Learnable Sparsity and Weak Supervision for Data-efficient, Transparent, and Compact Neural Models

Gonçalo M. Correia

Jury: André Martins, Mário Figueiredo, Ivan Titov, Wilker Aziz, Isabel Trancoso

Deep learning successes

Deep learning successes

- Subset of machine learning that uses **neural networks**

Deep learning successes

- Subset of machine learning that uses **neural networks**
- Powerful tool for learning representations of any data

Deep learning successes

- Subset of machine learning that uses **neural networks**
- Powerful tool for learning representations of any data
- Remarkable results

Deep learning successes

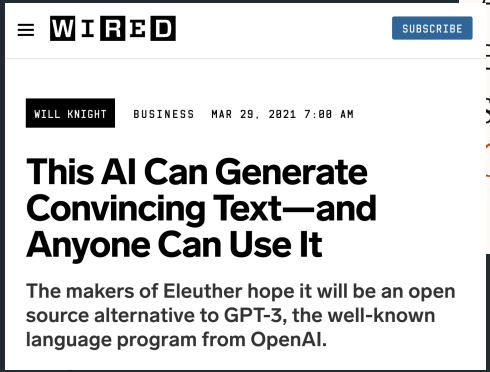
Deep learning successes

A robot wrote this entire article. Are you scared yet, human?

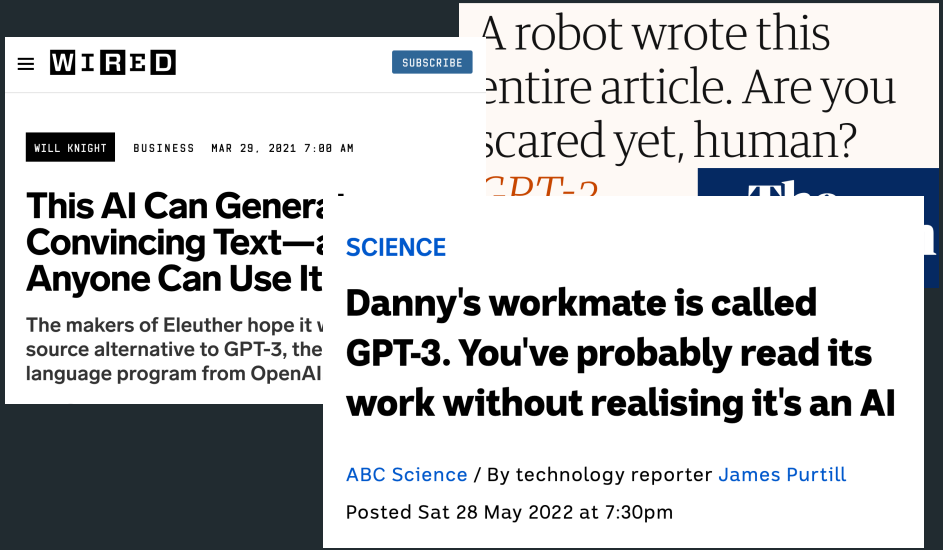
GPT-3

The Guardian
News website of the year

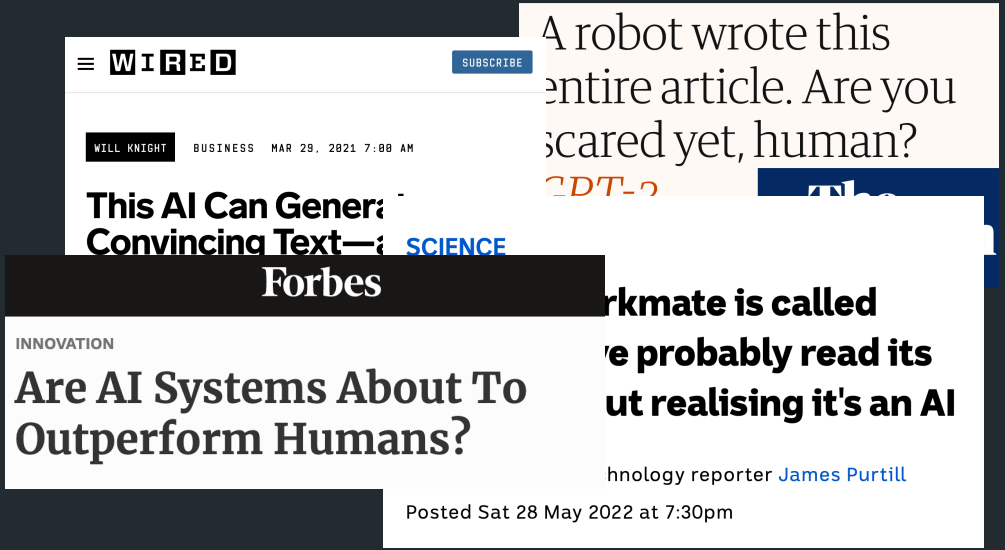
Deep learning successes



Deep learning successes



Deep learning successes



Deep learning successes

Artificial intelligence beats eight world champions at bridge

Victory marks milestone for AI as bridge requires more human skills than other strategy games

robot wrote this
ntire article. Are you
cared yet, human?
PT-2

INNOVATION

Are AI Systems About To Outperform Humans?

rkmate is called
e probably read its
ut realising it's an AI

hnology reporter James Purtill

Posted Sat 28 May 2022 at 7:30pm

Deep learning successes

Artificial intelligence beats eight world champions at bridge

Victory marks milestone for AI bridge requires more human sk than other strategy games

INNOVATION

Are AI Systems About T Outperform Humans?

Posted Sa

robot wrote this

AI 'outperforms' doctors diagnosing breast cancer

Fergus Walsh
Medical correspondent
@BBCFergusWalsh

Deep learning limitations and drawbacks

Deep learning limitations and drawbacks

- Requires a lot of data

Deep learning limitations and drawbacks

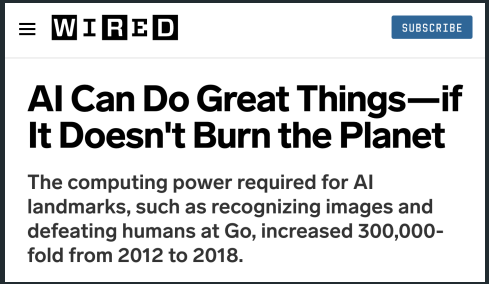
- Requires a lot of data
- Hard to understand and interpret reasons behind decisions

Deep learning limitations and drawbacks

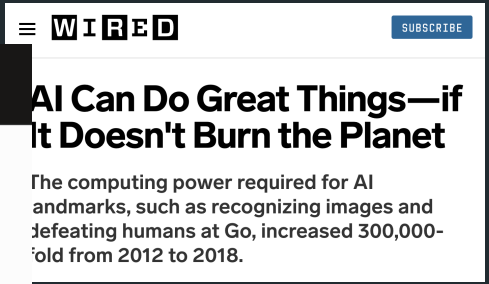
- Requires a lot of data
- Hard to understand and interpret reasons behind decisions
- Requires a lot of computation

Deep learning limitations and drawbacks

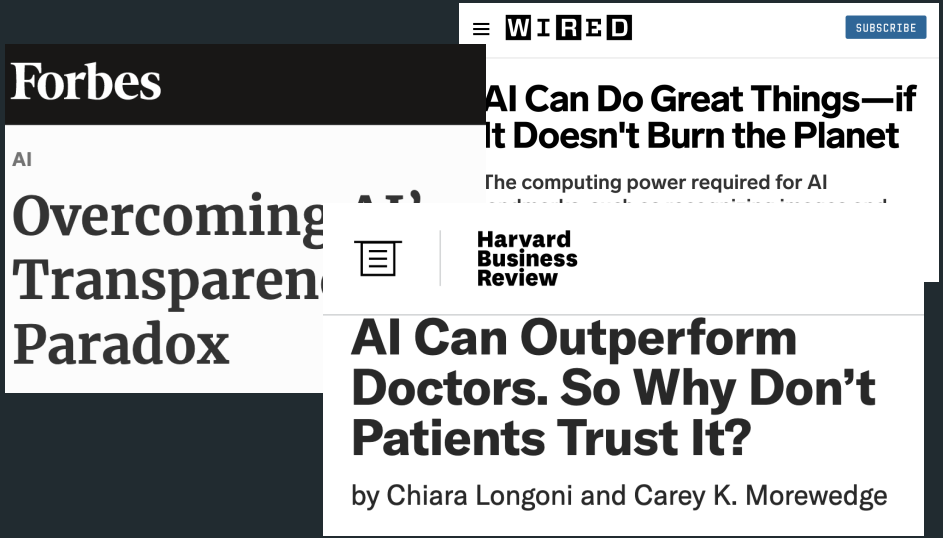
Deep learning limitations and drawbacks



Deep learning limitations and drawbacks



Deep learning limitations and drawbacks



Key concepts of this thesis

Learnable Sparsity and Weak Supervision for Data-efficient, Transparent, and Compact Neural Models

Key concepts of this thesis

Learnable Sparsity and Weak Supervision for Data-efficient, Transparent, and Compact Neural Models

Key concepts of this thesis

Learnable Sparsity and Weak Supervision for Data-efficient, Transparent, and Compact Neural Models

Key concepts of this thesis

Learnable Sparsity and Weak Supervision for Data-efficient, Transparent, and Compact Neural Models

Key concepts of this thesis

Learnable Sparsity and Weak Supervision for **Data-efficient**, **Transparent**, and **Compact** Neural Models

Key concepts of this thesis

Learnable Sparsity and **Weak Supervision**
for **Data-efficient**, **Transparent**, and **Compact**
Neural Models

Table of Contents

A Simple and Effective Approach to APE with Transfer Learning

Adaptively Sparse Transformers

Efficient Marg. of Discrete Latent Variables via Sparsity

Conclusions

(Vaswani et al., 2017)

A bit of context on transformers

What if... Attention is all you need?

