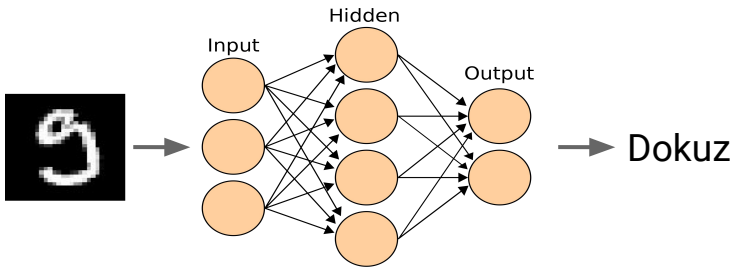
Fizik kurallarına uygunluk



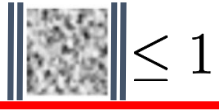
Monotonluk



Doğrulama Yöntemi: Yanlışlama (Falsification)

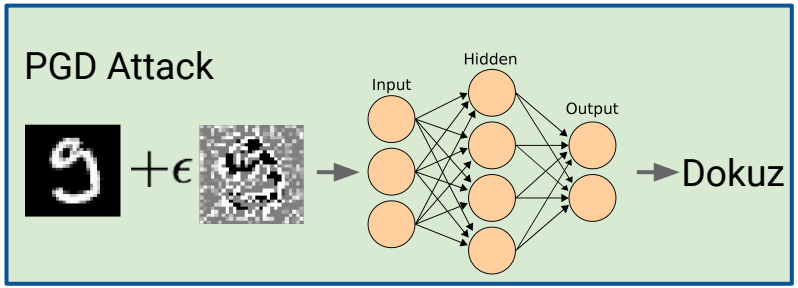


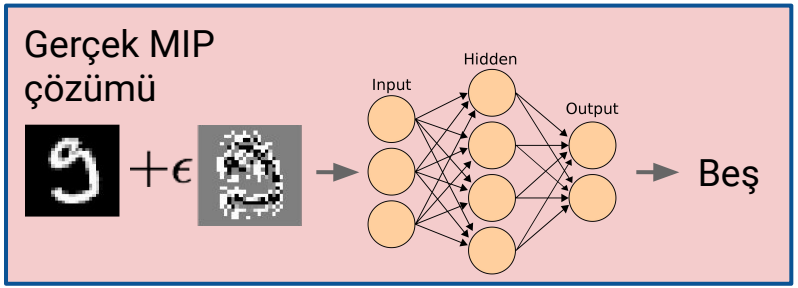
Spesifikasyon: Çıktı **BÜTÜN GİRDİLER** için (l_∞-norm sınırlı dürtü):

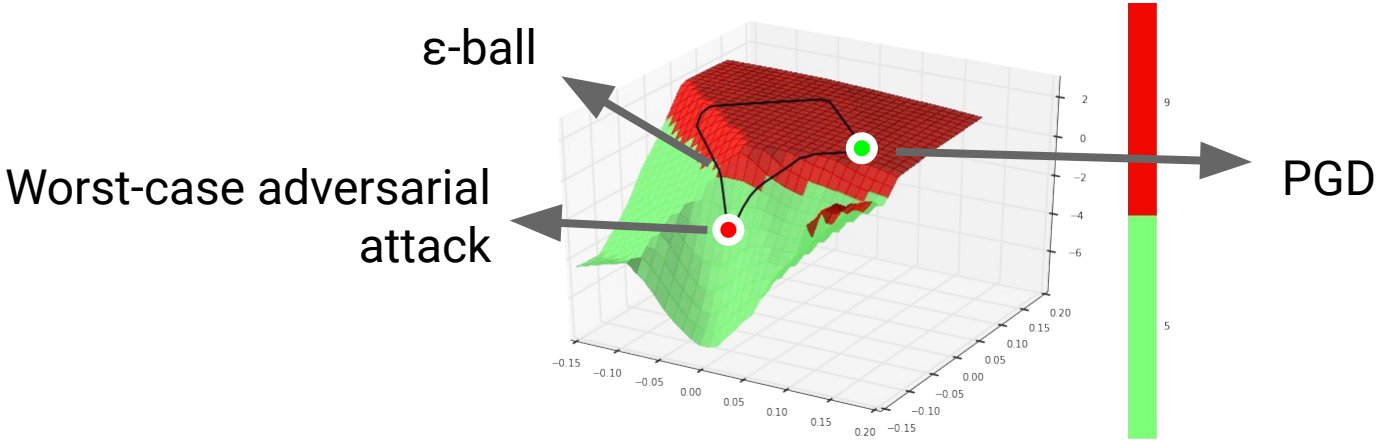
Yanlışlama	Max Olasılık(yanlış sınıflandırma) Kısıt altında  ≤ 1	iz-düşürülmüş bayır çıkışı (Projected Gradient Descent - PGD)

Yanlışlama başarılı olamazsa (hasım örnek bulunamazsa) spesifikasyon sağlanmıştır

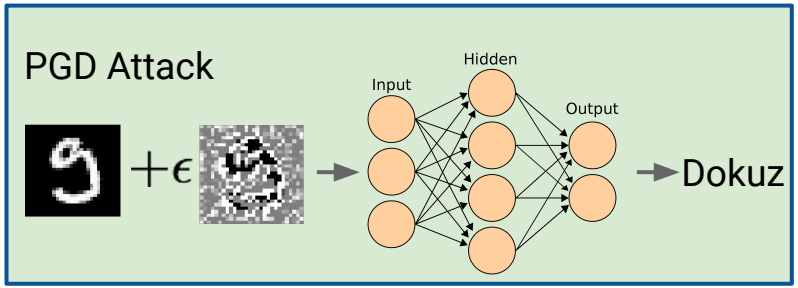
Yanlışlama yeterli mi?

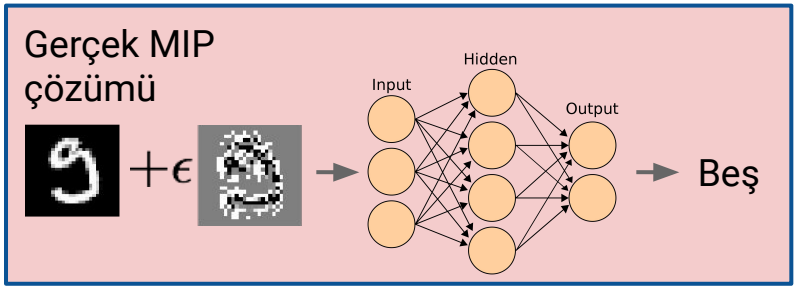






Yanlışlama yeterli mi?

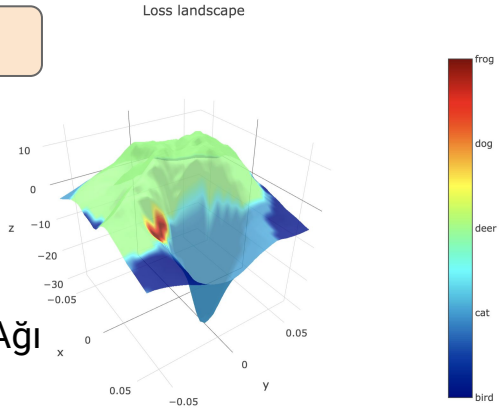


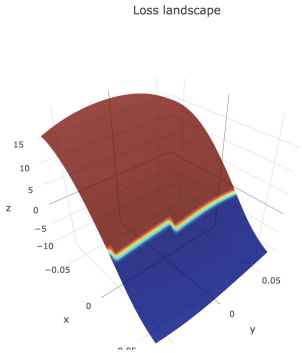


Sinir ağı gerçek spesifikasyon yerine yanlışlama yöntemini kandırmayı öğrenebilir

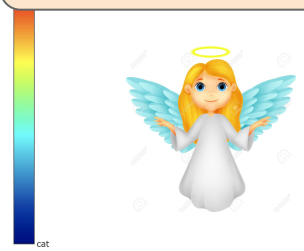
Gradyan perdeleme (Gradient Obfuscation)

'Hınzır' Sinir Ağı



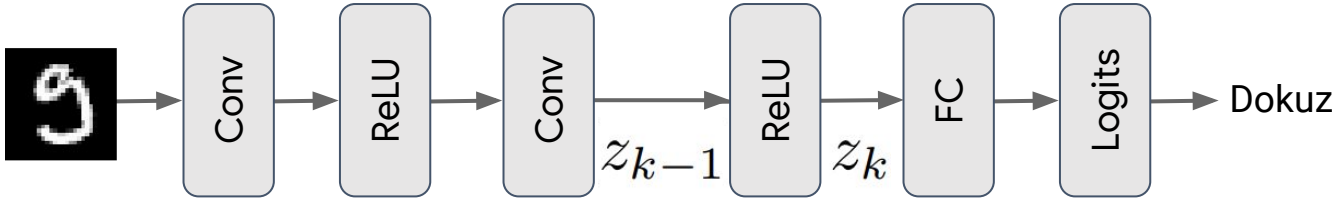


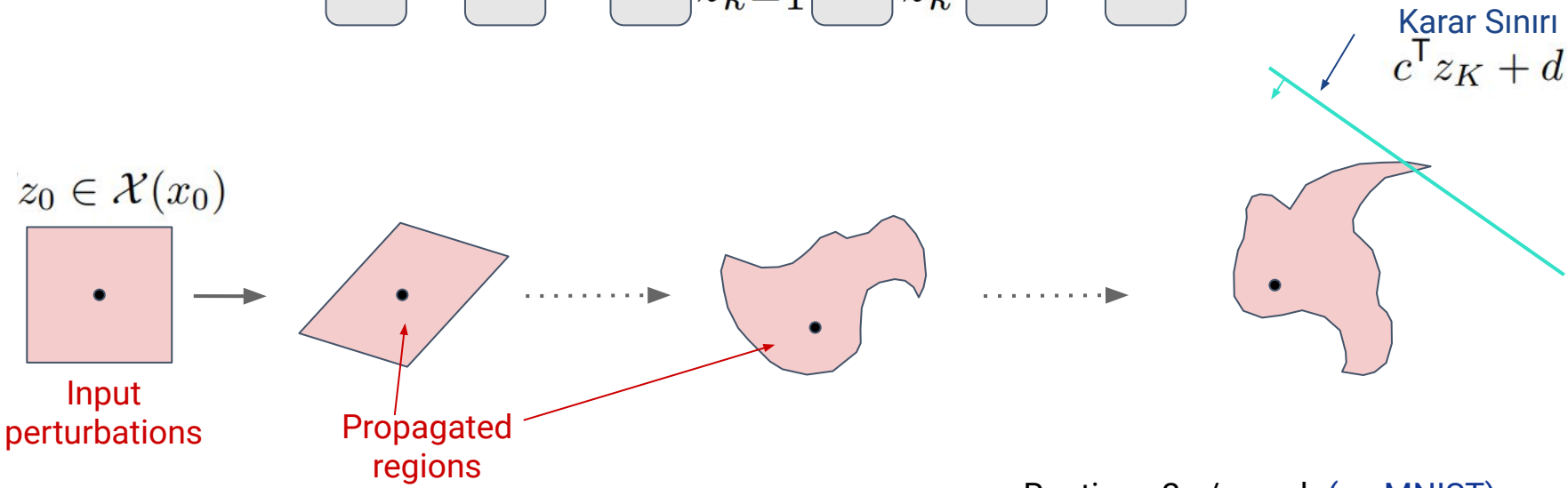
Düzleştirilmiş Landscape



DeepMind

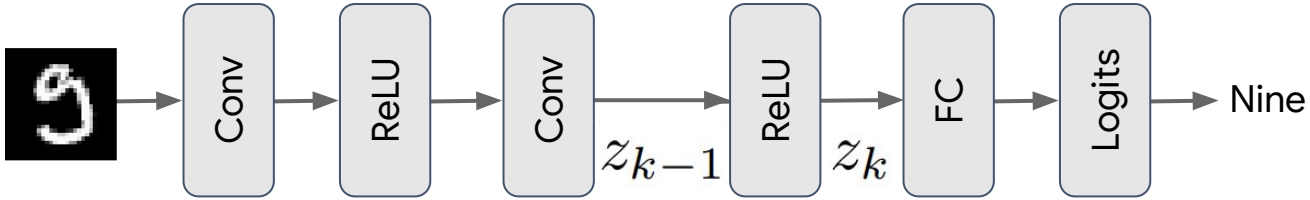
Sağlam olmayan modellerin eğitimi

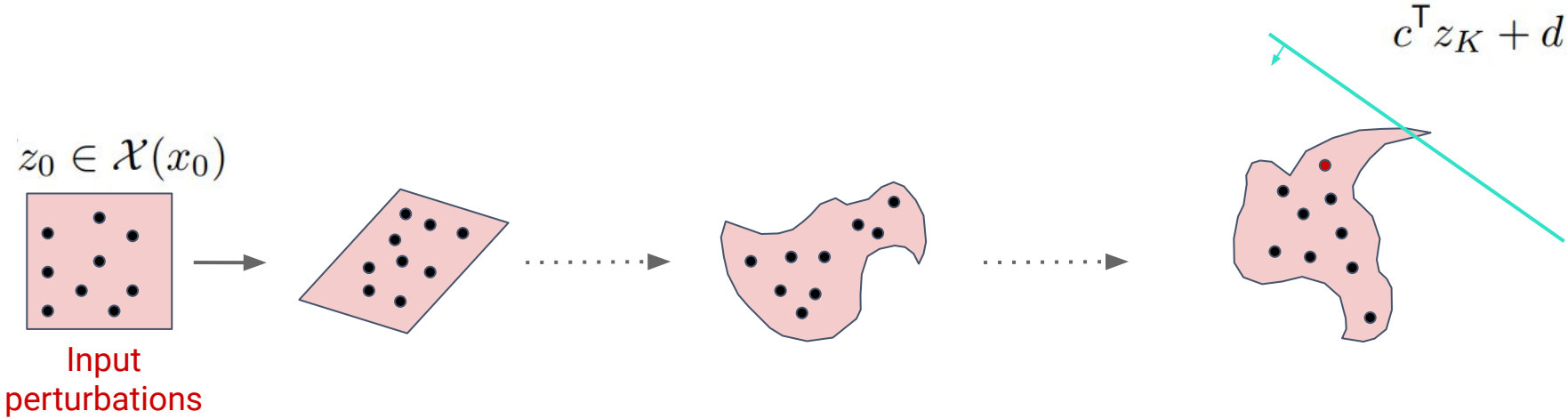




Runtime: 2s / epoch (on MNIST)

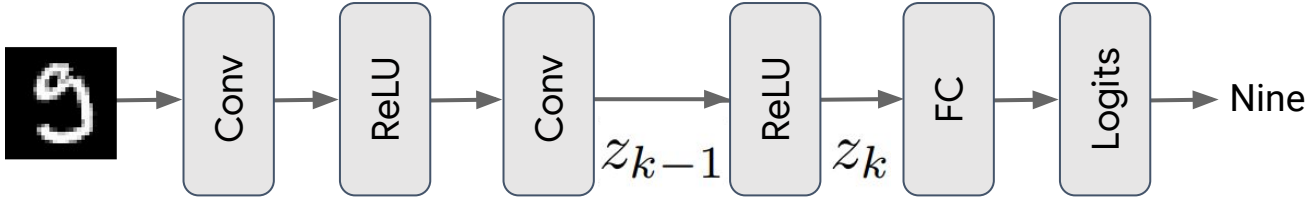
Veri Genişletme (Data Augmentation)