(Vaswani et al., 2017)

A bit of context on transformers

What if... Attention is all you need?

Key idea: Let's mainly use attention mechanisms!

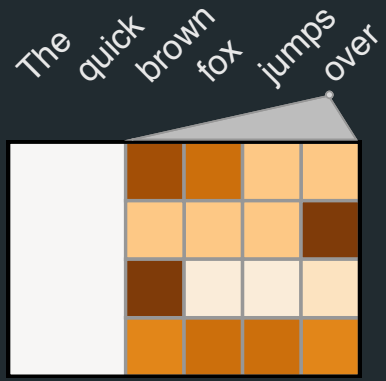
(Vaswani et al., 2017)

A bit of context on transformers

What if... Attention is all you need?

Key idea: Let's mainly use attention mechanisms!

- Do attention with multiple heads (i.e. attention mechanisms in parallel)



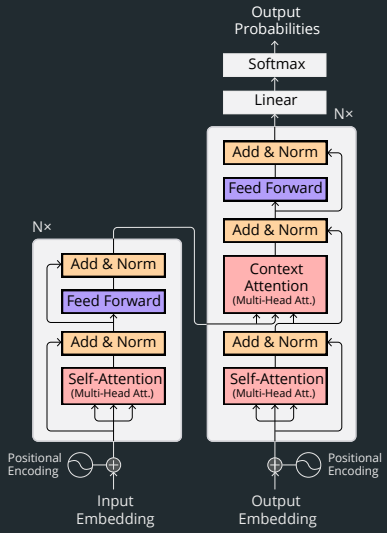
(Vaswani et al., 2017)

A bit of context on transformers

What if... Attention is all you need?

Key idea: Let's mainly use attention mechanisms!

- Do attention with multiple heads (i.e. attention mechanisms in parallel)
- ... and do it through several layers



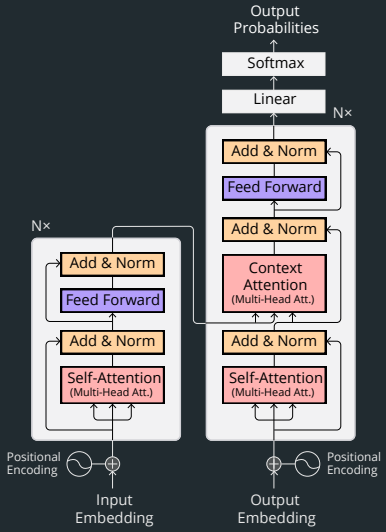
(Devlin et al., 2019)

A bit of context on transformers

What if... Attention is all you need?

Key idea: Let's mainly use attention mechanisms!

- Do attention with multiple heads (i.e. attention mechanisms in parallel)
- ... and do it through several layers
- Inspiration for big general-purpose models like BERT and GPT-3!



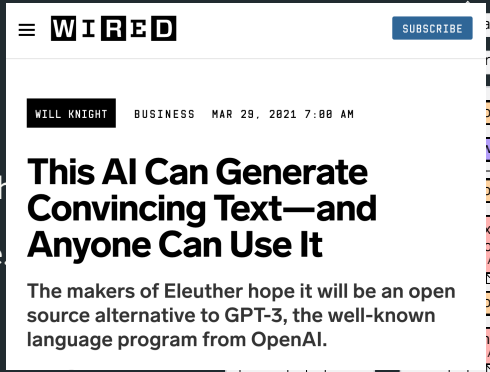
(Devlin et al., 2019)

A bit of context on transformers

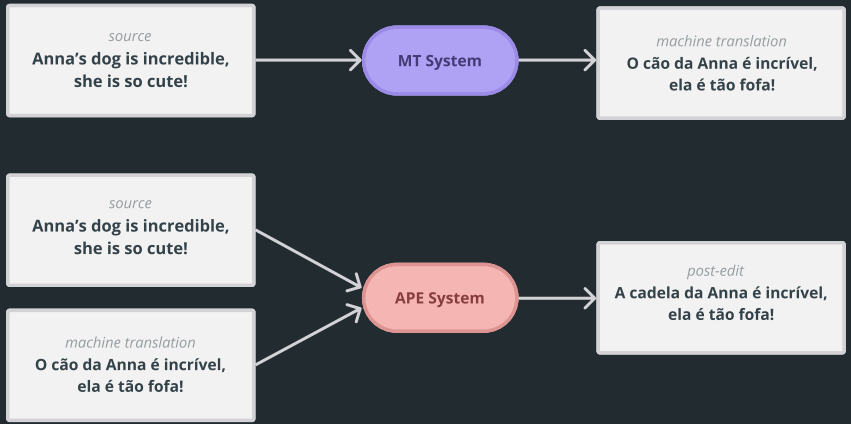
What if... Attention is all you need?

Key idea: Let's mainly use attention mech

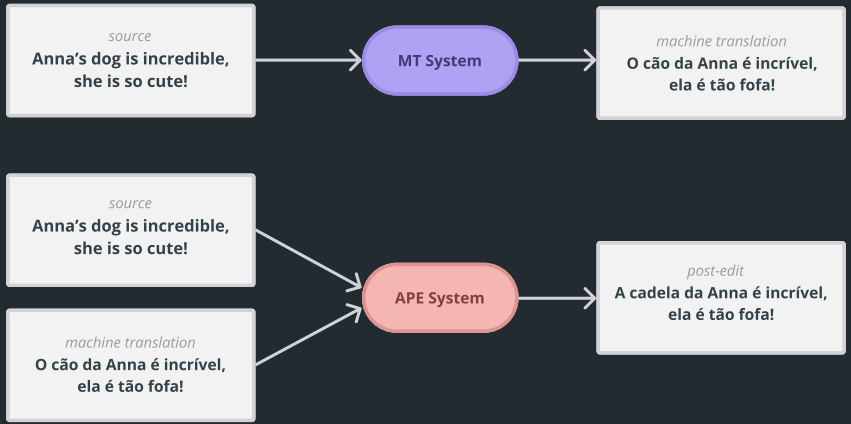
- Do attention with multiple heads (i.e. mechanisms in parallel)
- ... and do it through several layers
- Inspiration for big general-purpose models like BERT and GPT-3!



What is APE?

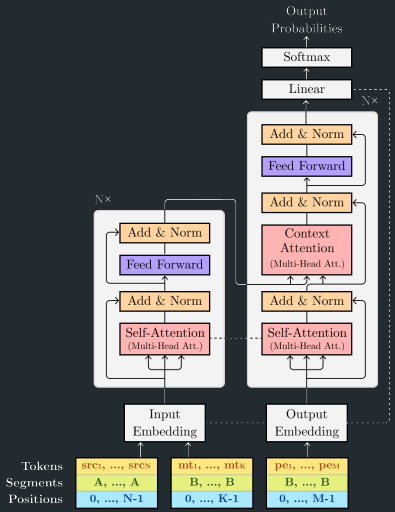

source
Anna's dog is incredible, she is so cute!
MT System
machine translation
O cão da Anna é incrível, ela é tão fofa!
source
Anna's dog is incredible, she is so cute!
machine translation
O cão da Anna é incrível, ela é tão fofa!
APE System
post-edit
A cadela da Anna é incrível, ela é tão fofa!

What is APE?



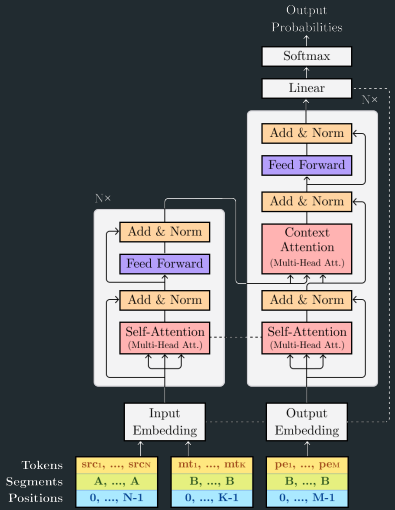
Challenge: APE data is very scarce! Need to create artificial data.

BERT for APE


Output Probabilities
Softmax
Linear
N×
Add & Norm
Feed Forward
Add & Norm
Context Attention
(Multi-Head Att.)
Add & Norm
Self-Attention
(Multi-Head Att.)
N×
Add & Norm
Feed Forward
Add & Norm
Self-Attention
(Multi-Head Att.)
Input Embedding
Output Embedding
Tokens
Segments
Positions
src1, ..., srcN
A, ..., A
0, ..., N-1
mt1, ..., mtK
B, ..., B
0, ..., K-1
pe1, ..., peM
B, ..., B
0, ..., M-1

BERT for APE

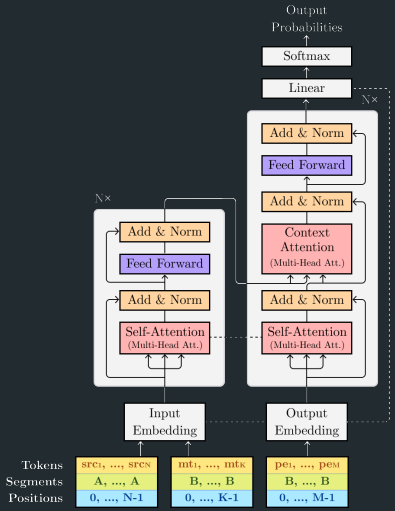
Key idea: Use BERT to do APE



BERT for APE

Key idea: Use BERT to do APE

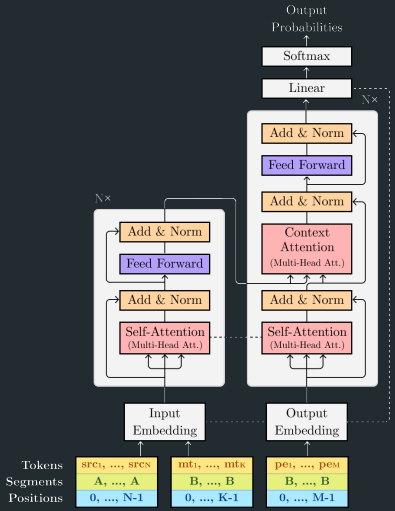
- Prior to this work, BERT was mainly used for simple classification tasks



BERT for APE

Key idea: Use BERT to do APE

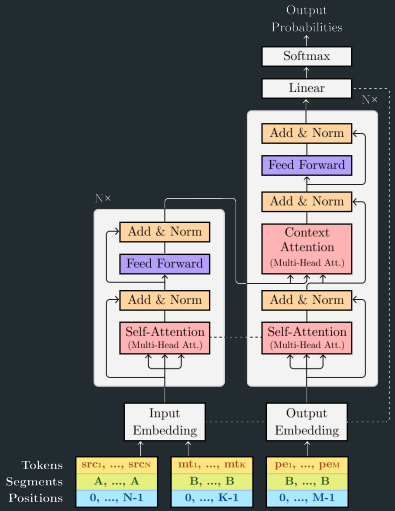
- Prior to this work, BERT was mainly used for simple classification tasks
- We introduced an effective method to use BERT in a generation task (APE)



BERT for APE

Key idea: Use BERT to do APE

- Prior to this work, BERT was mainly used for simple classification tasks
- We introduced an effective method to use BERT in a generation task (APE)
- Smart parameter sharing between encoder and decoder



Key results

model (data size)	TER↓	BLEU↑
mt baseline	24.48	62.49

(Junczys-Dowmunt and Grundkiewicz, 2018)

Key results

model (data size)	TER↓	BLEU↑
mt baseline	24.48	62.49
dual-source transformer (8M)	18.10	71.72

(Junczys-Dowmunt and Grundkiewicz, 2018)

Key results

model (data size)	TER↓	BLEU↑
mt baseline	24.48	62.49
dual-source transformer (8M)	18.10	71.72
dual-source transformer (23K)	27.73	59.78

Key results

model (data size)	TER↓	BLEU↑
mt baseline	24.48	62.49
dual-source transformer (8M)	18.10	71.72
dual-source transformer (23K)	27.73	59.78
ours (23K)	19.03	70.66

Key results

model (data size)	TER↓	BLEU↑
mt baseline	24.48	62.49
dual-source transformer (8M)	18.10	71.72
dual-source transformer (23K)	27.73	59.78
ours (23K)	19.03	70.66
ours (8M)	17.26	73.42

Key takeaways

Key takeaways

- One of pioneers in using pre-trained transformer encoders for a generation task

Key takeaways

- One of pioneers in using pre-trained transformer encoders for a generation task
- Massive improvement in low-resource scenario (data-efficiency)

(Lee et al., 2020)

Key takeaways

- One of pioneers in using pre-trained transformer encoders for a generation task
- Massive improvement in low-resource scenario (data-efficiency)
- Steered SOTA of APE towards using weak supervision through pre-trained models

Table of Contents

A Simple and Effective Approach to APE with Transfer Learning

Adaptively Sparse Transformers

Efficient Marg. of Discrete Latent Variables via Sparsity

Conclusions

Getting to know attention heads better

Attention heads may aid visualization but they are completely dense.

Getting to know attention heads better

Attention heads may aid visualization but they are completely dense.

Our solution is to bet on sparsity:

- for interpretability
- for discovering linguistic structure
- for efficiency

Getting to know attention heads better

Attention heads may aid visualization but they are completely dense.

Our solution is to bet on sparsity:

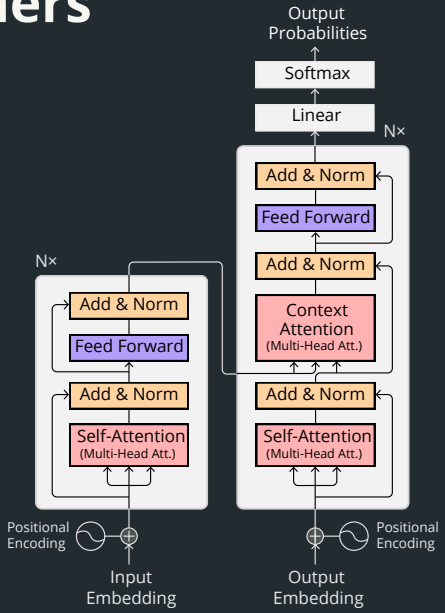
- for interpretability
- for discovering linguistic structure
- for efficiency

(Vaswani et al., 2017)

Transformers

In each attention head:

$$\bar{\mathbf{V}} = \mathbf{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}.$$



(Vaswani et al., 2017)

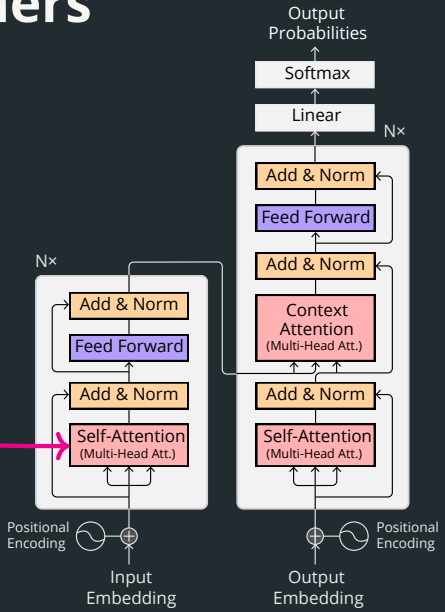
Transformers

In each attention head:

$$\bar{\mathbf{V}} = \mathbf{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}.$$

Attention in three places:

- Self-attention in the encoder



(Vaswani et al., 2017)

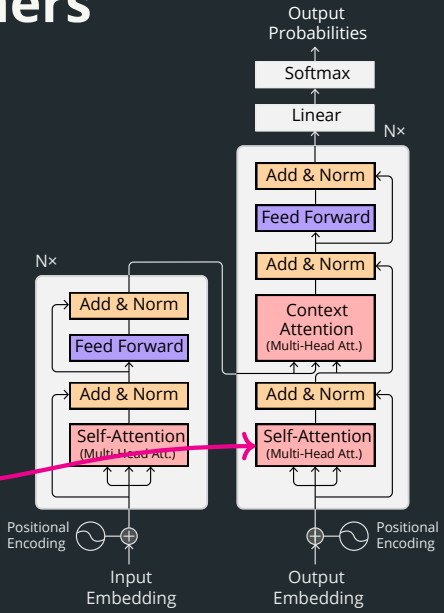
Transformers

In each attention head:

$$\bar{\mathbf{V}} = \mathbf{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}.$$

Attention in three places:

- Self-attention in the encoder
- Self-attention in the decoder



(Vaswani et al., 2017)

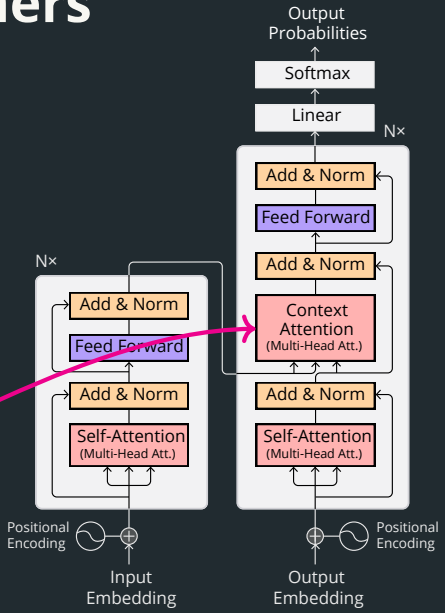
Transformers

In each attention head:

$$\bar{\mathbf{V}} = \mathbf{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d_k}}\right)\mathbf{V}.$$

Attention in three places:

- Self-attention in the encoder
- Self-attention in the decoder
- Contextual attention



Sparse Transformers

Sparse Transformers

Key idea: replace softmax in attention heads by a sparse normalizing function! 🙌

Adaptively Sparse Transformers

Key idea: replace softmax in attention heads by a sparse normalizing function! 🙌

Another key idea: use a normalizing function that is adaptively sparse via a learnable α! 🙌🙌🙌

What is softmax?