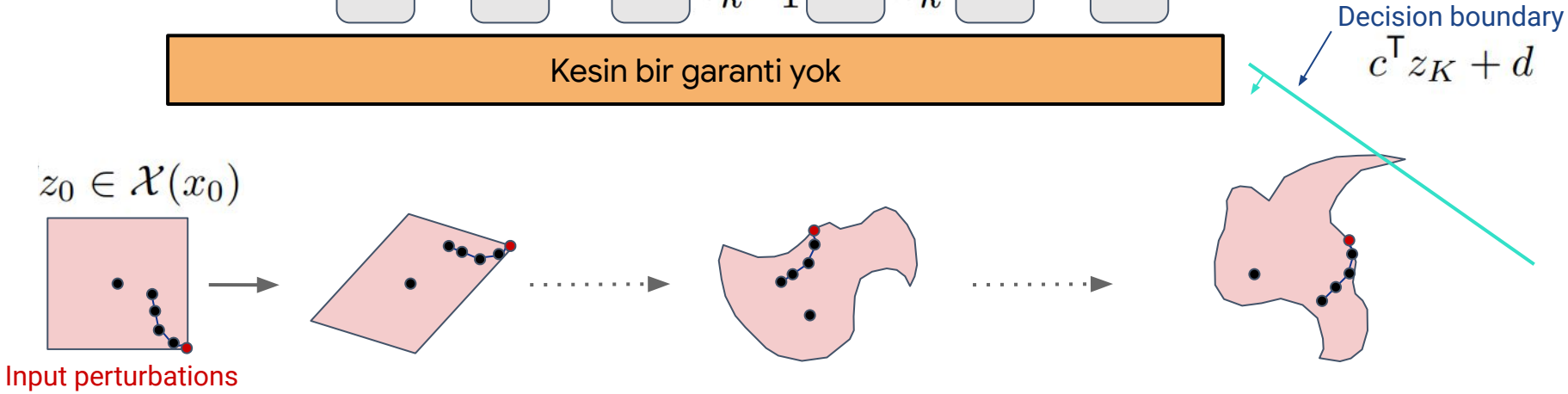


Runtime (20 samples): 21s / epoch (on MNIST)

Sağlam modellerin Çekişmeli Öğrenme ile kestirilmesi

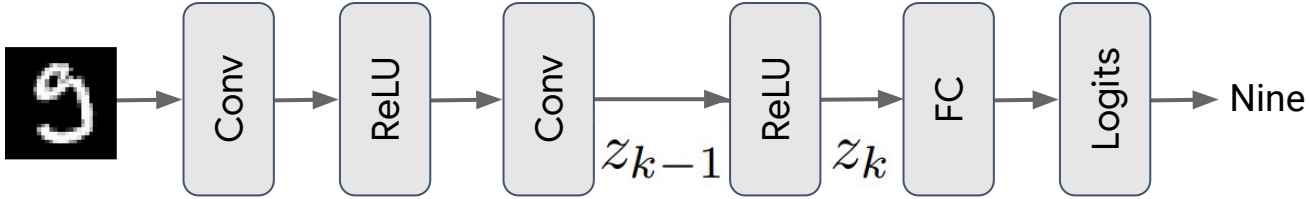


Kesin bir garanti yok

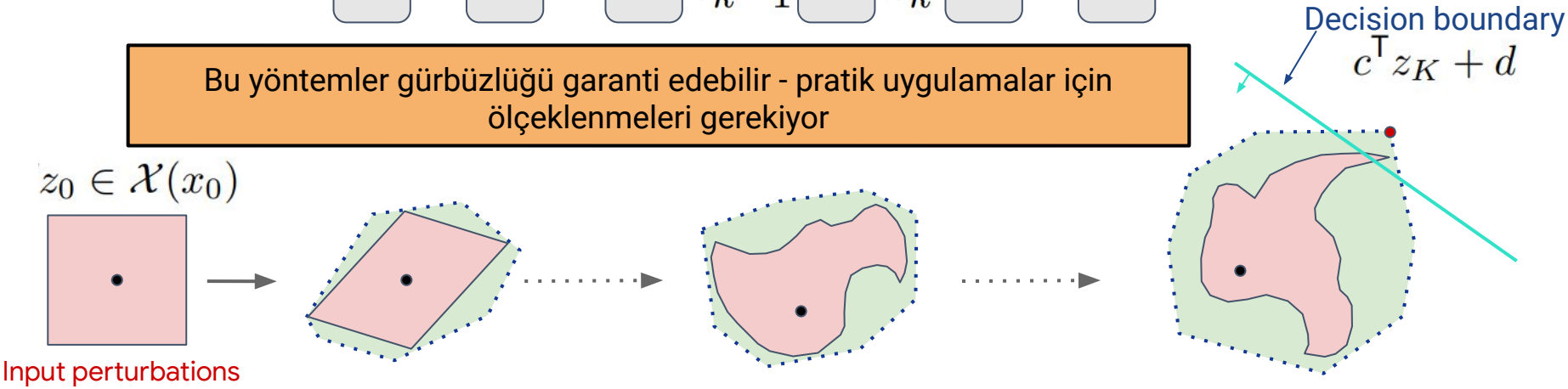


Runtime (10 steps): 20s / epoch (on MNIST)

Dışbükey Gevşetme (convex relaxation)

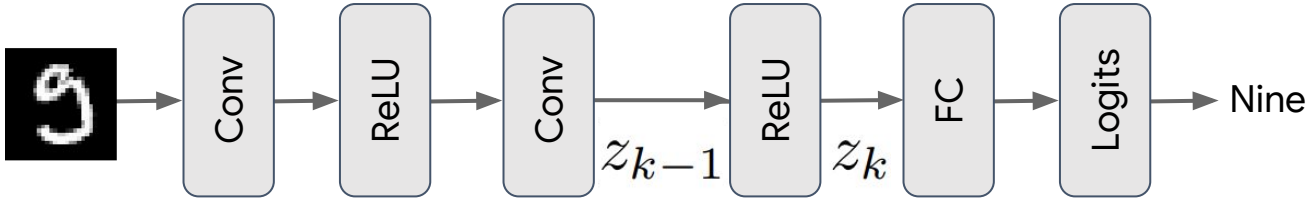


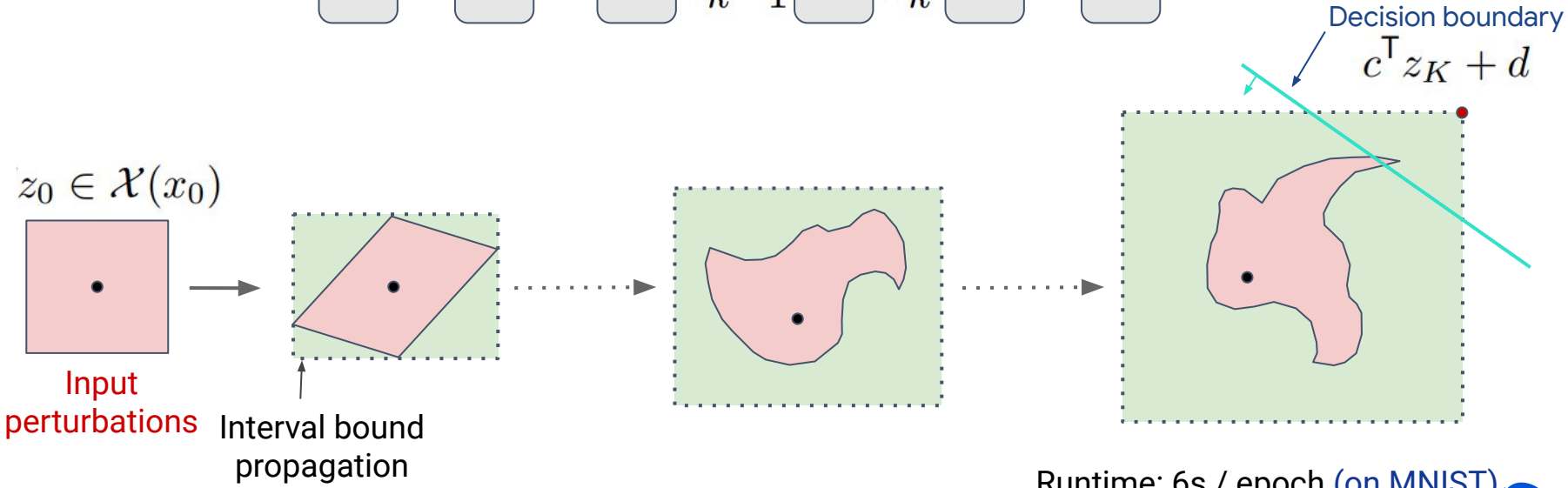
Bu yöntemler gürbüzlüğü garanti edebilir - pratik uygulamalar için ölçeklenmeleri gerekiyor



Runtime (Wong et al.): 160s / epoch (on MNIST)

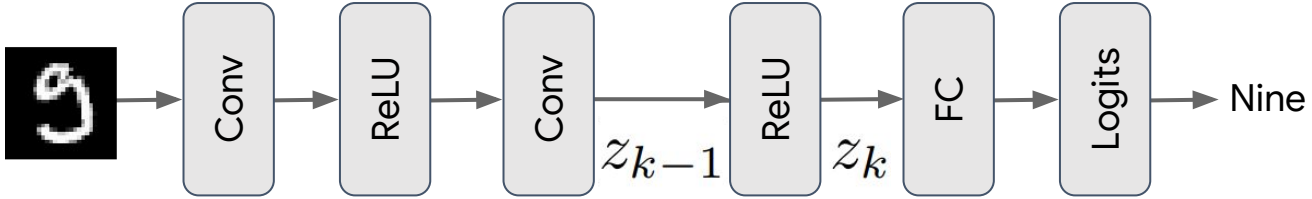
Interval Bound Propagation (IBP) (aralık yayma)

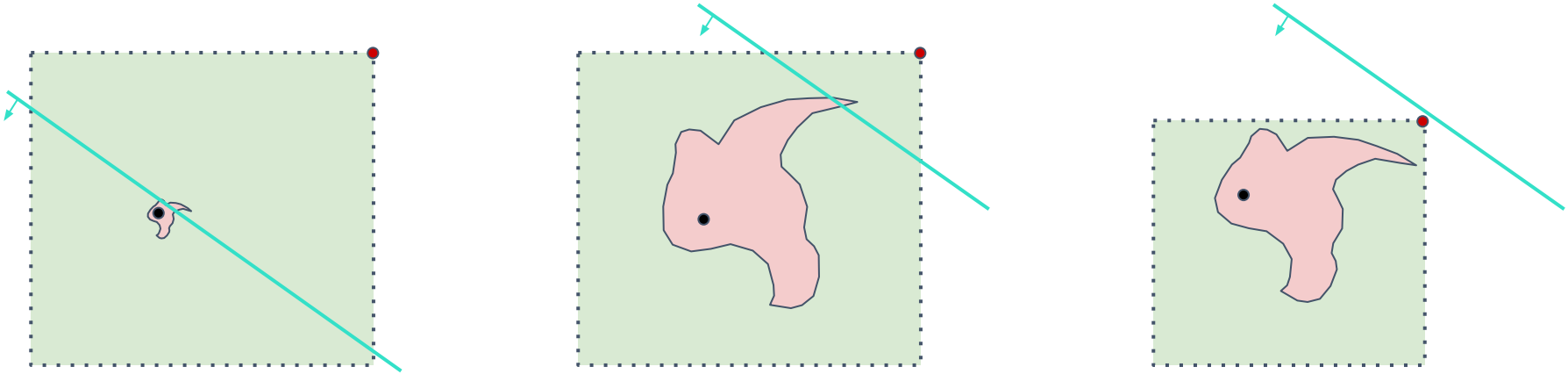




Runtime: 6s / epoch (on MNIST)

IBP





Runtime: 6s / epoch (on MNIST)

On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, Pushmeet Kohli

(Submitted on 30 Oct 2018 (v1), last revised 28 Jan 2019 (this version, v3))

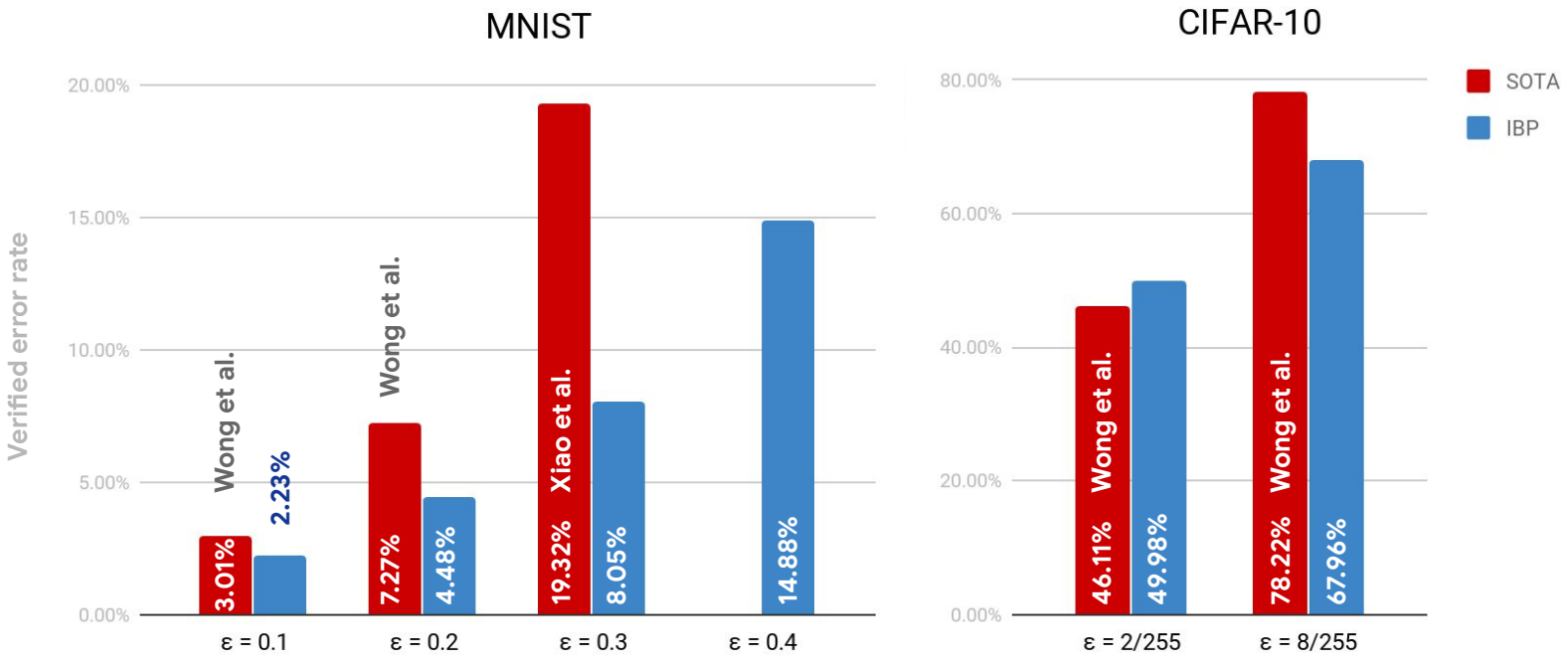
Different model sizes across a wide range of ε on MNIST and CIFAR-10

- Paper also includes results on SVHN and ImageNet (64x64)
- IBP is the only model that provides non-vacuous bounds on a non-trivial ImageNet model (i.e., WideResnet-10-10 with 1.8M hidden units) at ε = 1/255.