Softmax exponentiates and normalizes:

$$[\mathbf{softmax}(z)]_i := \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

What is softmax?

Softmax exponentiates and normalizes:

$$[\mathbf{softmax}(z)]_i := \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

It's fully dense: $\mathbf{softmax}(z) > 0$

(Peters et al., 2019)

α-entmax

Parametrized by $\alpha \geq 0$:

(Peters et al., 2019)

α-entmax

Parametrized by $\alpha \geq 0$:

- **Argmax** corresponds to $\alpha \rightarrow \infty$

(Peters et al., 2019)

α-entmax

Parametrized by $\alpha \geq 0$:

- **Argmax** corresponds to $\alpha \to \infty$
- **Softmax** amounts to $\alpha \to 1$

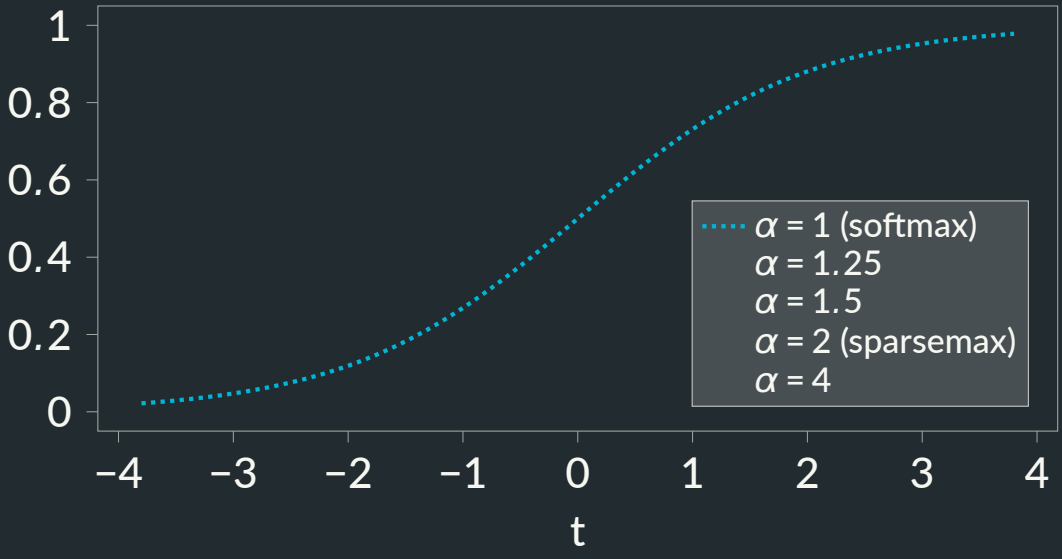
(Peters et al., 2019)

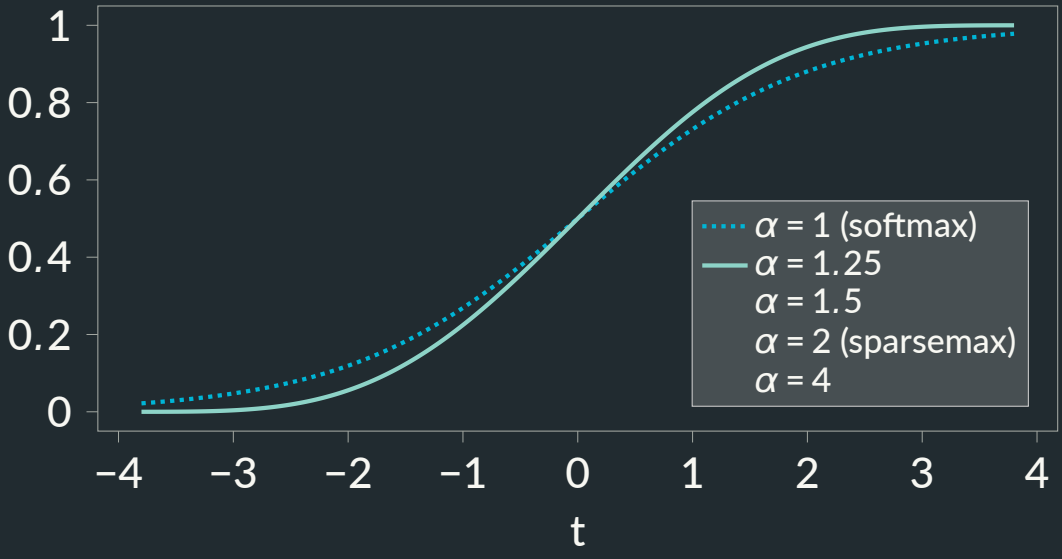
α-entmax

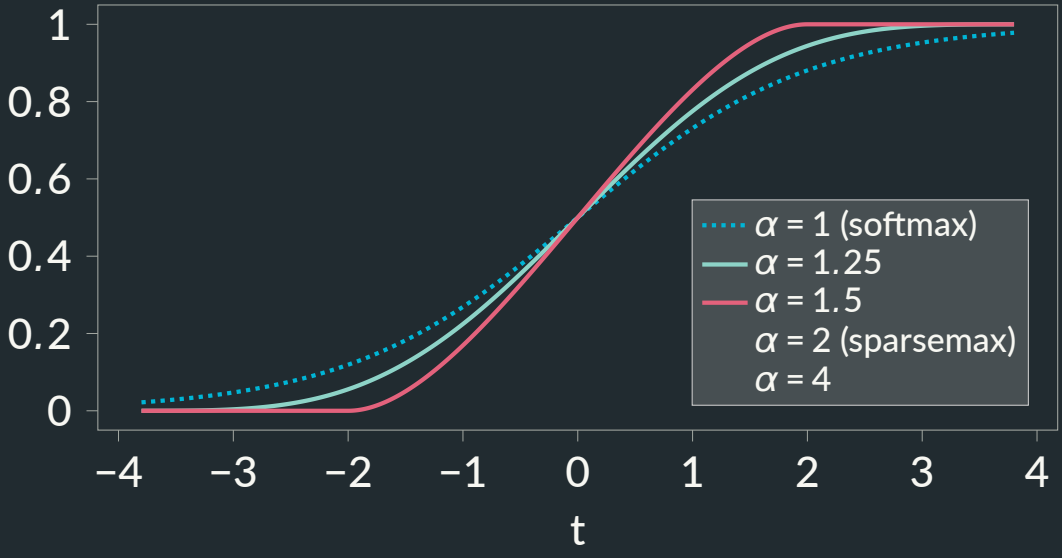
Parametrized by $\alpha \geq 0$:

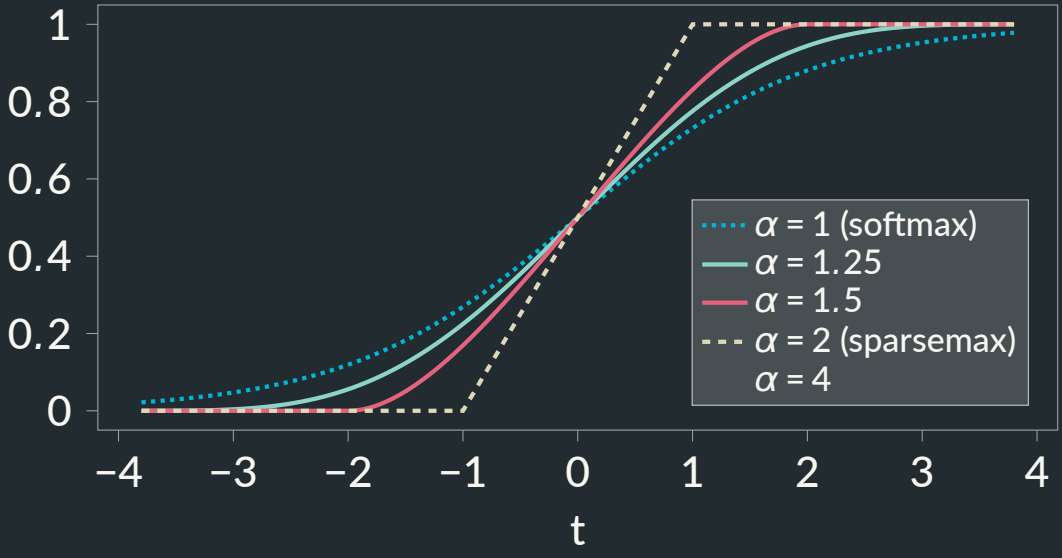
- **Argmax** corresponds to $\alpha \to \infty$
- **Softmax** amounts to $\alpha \to 1$

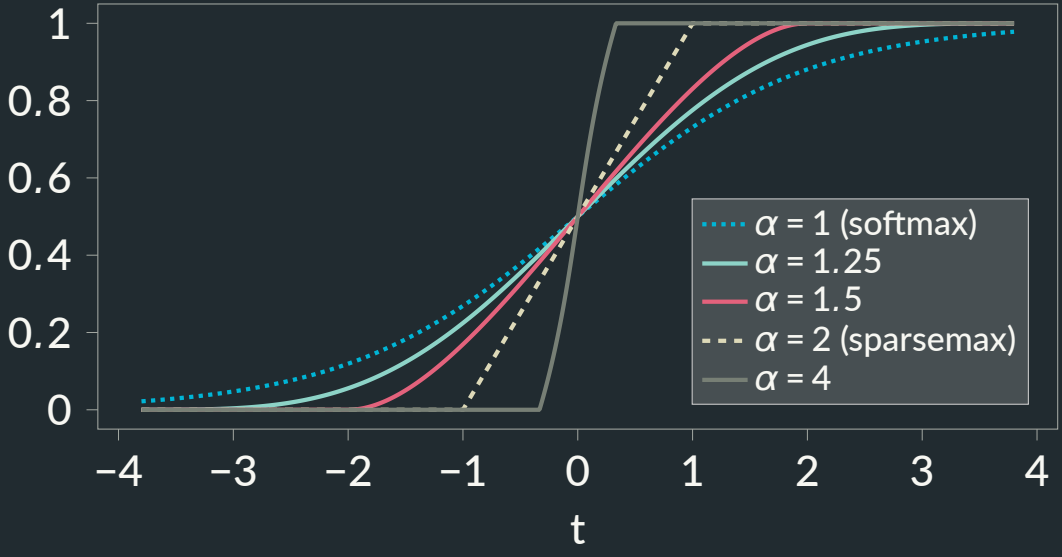
Key result: **can be sparse for** $\alpha > 1$, propensity for sparsity increases with α.


1
0.8
0.6
0.4
0.2
0
−4
−3
−2
−1
0
1
2
3
4
t
α = 1 (softmax)
α = 1.25
α = 1.5
α = 2 (sparsemax)
α = 4


1
0.8
0.6
0.4
0.2
0
−4
−3
−2
−1
0
1
2
3
4
t
α = 1 (softmax)
α = 1.25
α = 1.5
α = 2 (sparsemax)
α = 4


1
0.8
0.6
0.4
0.2
0
−4
−3
−2
−1
0
1
2
3
4
t
α = 1 (softmax)
α = 1.25
α = 1.5
α = 2 (sparsemax)
α = 4


1
0.8
0.6
0.4
0.2
0
−4
−3
−2
−1
0
1
2
3
4
t
α = 1 (softmax)
α = 1.25
α = 1.5
α = 2 (sparsemax)
α = 4


1
0.8
0.6
0.4
0.2
0
−4
−3
−2
−1
0
1
2
3
4
t
α = 1 (softmax)
α = 1.25
α = 1.5
α = 2 (sparsemax)
α = 4

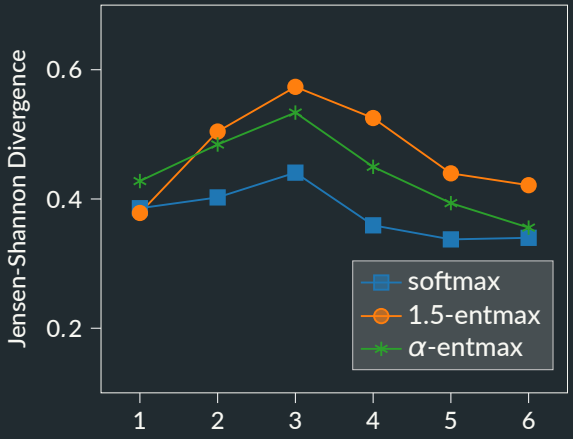
Learning α

Learning α

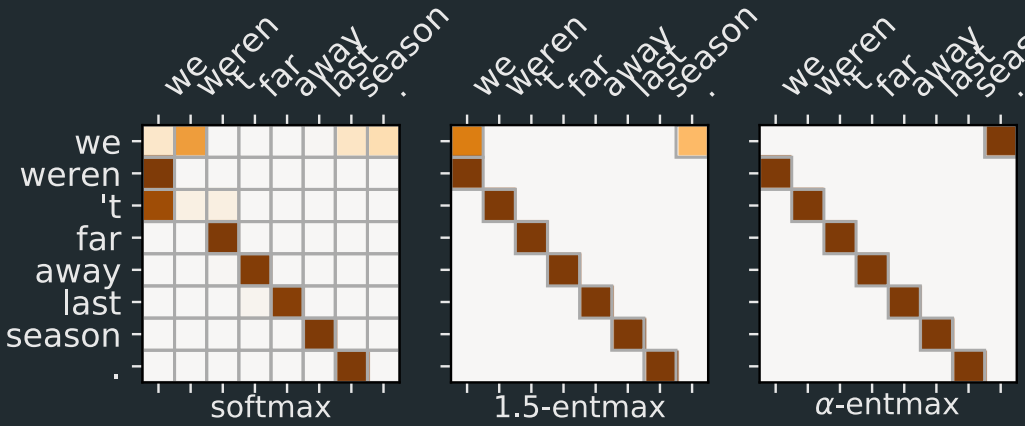
Key contribution:

a closed-form expression for $\frac{\partial\, \alpha\text{-entmax}(\boldsymbol{z})}{\partial \alpha}$ 😱

Head diversity per layer

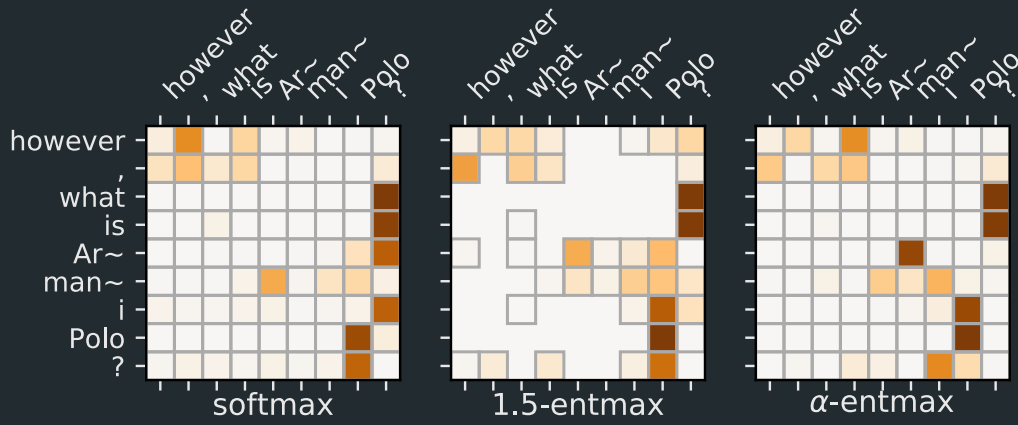


Previous position head



This head role was also found in Voita et al. (2019)! Learned $\alpha = 1.91$.

Interrogation-detecting head



Learned $\alpha = 1.05$.

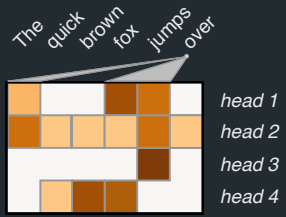
Key takeaways

Introduce adaptive sparsity
for Transformers via α-entmax with a gradient learnable α,
improving transparency.

Key takeaways

Introduce adaptive sparsity
for Transformers via α-entmax with a gradient learnable α,
improving transparency.