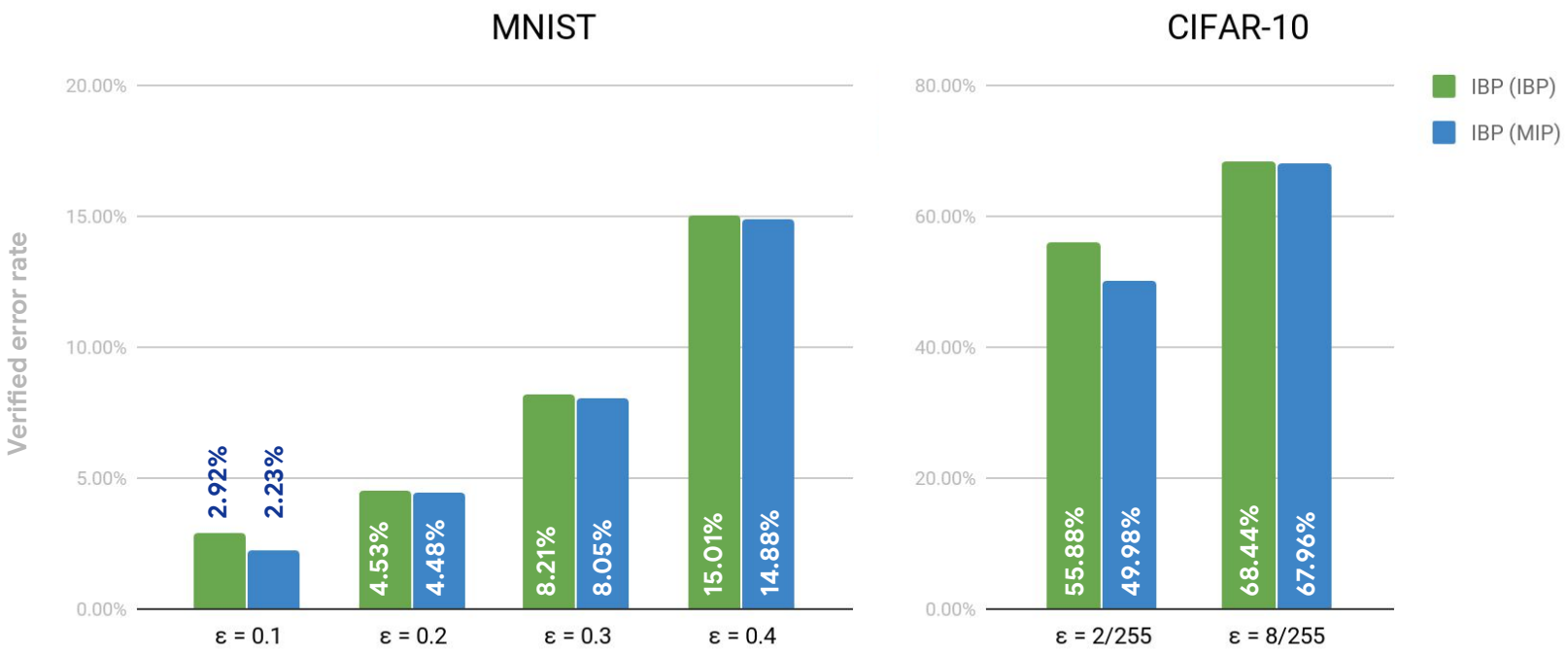
model	small	medium	large
# hidden units*	8.3K	47K	230K
# parameters*	471K	1.2M	7M

* On CIFAR-10

Doğrulanmış hata oranı



Tightness


MNIST
CIFAR-10
IBP (IBP)
IBP (MIP)
Verified error rate
20.00%
15.00%
10.00%
5.00%
0.00%
80.00%
60.00%
40.00%
20.00%
0.00%
2.92%
2.23%
4.53%
4.48%
8.21%
8.05%
15.01%
14.88%
55.88%
49.98%
68.44%
67.96%
ε = 0.1
ε = 0.2
ε = 0.3
ε = 0.4
ε = 2/255
ε = 8/255

Conclusion

- Despite the fact that IBP provides loose bounds on arbitrary networks, the models can adapt to make the IBP bound tight.

- IBP is the fastest, most scalable method that achieves SOTA verified accuracy on MNIST, CIFAR-10, SVHN and ImageNet (64x64).

- Open issues:
  - IBP reduces the expressivity of models - much sparser models/underfitting
  - Not easy to extend to other input constraints while preserving good performance

DeepMind

3

Sağlam Gösterimler (Robust Representations)

Motivasyon

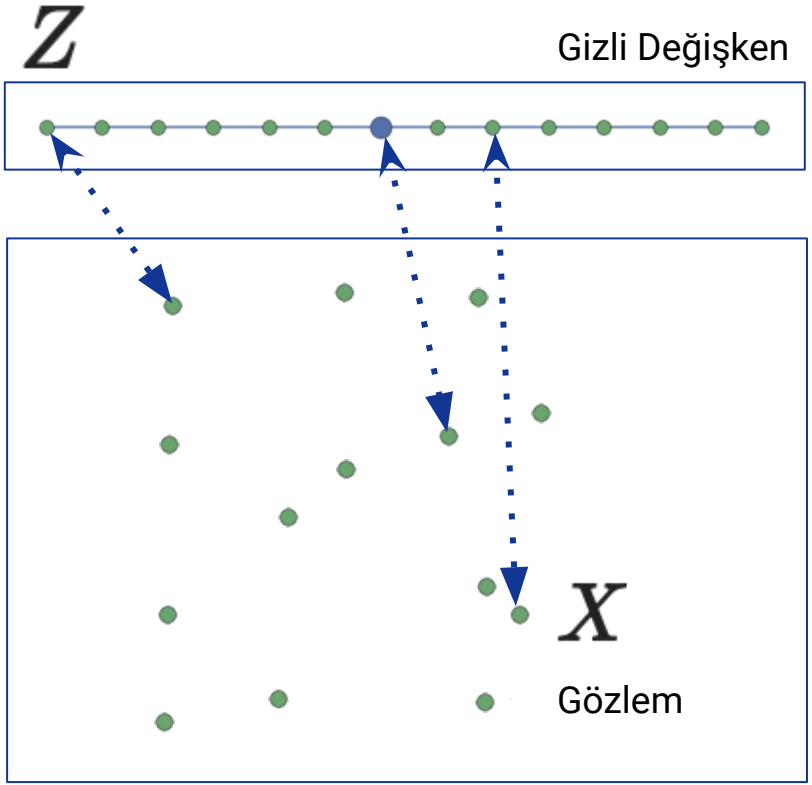
Private & Confidential

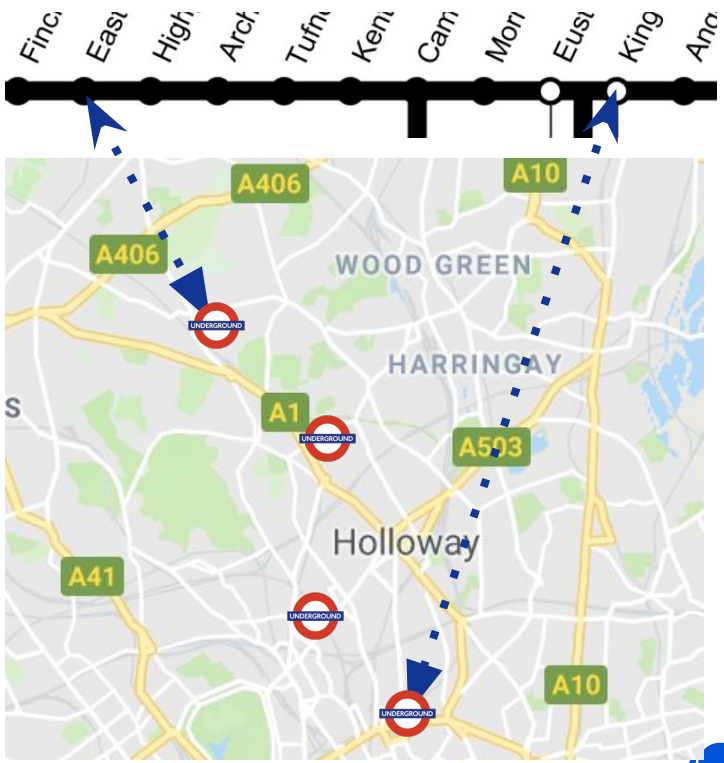
- **Güdümsüz** Sağlam gösterim öğrenmesi (Unsupervised Robust Representation)
  - Genel, herhangi bir belirlenmiş problem olmadan öğrenme
    - Metin: `word2vec, ELMO, BERT`
    - İlikisel veri: `GraphNET`

- Gösterim => Soyutlama yeteneği => (Yapay) zekanın temel taşı

- Gösterimler ile ilgili sorular
  - **İsabetli?**
  - **Sağlam?**
  - Doğrulanabilir?
  - Mahremiyet gözeten? (veri çalınabilir mi?)
  - Yansız? (gizli önyargılar içeriyor mu?)

Gösterim nedir?

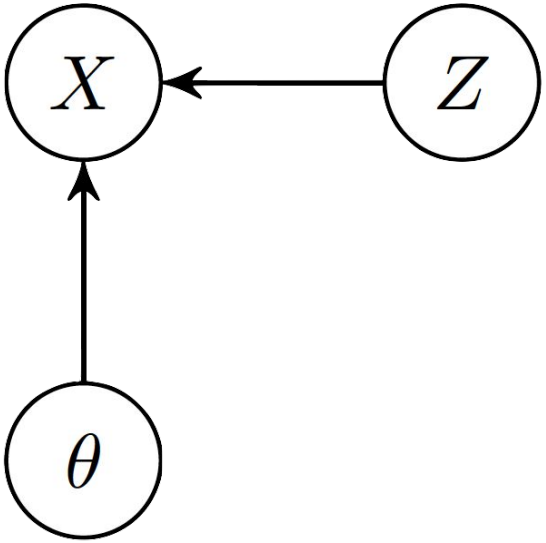




Varyasyonel Otokodlayıcı (Variational AutoEncoder - VAE)

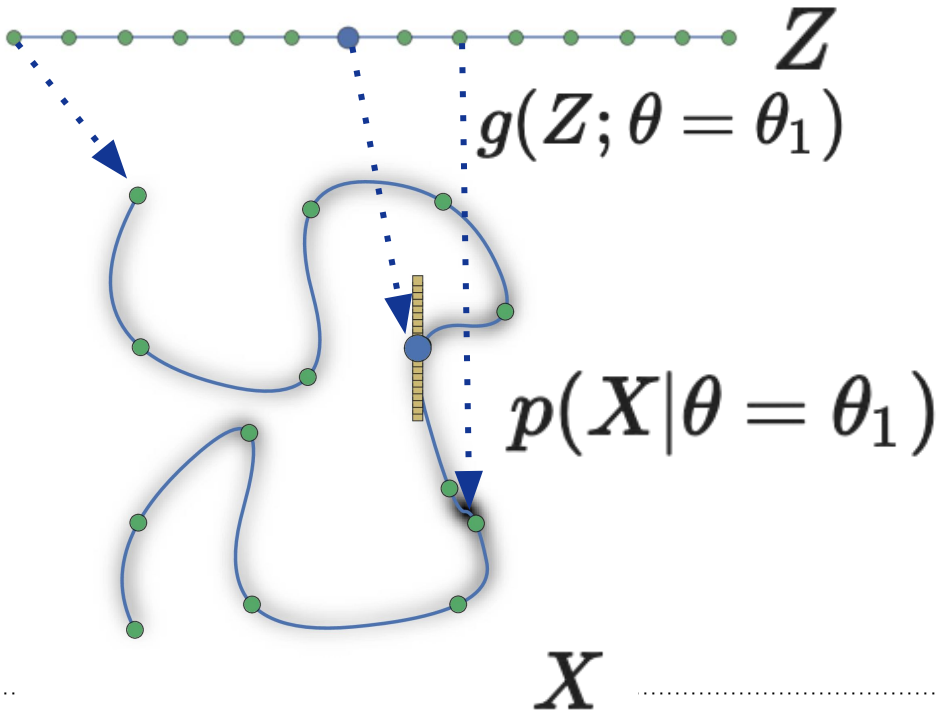
(Kingma & Welling 2014, Rezende, Mohamed & Wierstra, 2014)

$\mathcal{N}(X; g(Z;\theta), vI)$ $\mathcal{N}(Z; 0, I)$

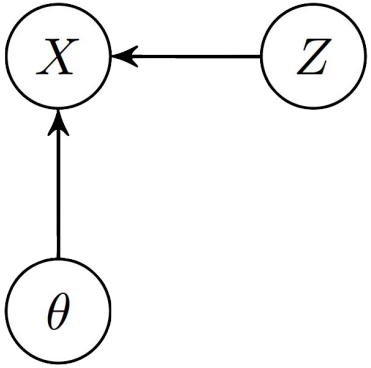


$$p(X|\theta) = \int p(X|Z,\theta)p(Z)dZ$$

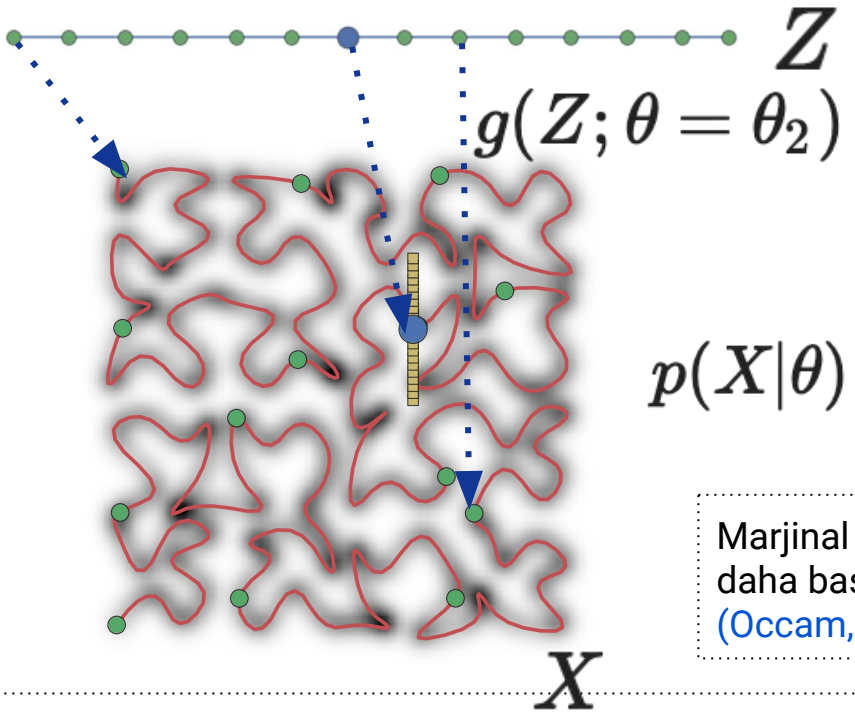
İyi bir gösterim nedir?