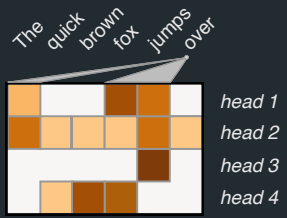
adaptive sparsity



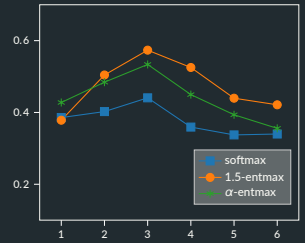
Key takeaways

Introduce adaptive sparsity
for Transformers via α-entmax with a gradient learnable α,
improving transparency.

adaptive sparsity



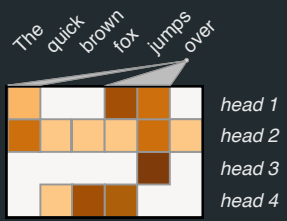
reduced head redundancy



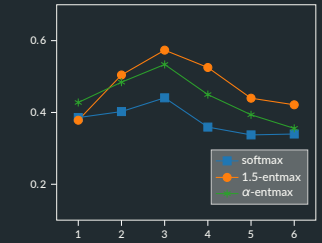
Key takeaways

Introduce adaptive sparsity
for Transformers via α-entmax with a gradient learnable α,
improving transparency.

adaptive sparsity



reduced head redundancy



clearer head roles

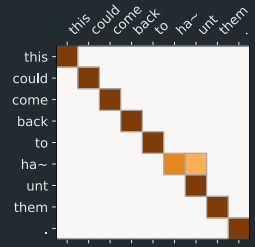


Table of Contents

A Simple and Effective Approach to APE with Transfer Learning

Adaptively Sparse Transformers

Efficient Marg. of Discrete Latent Variables via Sparsity

Conclusions

Latent variable models

Latent variable z can be

Latent variable models

Latent variable z can be continuous

Source: Bouges et al., 2013

Latent variable models

Latent variable z can be continuous, discrete

Source: valleyeyecareaz.com

Latent variable models

Latent variable z can be continuous, discrete, or structured

Source: Liu et al., 2015

Training discrete or structured latent variable models

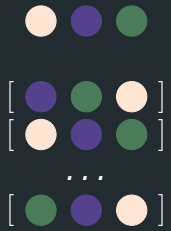
Latent variable z can be

Training discrete or structured latent variable models

Latent variable z can be discrete

Training discrete or structured latent variable models

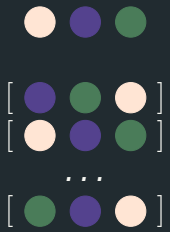
Latent variable z can be discrete or structured



Training discrete or structured latent variable models

Latent variable z can be discrete or structured

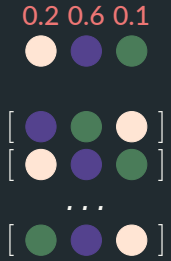
$\pi(z|x, \theta)$: distribution over possible z



Training discrete or structured latent variable models

Latent variable z can be discrete or structured

$\pi(z|x, \theta)$: distribution over possible z



Training discrete or structured latent variable models

Latent variable z can be discrete or structured

$\pi(z|x, \theta)$: distribution over possible z

Training discrete or structured latent variable models

Latent variable z can be discrete or structured

$\pi(z|x, \theta)$: distribution over possible z

$\ell(x, z; \theta)$: downstream loss: ELBO, Log-Likelihood, (...)

Training discrete or structured latent variable models

Latent variable z can be discrete or structured

$\pi(z|x, \theta)$: distribution over possible z

$\ell(x, z; \theta)$: downstream loss: ELBO, Log-Likelihood, (...)

To train, we need to compute the following expectation:

Training discrete or structured latent variable models

Latent variable z can be discrete or structured

$\pi(z|x, \theta)$: distribution over possible z

$\ell(x, z; \theta)$: downstream loss: ELBO, Log-Likelihood, (...)

To train, we need to compute the following expectation:

$$\mathcal{L}_x(\theta) = \sum_{z \in \mathcal{Z}} \pi(z|x, \theta)\, \ell(x, z; \theta)$$

Training discrete or structured latent variable models

Latent variable z can be discrete or structured

$\pi(z|x, \theta)$: distribution over possible z

$\ell(x, z; \theta)$: downstream loss: ELBO, Log-Likelihood, (...)

To train, we need to compute the following expectation:

$$\mathcal{L}_x(\theta) = \sum_{z \in \mathcal{Z}} \pi(z|x, \theta)\, \ell(x, z; \theta)$$

If $\mathcal{Z}$ is large, this sum can get very expensive due to $\ell(x, z; \theta)$! 😱

Training discrete or structured latent variable models

Latent variable z can be discrete or structured

$\pi(z|x, \theta)$: distribution over possible z

$\ell(x, z; \theta)$: downstream loss: ELBO, Log-Likelihood, (...)

To train, we need to compute the following expectation:

$$\mathcal{L}_x(\theta) = \sum_{z \in \mathcal{Z}} \pi(z|x, \theta)\, \ell(x, z; \theta)$$

If $\mathcal{Z}$ is combinatorial, this can be intractable to compute! 😱 😱 😱

Current solutions

Using emergent communication as example

Method	success (%)	# messages
Monte Carlo		
Marginalization		

Current solutions

Using emergent communication as example

Method	success (%)	# messages
Monte Carlo		
Marginalization		
Dense	93.37 ±0.42	256

Current solutions

Using emergent communication as example

Method	success (%)	# messages
Monte Carlo		
SFE	33.05 ±2.84	1
Marginalization		
Dense	93.37 ±0.42	256

Current solutions

Using emergent communication as example

Method	success (%)	# messages
Monte Carlo		
SFE	33.05 ±2.84	1
SFE+	44.32 ±2.72	2
Marginalization		
Dense	93.37 ±0.42	256

Current solutions

Using emergent communication as example

Method	success (%)	# messages
Monte Carlo		
SFE	33.05 ±2.84	1
SFE+	44.32 ±2.72	2
Gumbel	23.51 ±16.19	1
Marginalization		
Dense	93.37 ±0.42	256

Our solution

Using emergent communication as example

Method	success (%)	# messages
Monte Carlo		
SFE	33.05 ±2.84	1
SFE+	44.32 ±2.72	2
Gumbel	23.51 ±16.19	1
Marginalization		
Dense	93.37 ±0.42	256
Sparse		

Our solution

Using emergent communication as example

Method	success (%)	# messages
Monte Carlo		
SFE	33.05 ±2.84	1
SFE+	44.32 ±2.72	2
Gumbel	23.51 ±16.19	1
Marginalization		
Dense	93.37 ±0.42	256
Sparse		

Our solution

Using emergent communication as example

Method	success (%)	# messages
Monte Carlo		
SFE	33.05 ±2.84	1
SFE+	44.32 ±2.72	2
Gumbel	23.51 ±16.19	1
Marginalization		
Dense	93.37 ±0.42	256
Sparse	93.35 ±0.50	

Our solution

Using emergent communication as example

Method	success (%)	# messages
Monte Carlo		
SFE	33.05 ±2.84	1
SFE+	44.32 ±2.72	2
Gumbel	23.51 ±16.19	1
Marginalization		
Dense	93.37 ±0.42	256
Sparse	93.35 ±0.50	3.13 ±0.48

(Martins and Astudillo, 2016; Niculae, Martins, et al., 2018)

Our solution

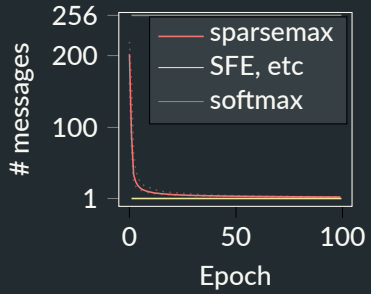
Using emergent communication as example

Method	success (%)	# messages
Monte Carlo		
SFE	33.05 ±2.84	1
SFE+	44.32 ±2.72	2
Gumbel	23.51 ±16.19	1
Marginalization		
Dense	93.37 ±0.42	256
Sparse	93.35 ±0.50	3.13 ±0.48

We use sparsemax, top-k sparsemax and SparseMAP to allow efficient marginalization

Our solution

Using emergent communication as example



We use sparsemax, top-k sparsemax and SparseMAP to allow efficient marginalization

Results

We test our methods for models with discrete latent variables,

Results

We test our methods for models with discrete latent variables,

- Semi-Supervised VAE

Results

We test our methods for models with discrete latent variables,

- Semi-Supervised VAE
- Emergent communication

Results

We test our methods for models with discrete latent variables,

- Semi-Supervised VAE
- Emergent communication

but also in models with an exponentially large set of $\mathcal{Z}$,

Results

We test our methods for models with discrete latent variables,

- Semi-Supervised VAE
- Emergent communication

but also in models with an exponentially large set of $\mathcal{Z}$,

- Bit-vector VAE

Results

We test our methods for models with discrete latent variables,

- Semi-Supervised VAE
- Emergent communication

but also in models with an exponentially large set of $\mathcal{Z}$,

- Bit-vector VAE

Our methods are top-performers and efficient!

Key takeaways

We introduce a new method
to train compact latent variable models,
using learned sparsity.