$\mathcal{N}(X; g(Z;\theta), vI)$ $\mathcal{N}(Z;0,I)$



İyi bir gösterim nedir?



$$p(X|\theta) = \int p(X|Z,\theta)p(Z)dZ$$

Marjinal olabilirlik veri uyumu benzer olanlar arasından daha basit olanını tercih eder
(Occam, Bayes, Jeffreys, Jaynes, MacKay)

Kullback-Leibler (KL) Iraksayı

$$\mathcal{KL}(Q(x)||P(x)) = \left\langle \log \frac{Q(x)}{P(x)} \right\rangle_{Q(x)} = \int dx Q(x) \log \frac{Q(x)}{P(x)}$$

Marjinal Olabilirliğin Eşlek (dual) Yapısı

$$\log p(X|\theta) = \langle \log p(X|Z,\theta) \rangle_{p(Z|X,\theta)} - \mathcal{KL}(p(Z|X,\theta)||p(Z))$$

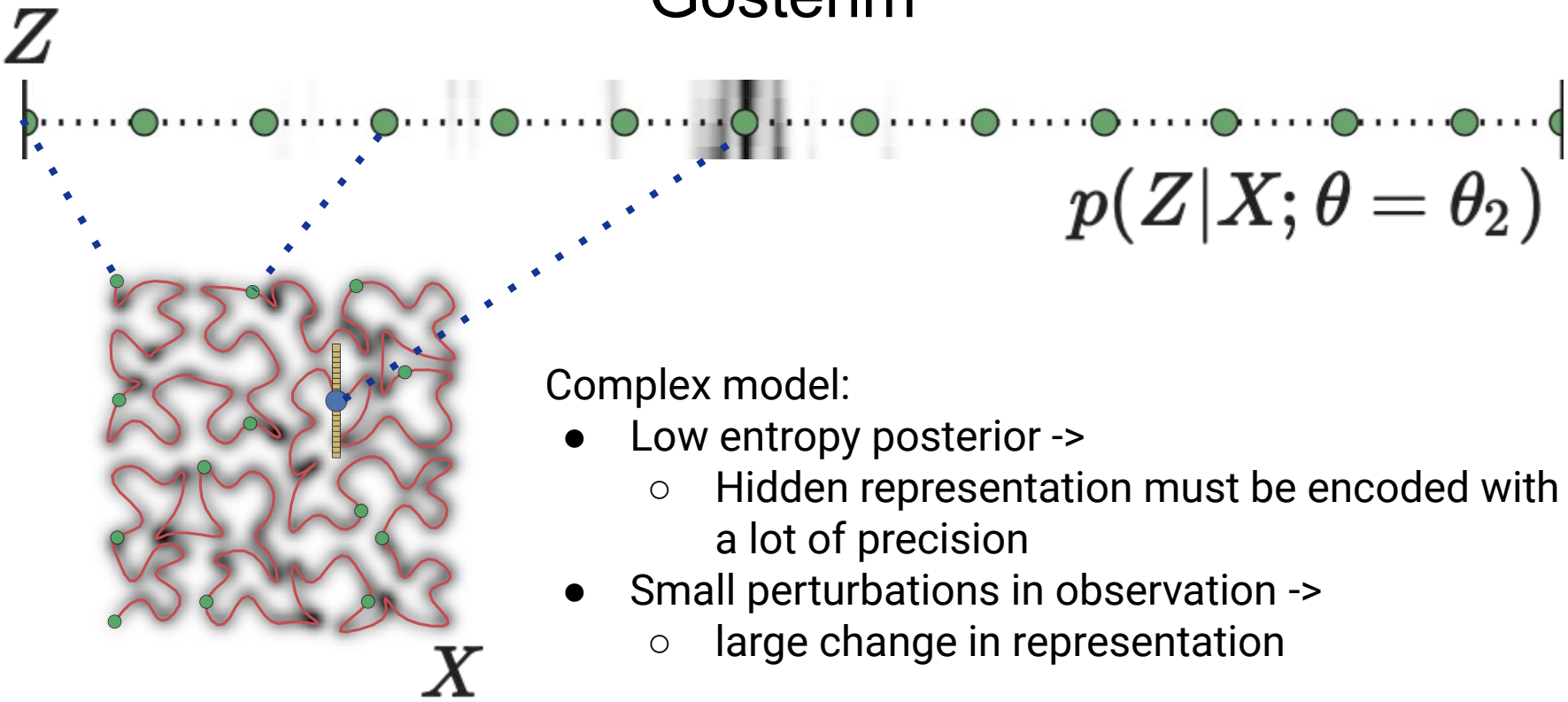
Veri Uyumu + Önsel uyum+Entropi

Gerçek sonsal dağılım ideal gürbüz gösterim

Bilinmeyen bir parametrenin (makul zamanda) **hesaplanamayan** fonksyonu

$$p(Z|X;\theta = \theta^*)$$

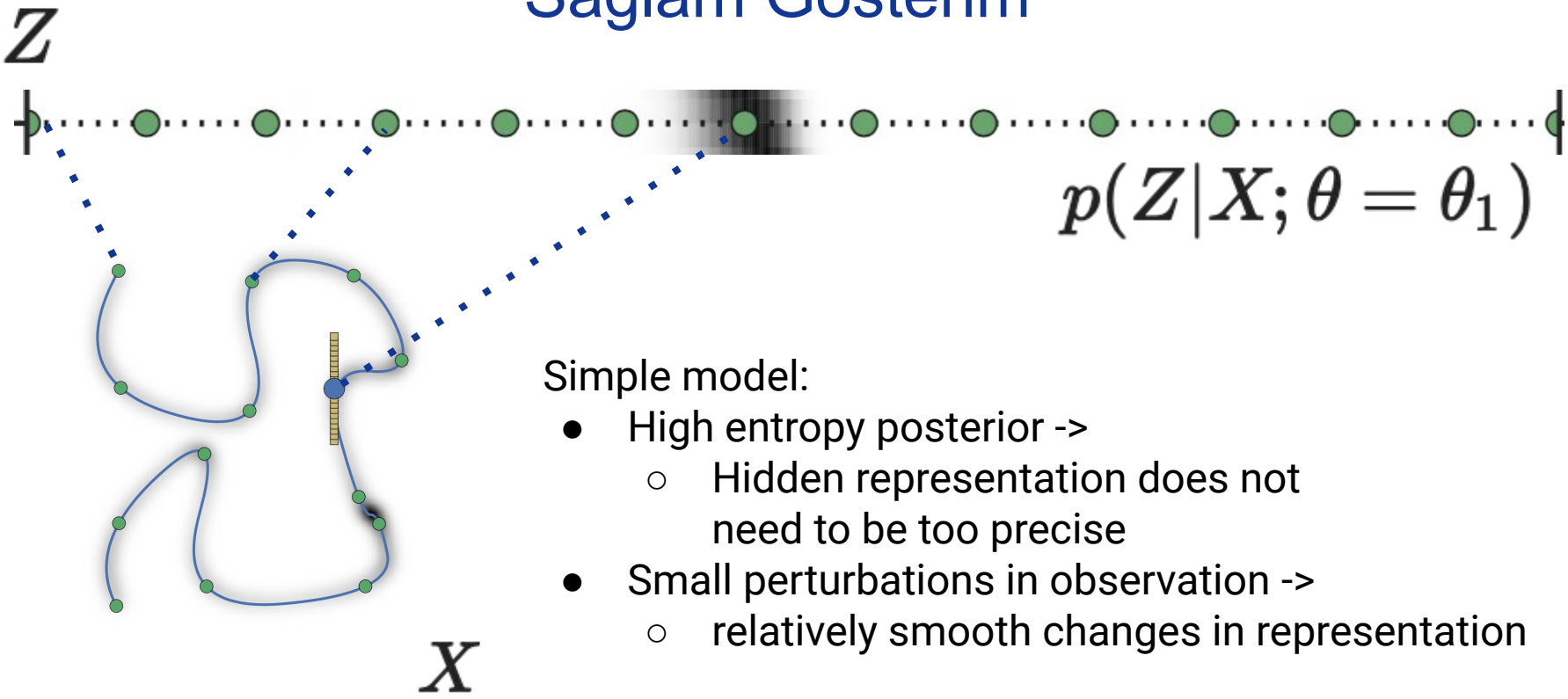
Gösterim



Complex model:

- Low entropy posterior ->
  - Hidden representation must be encoded with a lot of precision
- Small perturbations in observation ->
  - large change in representation

Sağlam Gösterim



Simple model:

- High entropy posterior ->
  - Hidden representation does not need to be too precise
- Small perturbations in observation ->
  - relatively smooth changes in representation

Varyasyonel Çıkarım (Variational Inference)

Wainwright and Jordan, 2008

Üretici modeli ve çıkarım algoritmasını aramak

$$\log p(X = x|\theta) \geq \langle \log p(X = x|Z, \theta) \rangle_{Q(Z)} - \mathcal{KL}(Q(Z)||p(Z))$$

Evidence Lower Bound (ELBO)

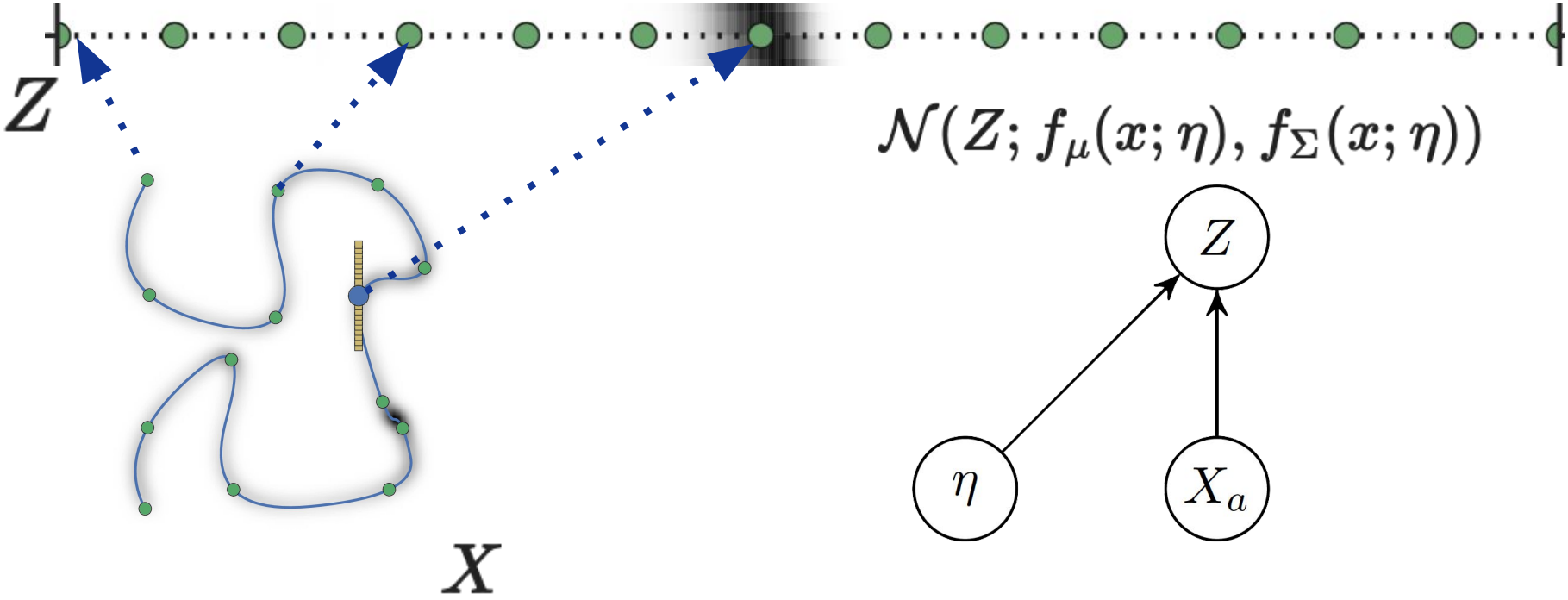
ELBO = Veri uyumu + Önsel dağılım uyumu +Entropi

Bütün dağılımlar için geçerli

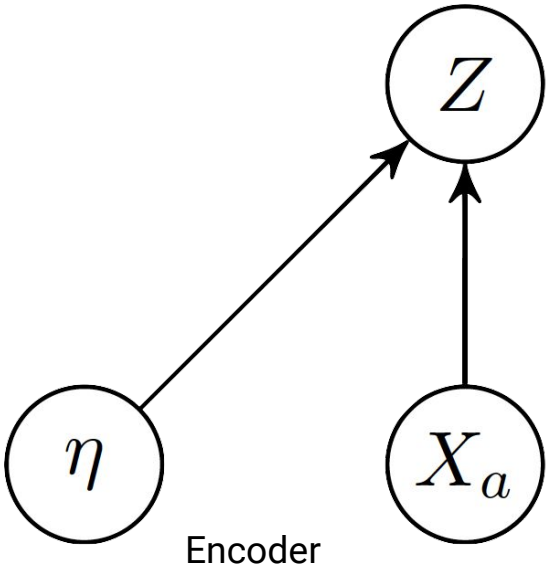
$$Q(Z)$$

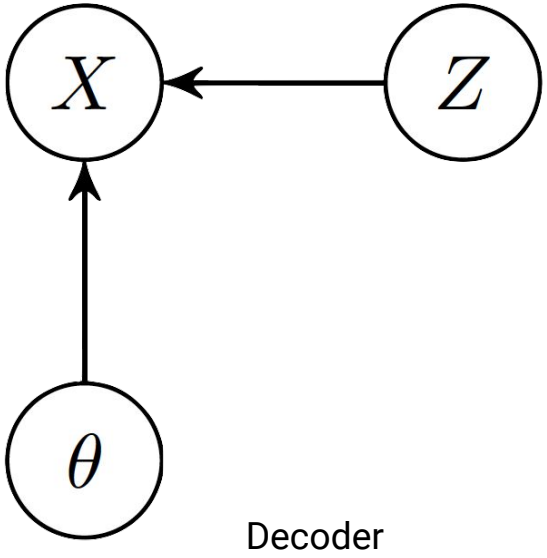
Kodlayıcı

(Kingma & Welling 2014, Rezende, Mohamed & Wierstra, 2014)



VAE (Encoder and Decoder)

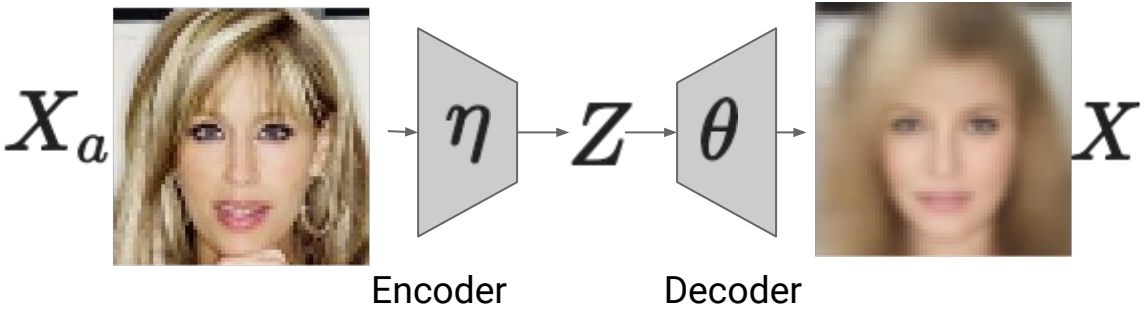


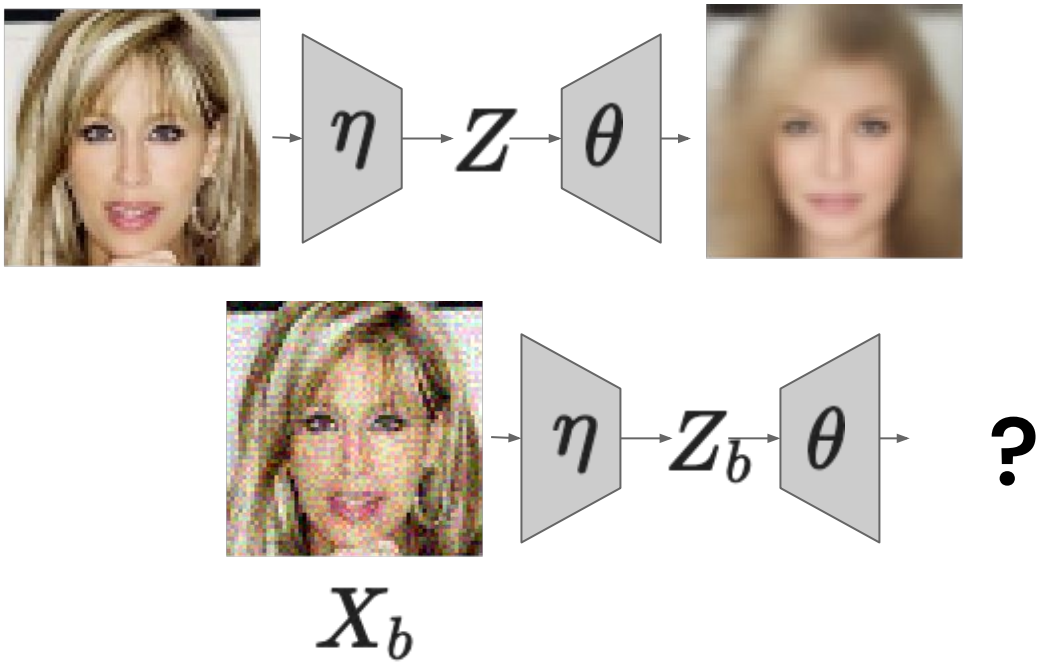


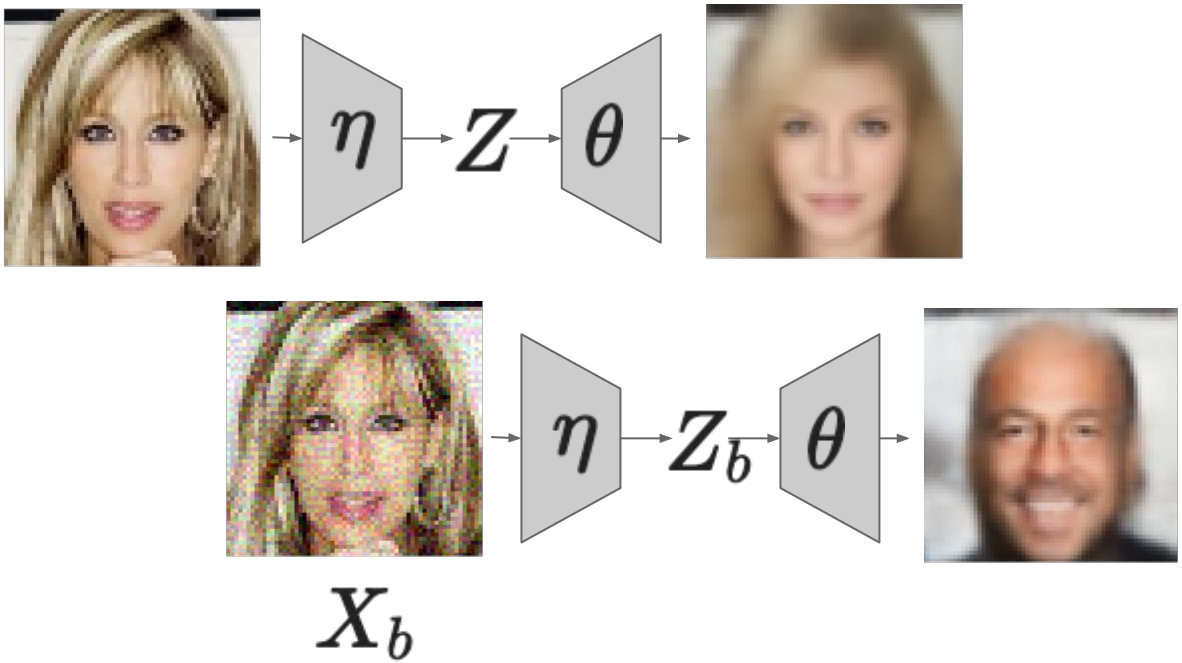
$$\log p(X = x|\theta) \geq \langle \log p(X = x|Z, \theta) \rangle_{q(Z|X_a = x, \eta)} - \mathcal{KL}(q(Z|X_a = x, \eta)||p(Z))$$

VAE ile öğrenilen gösterimler gürbüz mü?

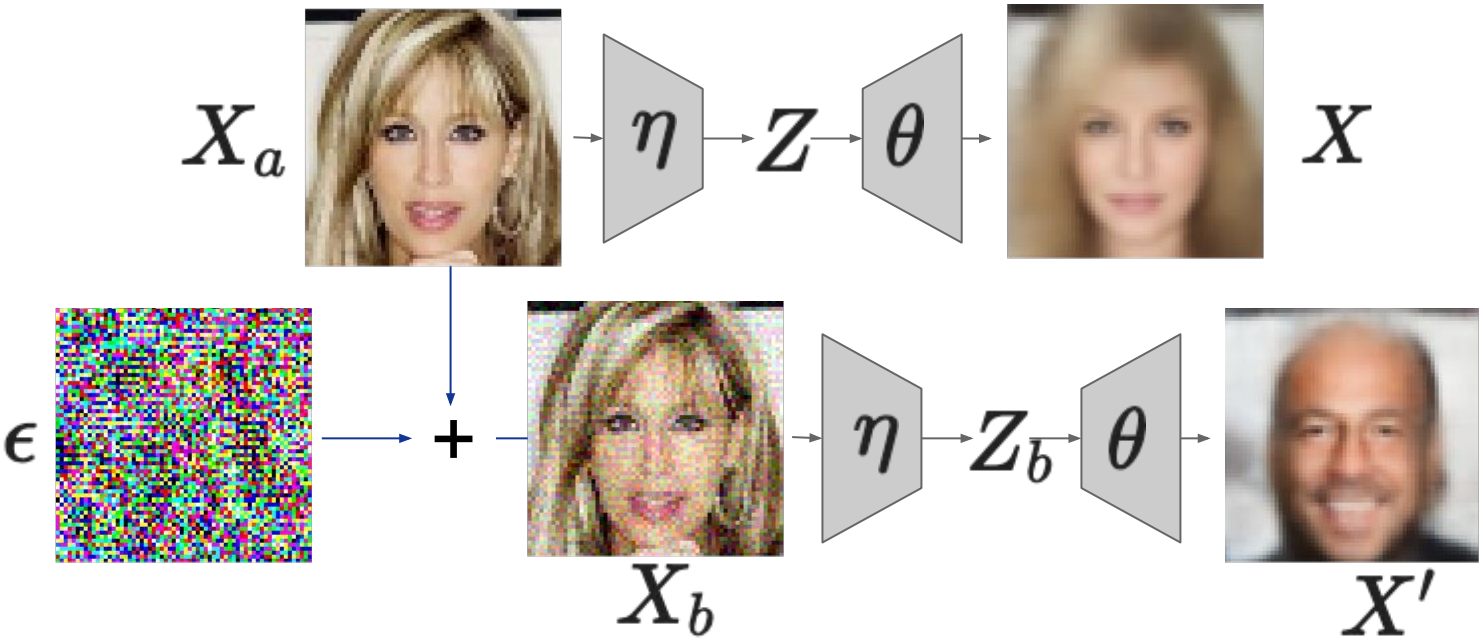
(Kingma & Welling 2014, Rezende, Mohamed & Wierstra, 2014)




η
Z
θ
η
Z_b
θ
?
X_b

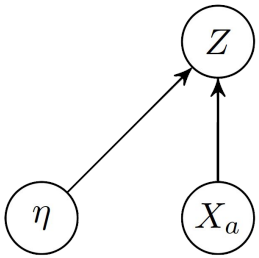

η
Z
θ
η
Z_b
θ
X_b

Is a VAE Robust to Adversarial Perturbations?

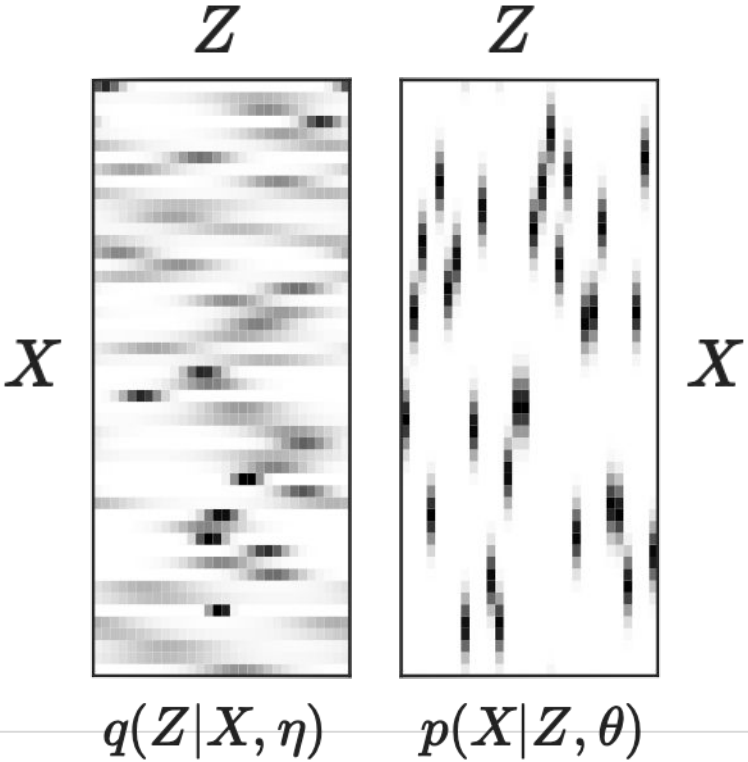


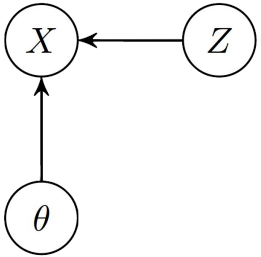
Unfortunately not

A Discrete VAE



Encoder

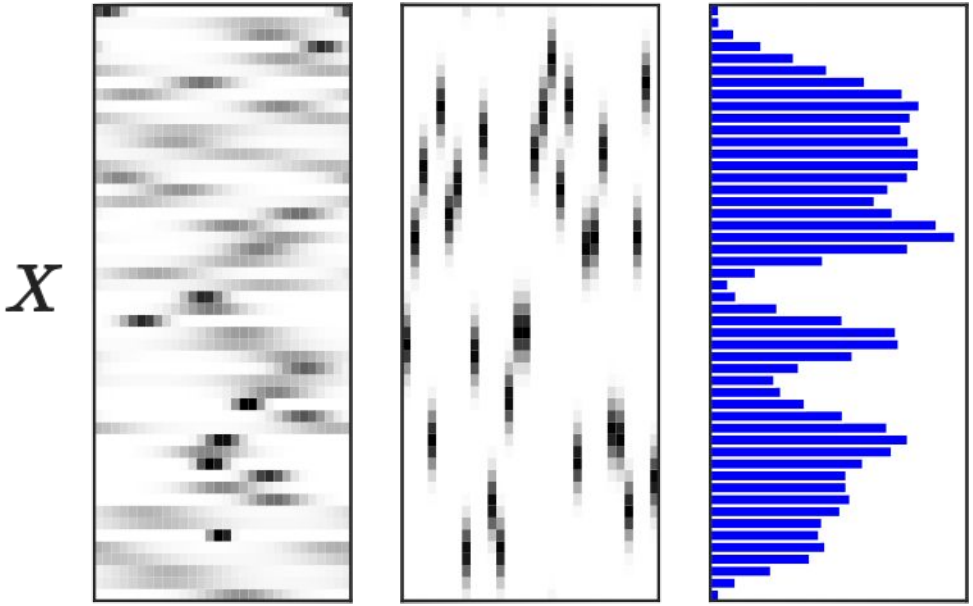




Decoder

A Discrete VAE

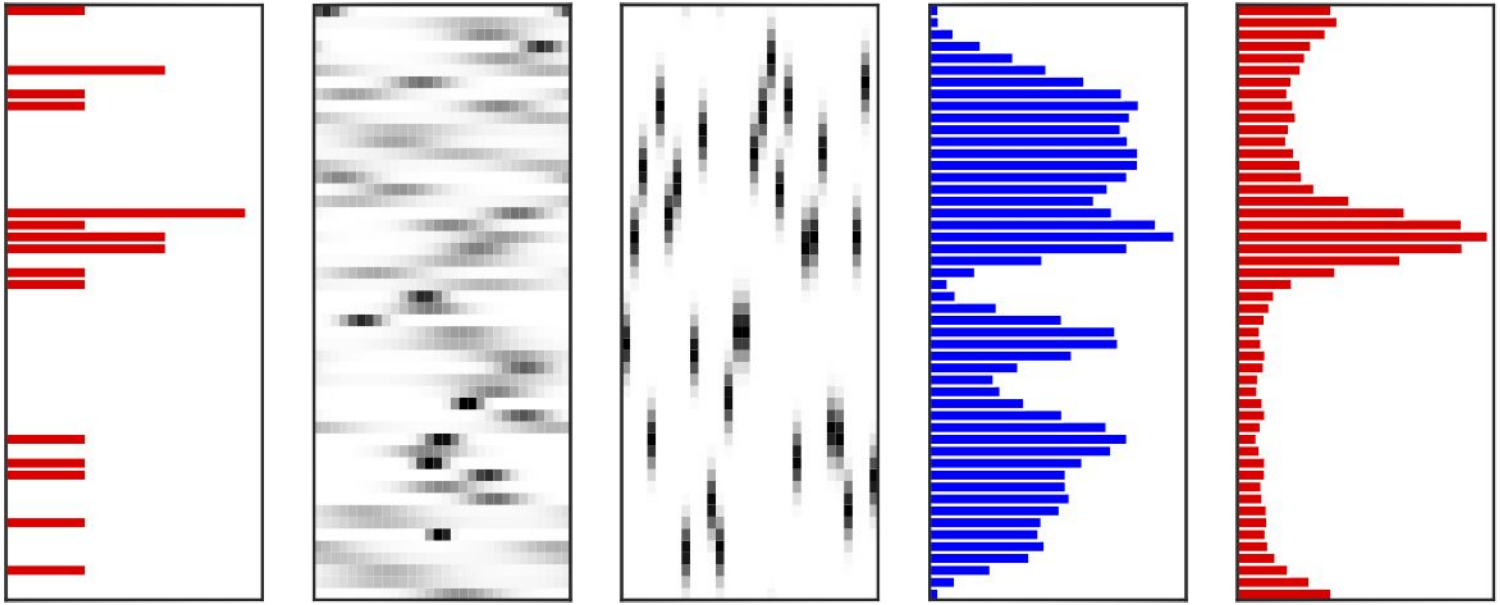
$$p(X|\theta) = \sum_Z p(X|Z,\theta)p(Z)$$