Key takeaways

We introduce a new method
to train compact latent variable models,
using learned sparsity.

discrete and structured

Key takeaways

We introduce a new method
to train compact latent variable models,
using learned sparsity.

discrete and structured

deterministic, yet efficient

Key takeaways

We introduce a new method
to train compact latent variable models,
using learned sparsity.

discrete and structured

deterministic, yet efficient

sparse, as needed

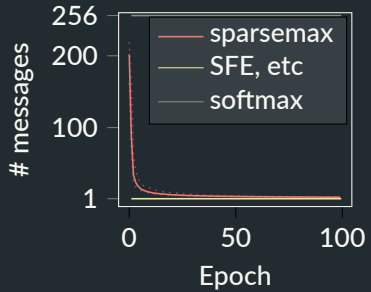


Table of Contents

A Simple and Effective Approach to APE with Transfer Learning

Adaptively Sparse Transformers

Efficient Marg. of Discrete Latent Variables via Sparsity

Conclusions

Conclusions

Using learned sparsity and weak supervision
we took steps to take neural models closer to version 2.0

Conclusions

Using learned sparsity and weak supervision
we took steps to take neural models closer to version 2.0

data-efficiency

model (data size)	BLEU↑
dual-source transformer (8M)	71.72
dual-source transformer (23K)	59.78
ours (23K)	70.66

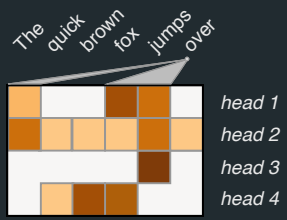
Conclusions

Using learned sparsity and weak supervision
we took steps to take neural models closer to version 2.0

data-efficiency

model (data size)	BLEU↑
dual-source transformer (8M)	71.72
dual-source transformer (23K)	59.78
ours (23K)	70.66

transparency



Conclusions

Using learned sparsity and weak supervision
we took steps to take neural models closer to version 2.0

data-efficiency

model (data size)	BLEU↑
dual-source transformer (8M)	71.72
dual-source transformer (23K)	59.78
ours (23K)	70.66

transparency

better & efficient compactness

References I

- Bouges, Pierre, Thierry Chateau, Christophe Blanc, and Gaëlle Loosli (Dec. 2013). “Handling missing weak classifiers in boosted cascade: application to multiview and occluded face detection”. In: *EURASIP Journal on Image and Video Processing* 2013, p. 55. DOI: 10.1186/1687-5281-2013-55.
- Correia, Gonçalo M., Vlad Niculae, Wilker Aziz, and André F. T. Martins (2020). “Efficient Marginalization of Discrete and Structured Latent Variables via Sparsity”. In: *Proc. NeurIPS*. URL: https://arxiv.org/abs/2007.01919.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *Proc. NAACL-HLT*.
- Junczys-Dowmunt, Marcin and Roman Grundkiewicz (2018). “MS-UEdin Submission to the WMT2018 APE Shared Task: Dual-Source Transformer for Automatic Post-Editing”. In: *Proceedings of WMT18*.
- Lazaridou, Angeliki, Alexander Peysakhovich, and Marco Baroni (2017). “Multi-agent cooperation and the emergence of (natural) language”. In: *Proc. ICLR*.
- Lee, Jihyung, WonKee Lee, Jaehun Shin, Baikjin Jung, Young-Kil Kim, and Jong-Hyeok Lee (2020). “POSTECH-ETRI’s Submission to the WMT2020 APE Shared Task: Automatic Post-Editing with Cross-lingual Language Model”. In: *Proceedings of WMT*.
- Liu, Ziwei, Ping Luo, Xiaogang Wang, and Xiaoou Tang (Dec. 2015). “Deep Learning Face Attributes in the Wild”. In: *Proceedings of International Conference on Computer Vision (ICCV)*.

References II

- Martins, André FT and Ramón Fernandez Astudillo (2016). “From softmax to sparsemax: A sparse model of attention and multi-label classification”. In: *Proc. of ICML*.
- Niculae, Vlad and Mathieu Blondel (2017). “A Regularized Framework for Sparse and Structured Neural Attention”. In: *arXiv preprint arXiv:1705.07704*.
- Niculae, Vlad, André FT Martins, Mathieu Blondel, and Claire Cardie (2018). “SparseMAP: Differentiable sparse structured inference”. In: *Proc. of ICML*.
- Peters, Ben, Vlad Niculae, and André F. T. Martins (2019). “Sparse Sequence-to-Sequence Models”. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). “Attention Is All You Need”. In: *Proc. of NeurIPS*.
- Voita, Elena, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov (2019). “Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned”. In: *Proc. ACL*.

Parameter sharing analysis

	TER↓	BLEU↑
MT Baseline	24.76	62.11
Transformer	27.80	60.76
Transformer decoder	20.33	69.31
Pre-trained BERT	20.83	69.11
with CA ← SA	18.91	71.81
and SA ↔ Encoder SA	**18.44**	**72.25**
and CA ↔ SA	18.75	71.83
and FF ↔ Encoder FF	19.04	71.53

(Niculae and Blondel, 2017)

Ω-Regularized Argmax

For convex Ω, define the **Ω-regularized argmax transformation**:

$$\textbf{argmax}_{\Omega}(\boldsymbol{z}) := \operatorname*{arg\,max}_{\boldsymbol{p} \in \triangle} \boldsymbol{z}^\top \boldsymbol{p} - \Omega(\boldsymbol{p})$$

(Niculae and Blondel, 2017)

Ω-Regularized Argmax

For convex Ω, define the **Ω-regularized argmax transformation**:

$$\textbf{argmax}_{\Omega}(\boldsymbol{z}) := \operatorname*{arg\,max}_{\boldsymbol{p} \in \triangle} \boldsymbol{z}^\top \boldsymbol{p} - \Omega(\boldsymbol{p})$$

- Argmax corresponds to **no regularization**, $\Omega \equiv 0$

(Niculae and Blondel, 2017)

Ω-Regularized Argmax

For convex Ω, define the **Ω-regularized argmax transformation**:

$$\textbf{argmax}_{\Omega}(\boldsymbol{z}) := \operatorname*{arg\,max}_{\boldsymbol{p} \in \triangle} \boldsymbol{z}^\top \boldsymbol{p} - \Omega(\boldsymbol{p})$$

- Argmax corresponds to **no regularization**, $\Omega \equiv 0$
- Softmax amounts to **entropic regularization**, $\Omega(\boldsymbol{p}) = \sum_{i=1}^{K} p_i \log p_i$

(Niculae and Blondel, 2017)

Ω-Regularized Argmax

For convex Ω, define the **Ω-regularized argmax transformation**:

$$\textbf{argmax}_{\Omega}(\boldsymbol{z}) := \operatorname*{arg\,max}_{\boldsymbol{p} \in \triangle} \boldsymbol{z}^\top \boldsymbol{p} - \Omega(\boldsymbol{p})$$

- Argmax corresponds to **no regularization**, $\Omega \equiv 0$
- Softmax amounts to **entropic regularization**, $\Omega(\boldsymbol{p}) = \sum_{i=1}^{K} p_i \log p_i$
- Sparsemax amounts to **ℓ_2-regularization**, $\Omega(\boldsymbol{p}) = \frac{1}{2}\|\boldsymbol{p}\|^2$.

(Martins and Astudillo, 2016)

(Niculae and Blondel, 2017)

Ω-Regularized Argmax

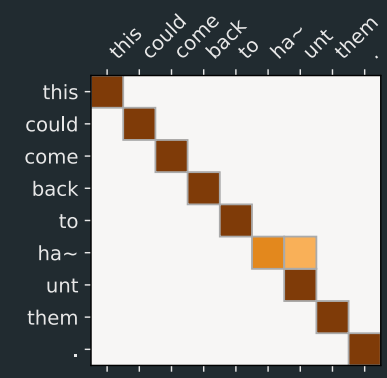
For convex Ω, define the **Ω-regularized argmax transformation**:

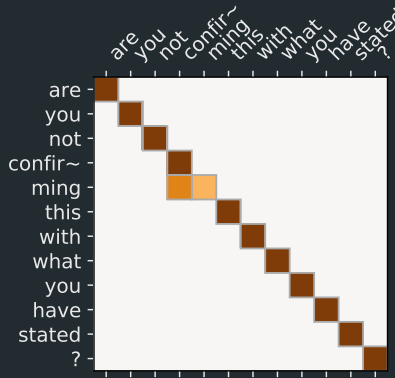
$$\textbf{argmax}_{\Omega}(\boldsymbol{z}) := \operatorname*{arg\,max}_{\boldsymbol{p} \in \triangle} \boldsymbol{z}^\top \boldsymbol{p} - \Omega(\boldsymbol{p})$$

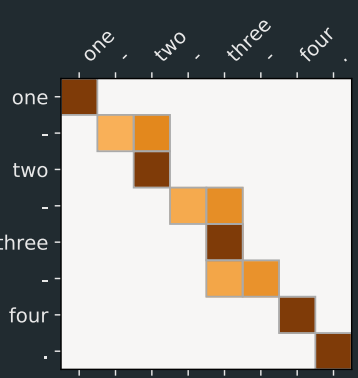
- Argmax corresponds to **no regularization**, $\Omega \equiv 0$
- Softmax amounts to **entropic regularization**, $\Omega(\boldsymbol{p}) = \sum_{i=1}^{K} p_i \log p_i$
- Sparsemax amounts to **ℓ_2-regularization**, $\Omega(\boldsymbol{p}) = \frac{1}{2}\|\boldsymbol{p}\|^2$.