$q(Z|X,\eta)$ $p(X|Z,\theta)$

Fitting a VAE - Initialization

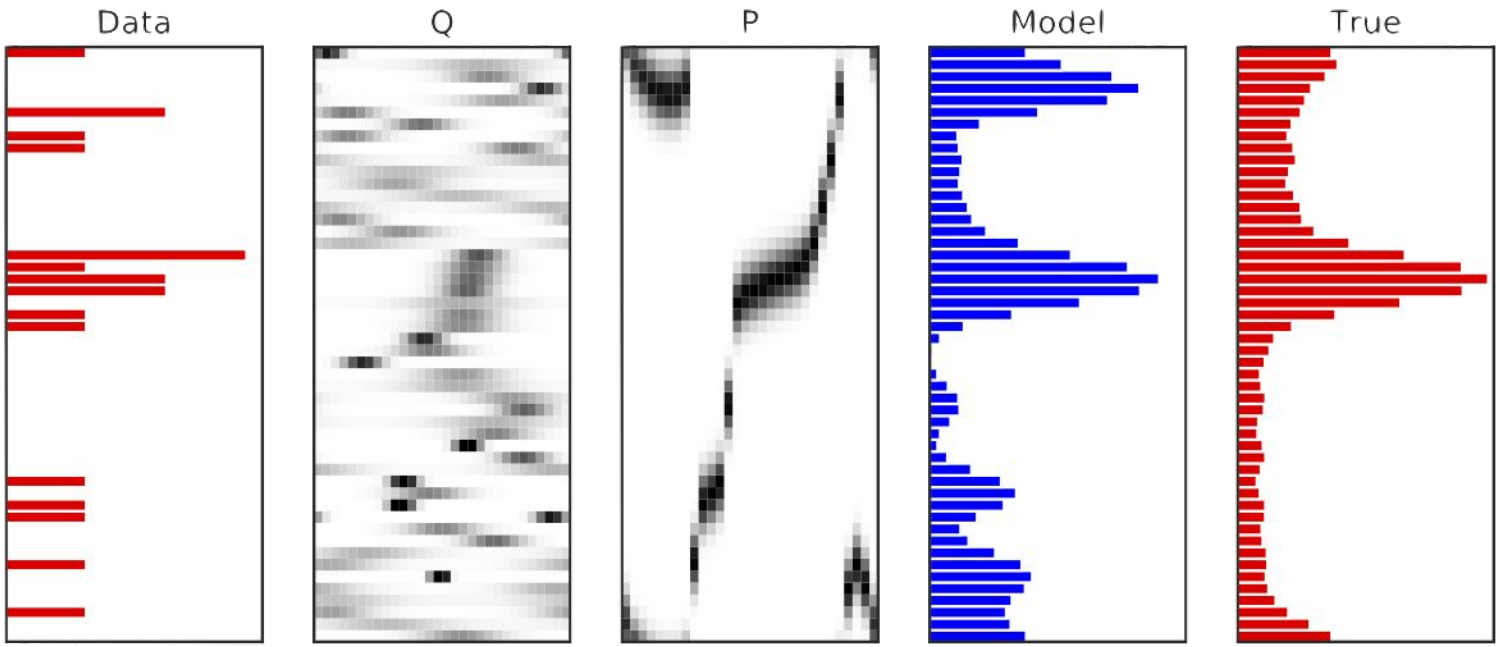
minimize $\mathcal{KL}(\hat{\pi}(X)q(Z|X,\eta)||p(X|Z,\theta)p(Z))$



$\hat{\pi}(X)$ $q(Z|X,\eta)$ $p(X|Z,\theta)$ $p(X|\theta)$ $\pi(X)$

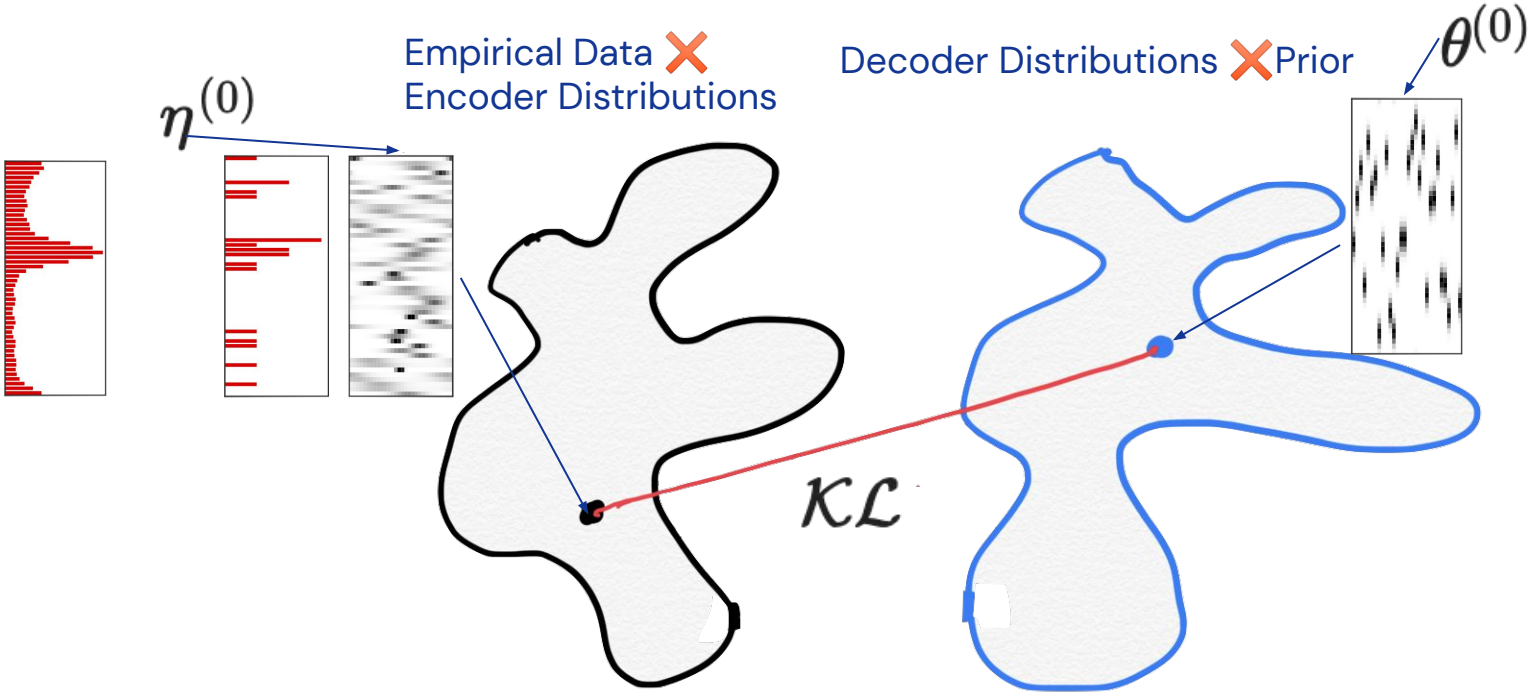
VAE - After Training

minimize $\mathcal{KL}(\hat{\pi}(X)q(Z|X,\eta)||p(X|Z,\theta)p(Z))$



$\hat{\pi}(X)$ $q(Z|X,\eta)$ $p(X|Z,\theta)$ $p(X|\theta)$ $\pi(X)$

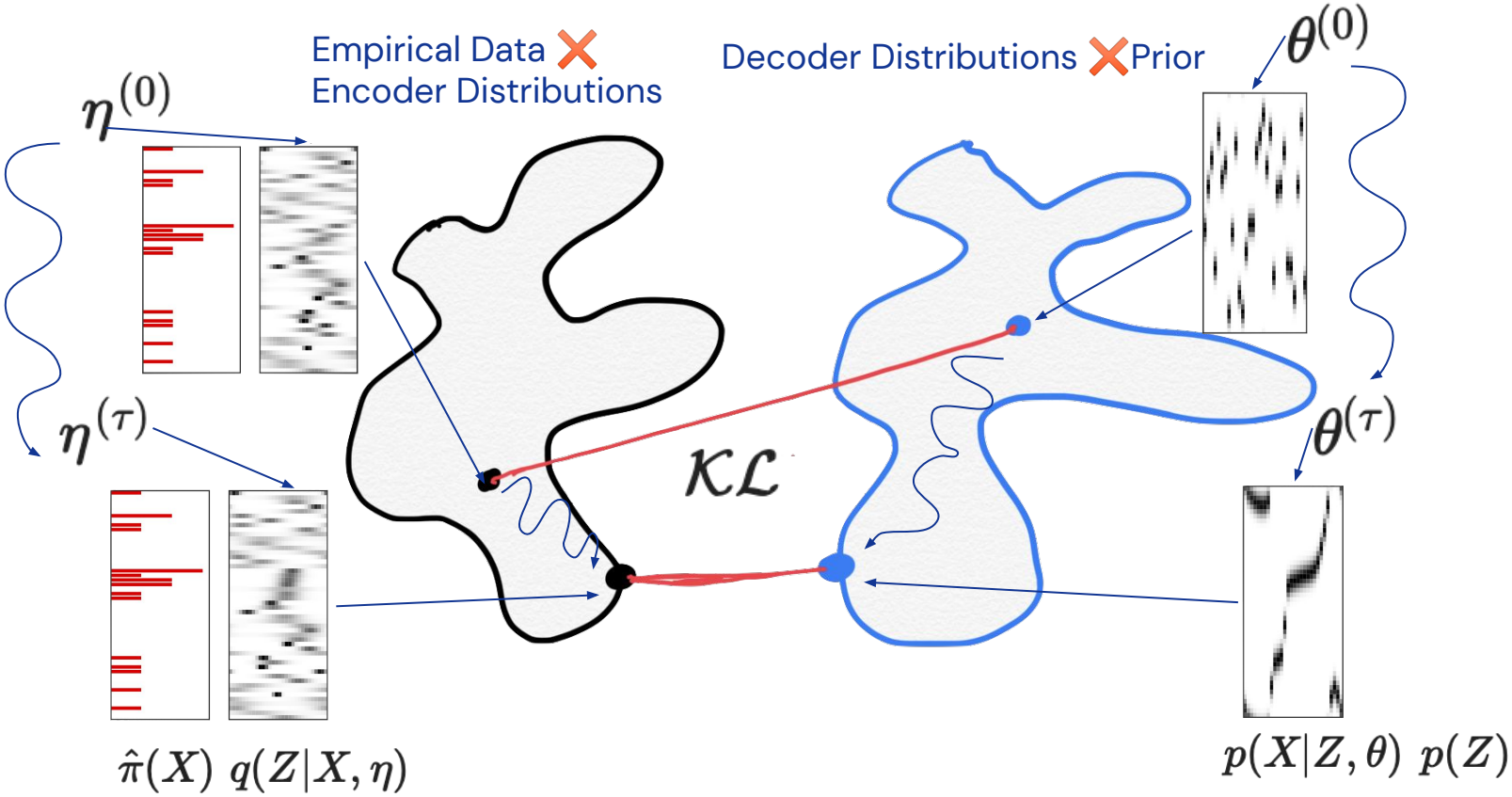
VAE Training - Information Geometric Picture



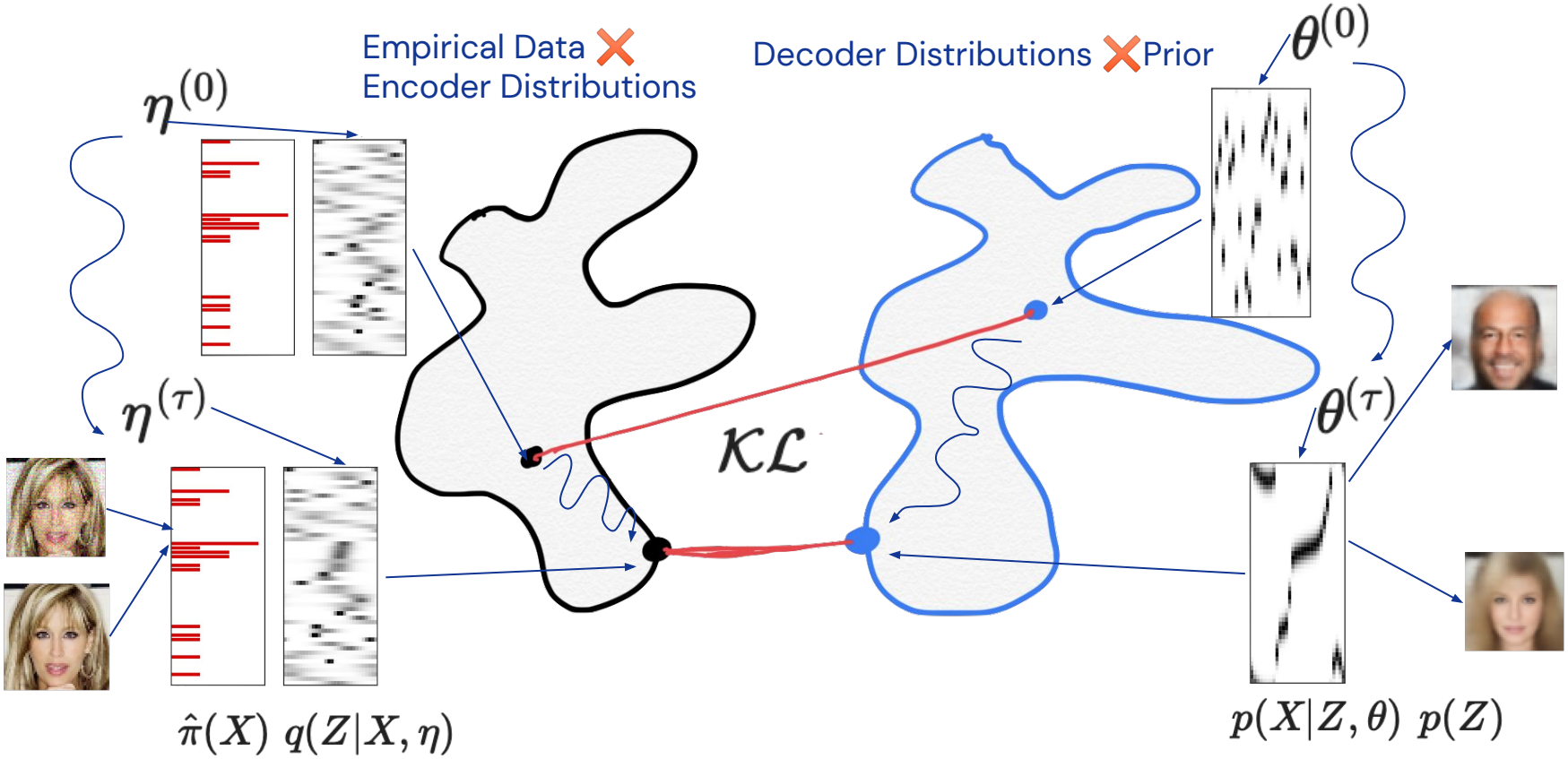
$\hat{\pi}(X)\ q(Z|X,\eta)$

$p(X|Z,\theta)\ p(Z)$

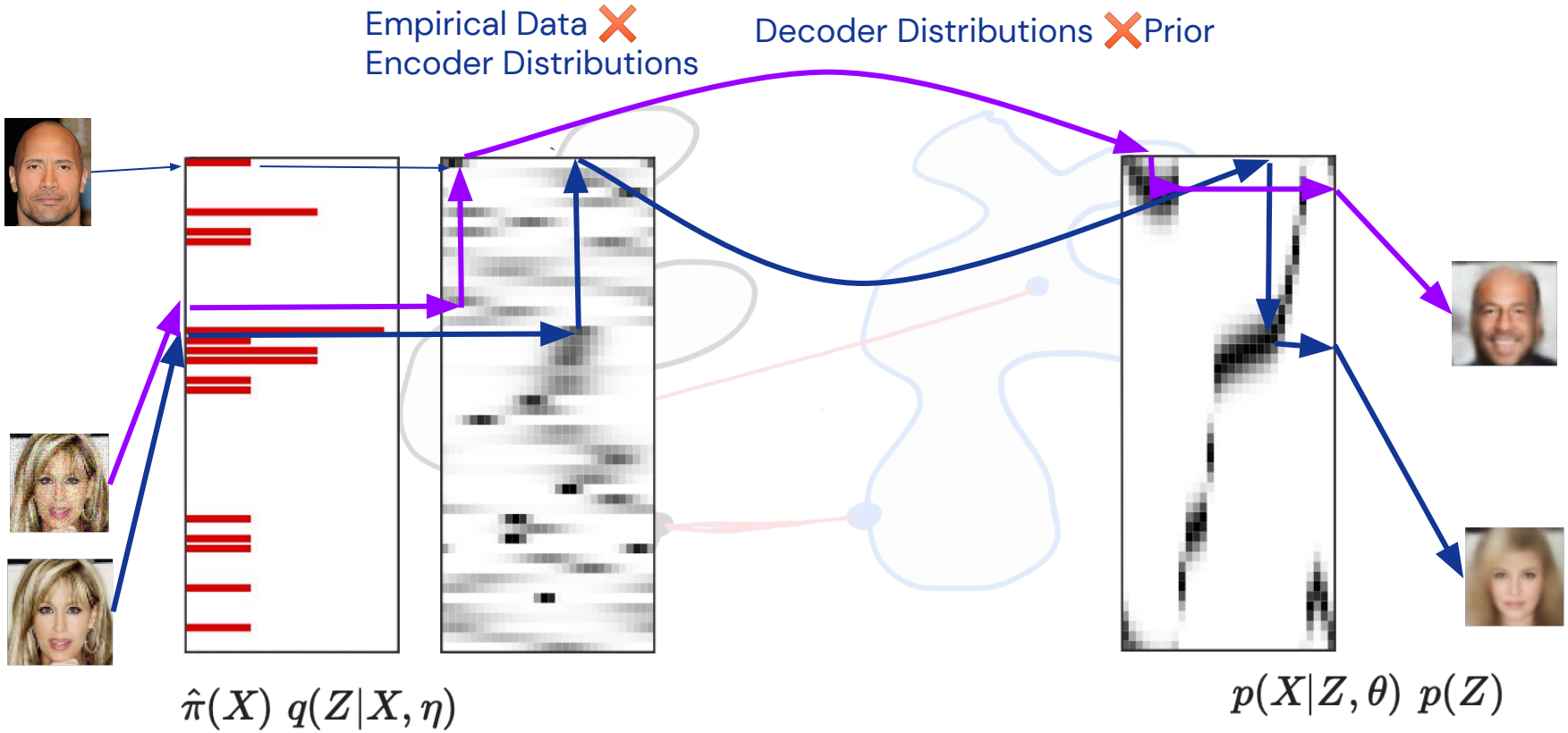
VAE Training - Information Geometric Picture



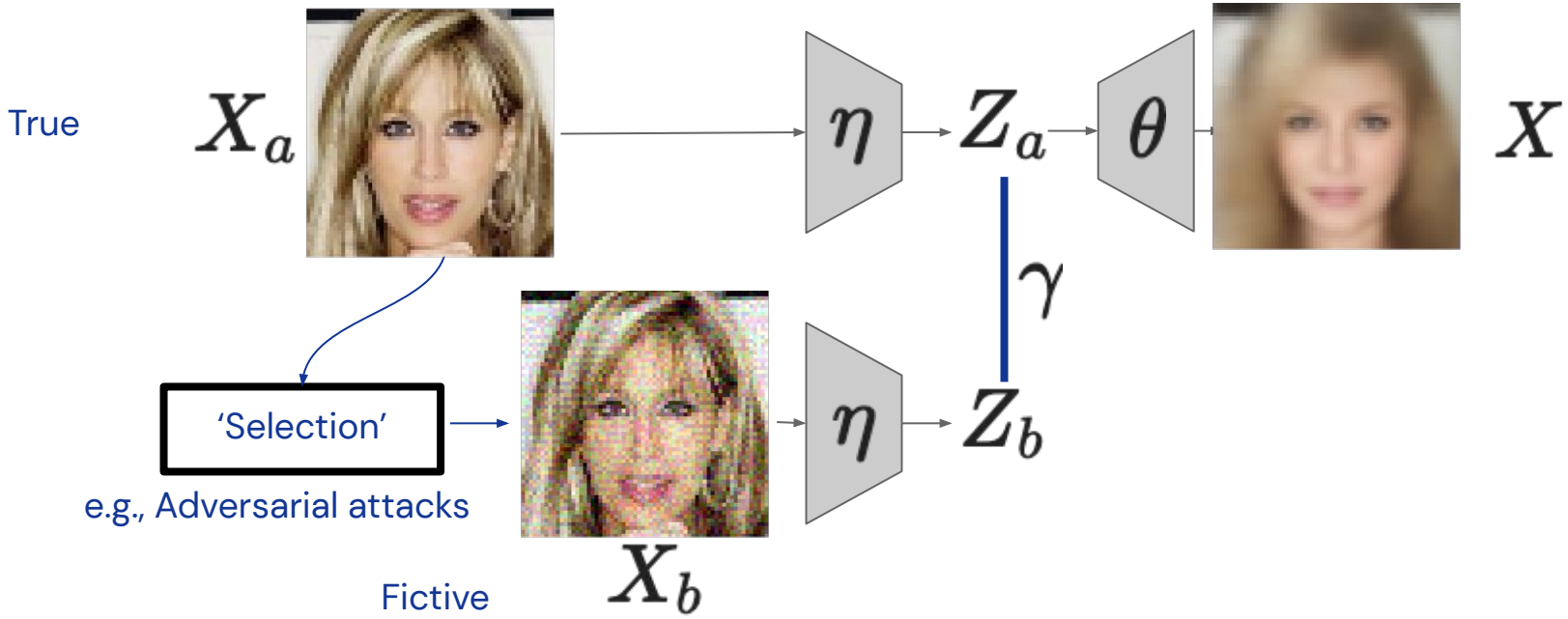
VAE Training - Information Geometric Picture



VAE Training - Information Geometric Picture



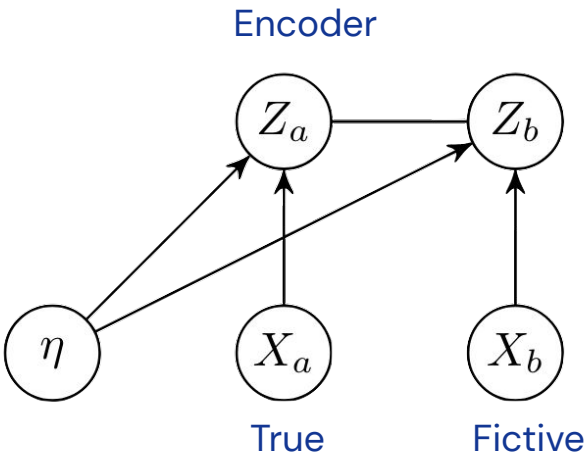
Smooth Encoders (SE)

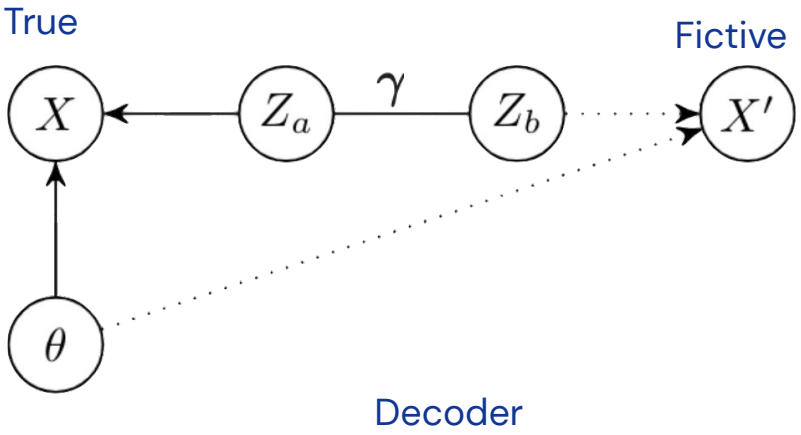


$$\|X_a - X_b\|_p < \epsilon \Rightarrow \| \langle Z_a | X_a \rangle - \langle Z_b | X_b \rangle \|_2 < \gamma$$

Smooth Encoder

Key Idea: Bias the search towards encoders with a smoother representation





- Target for the encoder: the extended distribution where we integrate out the fictive sample but not the corresponding representation
- Fictive samples come from an external selection mechanism

VAE/SE amaç fonksyonlarının karşılaştırılması

Veri uyumu — Önsel dağılım uyumu

$$\mathcal{B}_{\text{VAE}}(\eta,\theta) = \boxed{\langle \log p(X=x|Z_a,\theta)\rangle_{q(Z_a|x_a,\eta)}} + \boxed{\langle \log p(Z_a)\rangle_{q(Z_a|x_a,\eta)}} \boxed{+H(q(Z_a|x_a,\eta))}$$

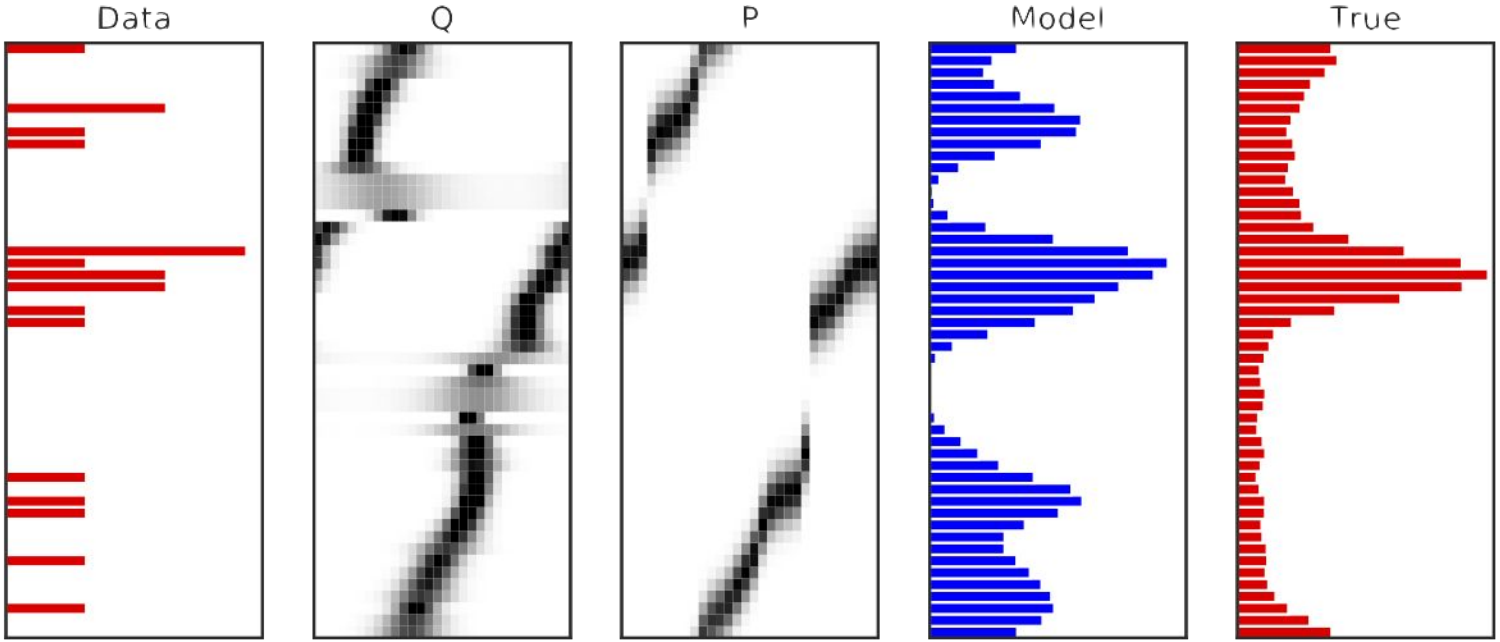
Entropi

Veri uyumu — Önsel dağılım uyumu

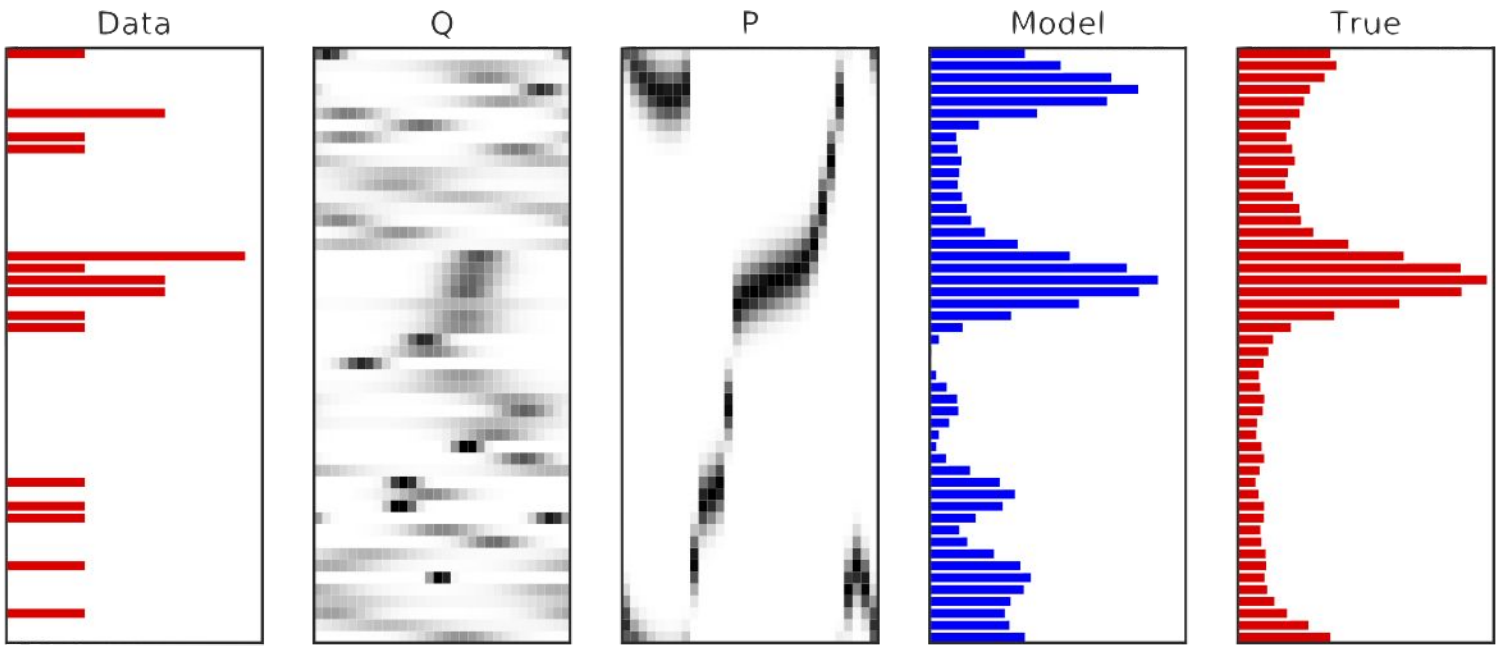
$$\mathcal{B}_{SE}(\eta,\theta) = \boxed{\langle \log p(X=x|Z_a,\theta)\rangle_{q(Z_a|x_a,\eta)}} + \boxed{\langle \log p(Z_a)\rangle_{q(Z_a|x_a,\eta)} + \langle \log p(Z_b)\rangle_{q(Z_b|x_b,\eta)}} \boxed{-\frac{\gamma}{2}\langle c(Z_a,Z_b)\rangle_{q(Z_a,Z_b|x_a,x_b,\eta)} + H(q(Z_a,Z_b|x_a,x_b,\eta))}$$

Entropi düzenlileştirilmiş Wasserstein (Düzlük + Entropi)

Düzleştirilmiş kodlayıcı



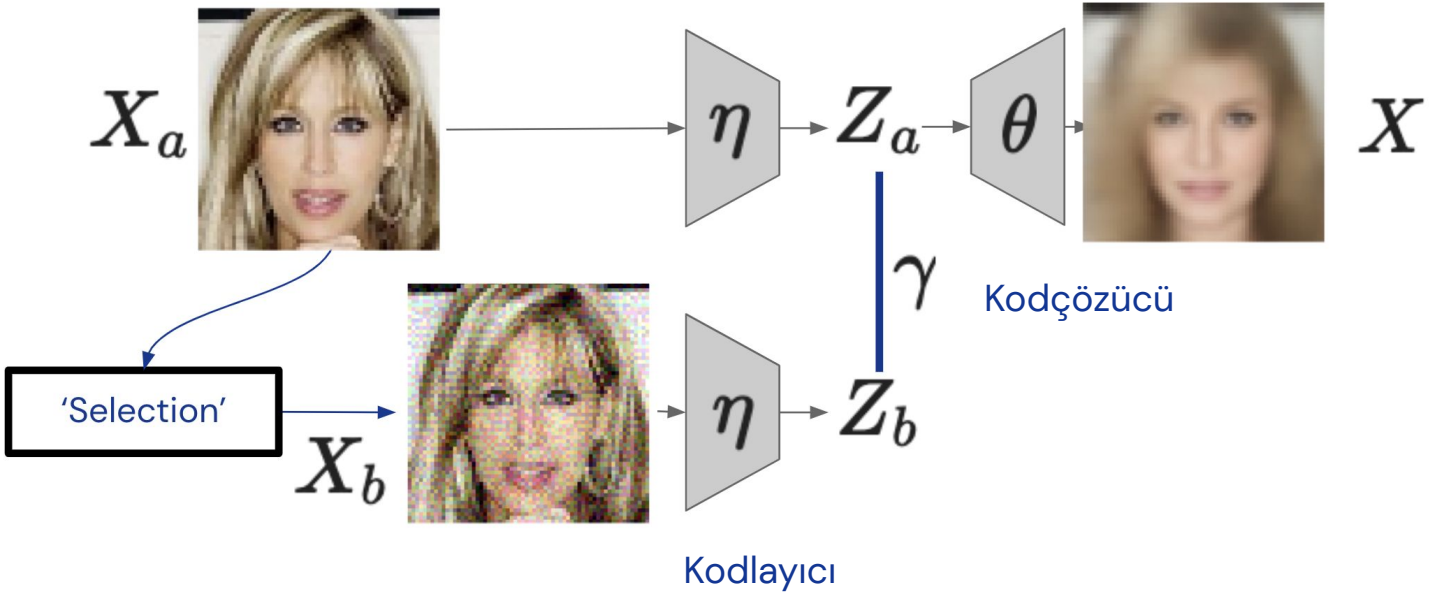
VAE



$\hat{\pi}(X)$ $q(Z|X,\eta)$ $p(X|Z,\theta)$ $p(X|\theta)$ $\pi(X)$

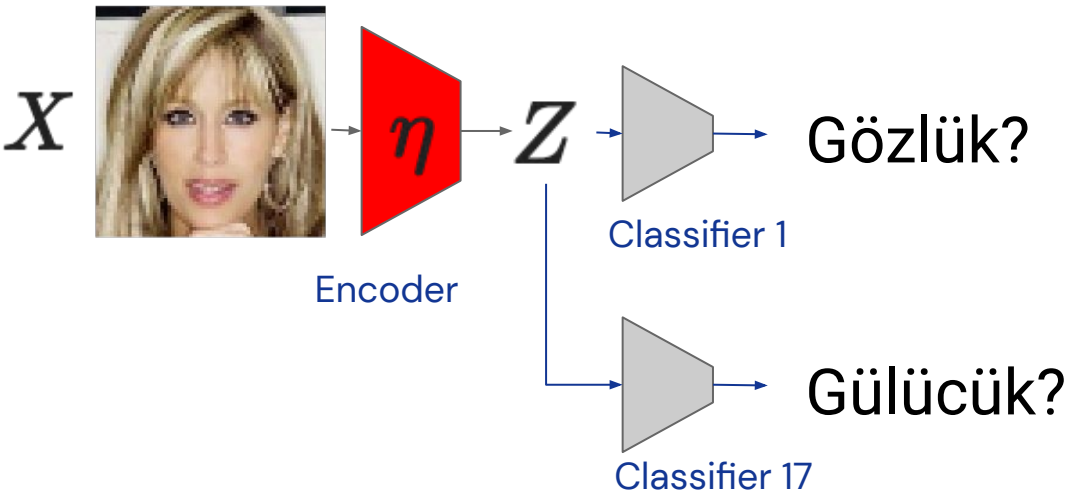
Deney Protokolü

Adım 1: Kodlayıcı-Kodçözücü çifti öğren



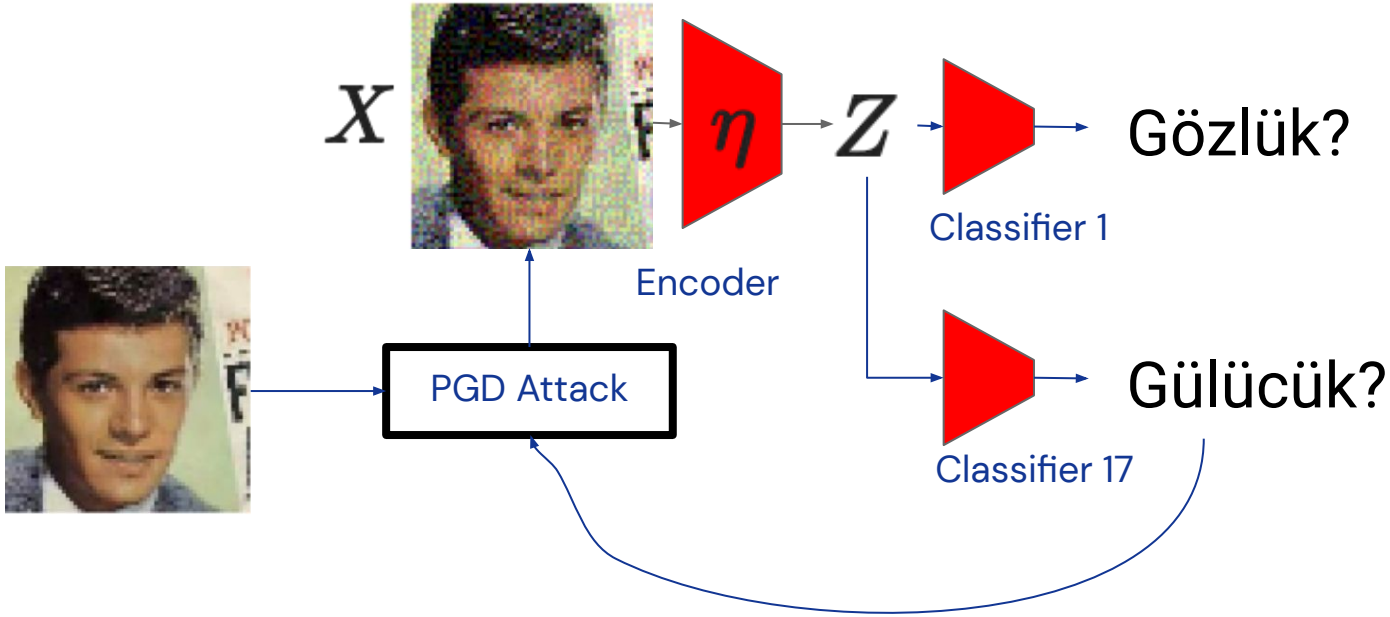
Deney Protokolü

Adım 2: Kodlayıcı parametrelerini **sabitle**, her problem için normal şekilde sınıflandırıcı öğren



Deney Protokolü

Step 3: Başarı ölçümü: Çekişmeli başarım (adversarial accuracy)
İz düşürülmüş bayır inişi (Projected Gradient Descent PGD) saldırısı ile



Sonuçlar (CelebA)

Private & Confidential

Kel?	Bıyık?	Kolye?	...	Gülme?	Gözlük?
No	No	No		Yes	No

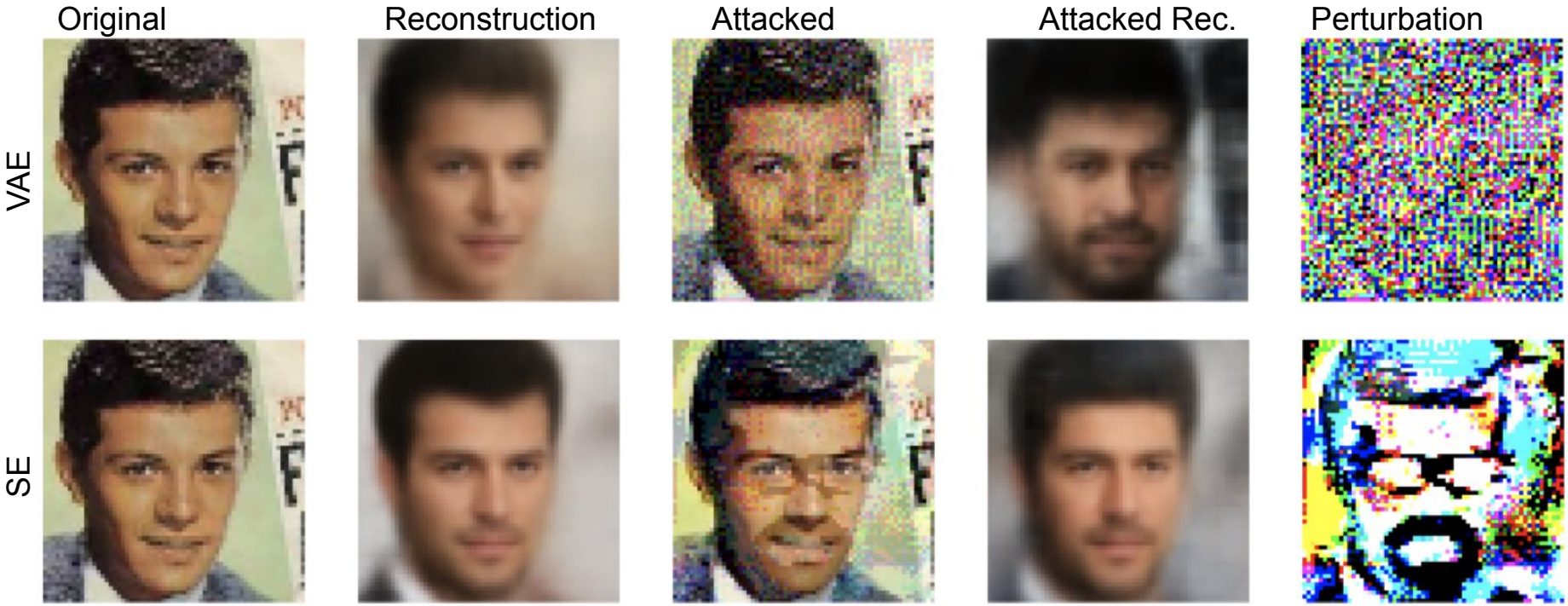
Evaluation: Nom : Attack Radius = 0.0 (No Attack)
Adv : Attack Radius = 0.1 Budget: 100 iterations with 10 restarts

	VAE		SE	
Task	Nom	Adv	Nom	Adv
Bald	97.8	0.0	97.4	86.5
Mustache	96.1	0.0	95.7	84.4
Necklace	86.2	0.0	88.0	78.9
Straight Hair	79.1	0.0	78.7	77.3
Wearing Hat	96.6	0.0	96.4	77.3
Earrings	79.3	0.0	81.3	55.3
Blond Hair	92.4	0.0	90.7	53.5
Necktie	93.2	0.0	92.7	51.7

	VAE		SE	
Task	Nom	Adv	Nom	Adv
Brown Hair	83.1	0.0	80.5	41.5
Eyeglasses	95.6	0.0	95.7	33.0
Black Hair	79.7	0.0	81.4	31.4
Bangs	90.3	0.0	89.6	27.0
No Beard	86.5	0.0	85.3	24.3
Wavy Hair	71.4	0.0	72.8	10.2
Smiling	82.7	0.0	85.7	1.1
Gender	81.1	0.0	81.6	0.7
Lipstick	79.2	0.0	80.3	0.6

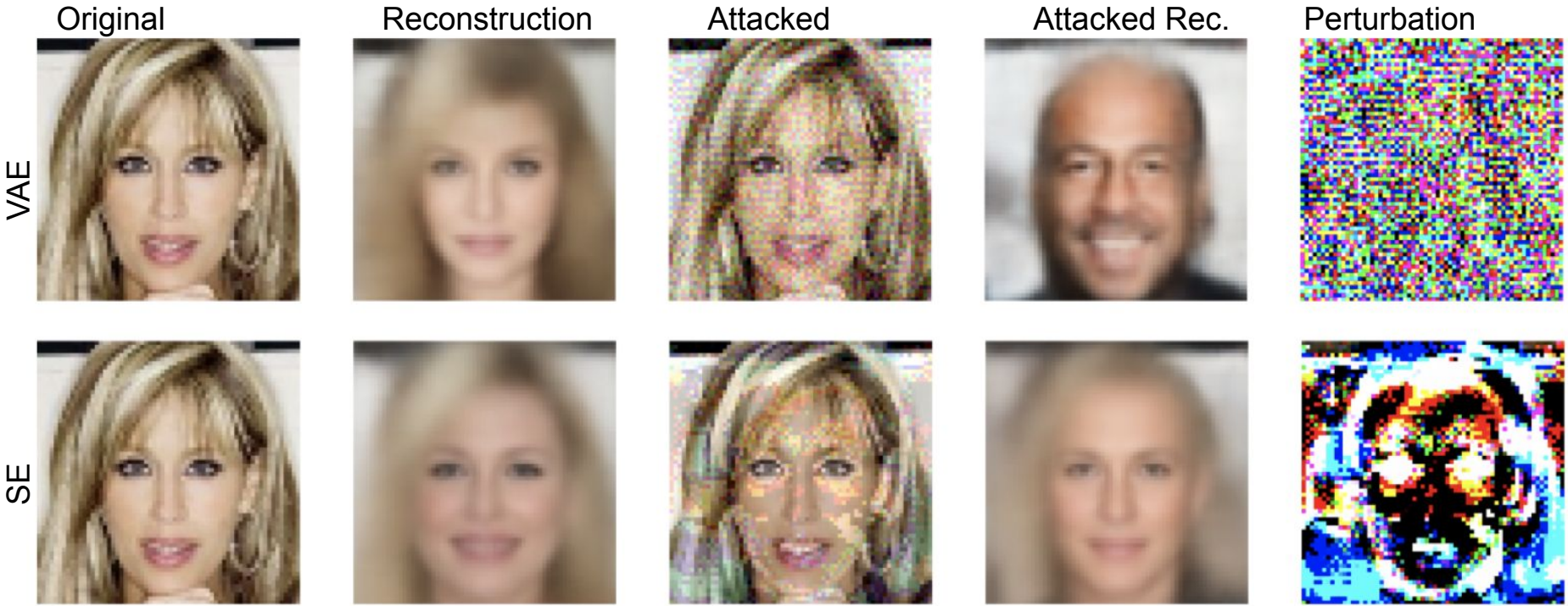
Sonuçlar (Bıyık tanıma)

Private & Confidential



Sonuçlar (Kellik tanıma)

Private & Confidential



Özet

Private & Confidential

- Genel Yapay Zeka (AGI – Artificial General Intelligence)
- Sağlamlük (Robustness)
  - Saldırı (attack) ve Korunma (defense)
  - Spesifikasyonlar ve garantiler (certificates for specifications)
  - Gösterimler (representations)
- Doğrulama (Verification)
- Gösterim öğrenmesi (VAE ve düzleştirlmiş otokodlayıcılar)

- Pekiştirmeli öğrenme:
  - Bertsekas, 2019, REINFORCEMENT LEARNING AND OPTIMAL CONTROL https://web.mit.edu/dimitrib/www/RLbook.html
  - Sutton and Barto, Reinforcement learning, 2nd edition
- Doğrulama
  - Gowal et. al. On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models https://arxiv.org/abs/1810.12715
- Gösterim Öğrenmesi
  - Tschannen et. al. Recent Advances in Autoencoder-Based Representation Learning https://arxiv.org/abs/1812.05069
  - Chen et. al. A Simple Framework for Contrastive Learning of Visual Representations https://arxiv.org/abs/2002.05709
  - Cemgil et. al. Adversarially Robust Representations with Smooth Encoders https://openreview.net/forum?id=H1gfFaEYDS