Is there something in-between?

Subword-Merging Head







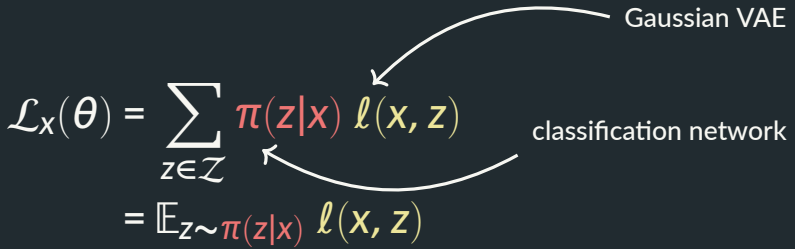
Learned $\alpha = 1.91$.

Semi-Supervised VAE

$$\mathcal{L}_x(\theta) = \sum_{z \in \mathcal{Z}} \pi(z|x)\, \ell(x, z)$$
$$= \mathbb{E}_{z \sim \pi(z|x)}\, \ell(x, z)$$

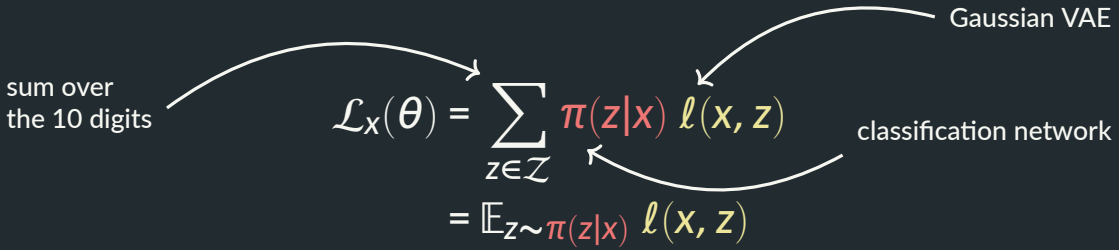
- Semi-Supervised VAE on MNIST: z is one of 10 categories

Semi-Supervised VAE



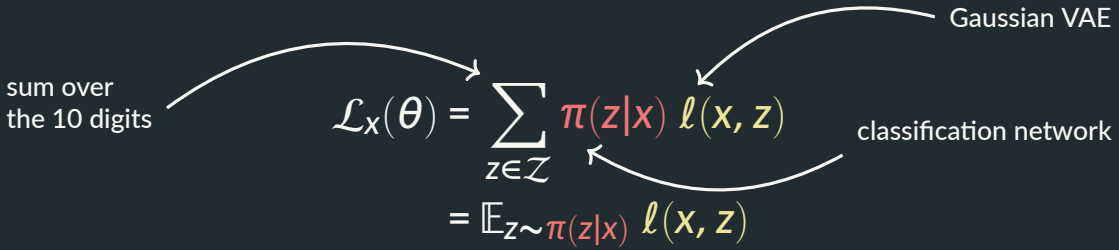
- Semi-Supervised VAE on MNIST: z is one of 10 categories

Semi-Supervised VAE



- Semi-Supervised VAE on MNIST: z is one of 10 categories

Semi-Supervised VAE



- Semi-Supervised VAE on MNIST: z is one of 10 categories
- Train this with 10% labeled data

Semi-Supervised VAE

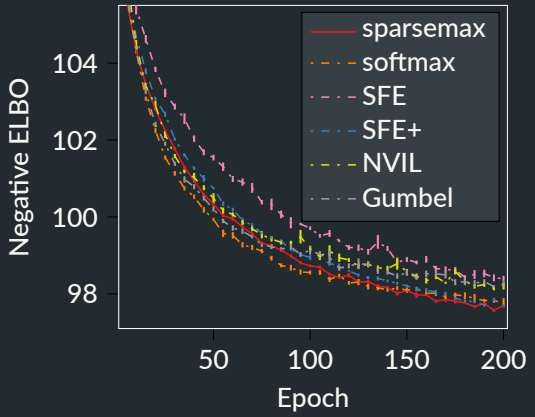
Method	Accuracy (%)	Dec. calls
Monte Carlo		
SFE	94.75±.002	1
SFE+	96.53±.001	2
NVIL	96.01±.002	1
Gumbel	95.46±.001	1
Marginalization		
Dense	96.93±.001	10

Semi-Supervised VAE

Method	Accuracy (%)	Dec. calls
Monte Carlo		
SFE	94.75±.002	1
SFE+	96.53±.001	2
NVIL	96.01±.002	1
Gumbel	95.46±.001	1
Marginalization		
Dense	96.93±.001	10
Sparse	96.87±.001	1.01±0.01

Semi-Supervised VAE

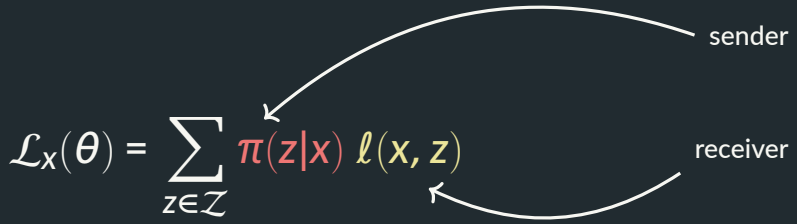
Method	Accuracy (%)	Dec. calls
Monte Carlo		
SFE	94.75±.002	1
SFE+	96.53±.001	2
NVIL	96.01±.002	1
Gumbel	95.46±.001	1
Marginalization		
Dense	96.93±.001	10
Sparse	96.87±.001	1.01±0.01



Emergent communication

$$\mathcal{L}_x(\theta) = \sum_{z \in \mathcal{Z}} \pi(z|x)\, \ell(x, z)$$
$$= \mathbb{E}_{z \sim \pi(z|x)}\, \ell(x, z)$$

Emergent communication

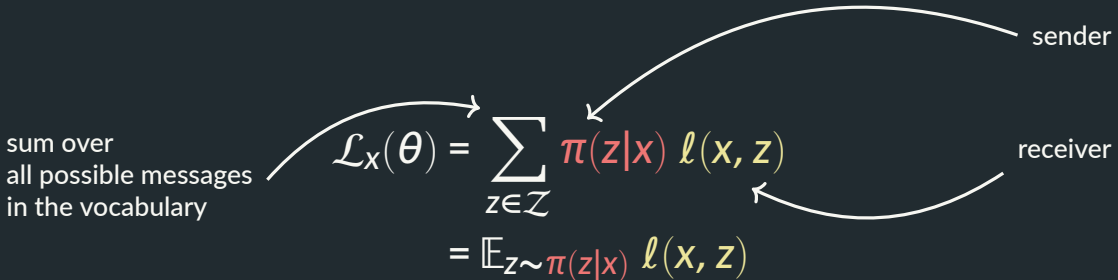


$$\mathcal{L}_x(\theta) = \sum_{z \in \mathcal{Z}} \pi(z|x)\, \ell(x, z)$$

$$= \mathbb{E}_{z \sim \pi(z|x)}\, \ell(x, z)$$

- receiver picks image from a set $\mathcal{V}$ based on message

Emergent communication



- receiver picks image from a set $\mathcal{V}$ based on message
- images come from ImageNet

(Lazaridou et al., 2017; Correia et al., 2020)

Emergent Communication

... but make it harder: $|\mathcal{Z}| = 256$, $|\mathcal{V}| = 16$

Method	success (%)	Dec. calls
Monte Carlo		
SFE	33.05 ±2.84	1
SFE+	44.32 ±2.72	2
NVIL	37.04 ±1.61	1
Gumbel	23.51 ±16.19	1
ST Gumbel	27.42 ±13.36	1
Marginalization		

(Lazaridou et al., 2017; Correia et al., 2020)

Emergent Communication

... but make it harder: $|\mathcal{Z}| = 256$, $|\mathcal{V}| = 16$

Method	success (%)	Dec. calls
Monte Carlo		
SFE	33.05 ±2.84	1
SFE+	44.32 ±2.72	2
NVIL	37.04 ±1.61	1
Gumbel	23.51 ±16.19	1
ST Gumbel	27.42 ±13.36	1
Marginalization		
Dense	93.37 ±0.42	256

(Lazaridou et al., 2017; Correia et al., 2020)

Emergent Communication

... but make it harder: $|\mathcal{Z}| = 256$, $|\mathcal{V}| = 16$

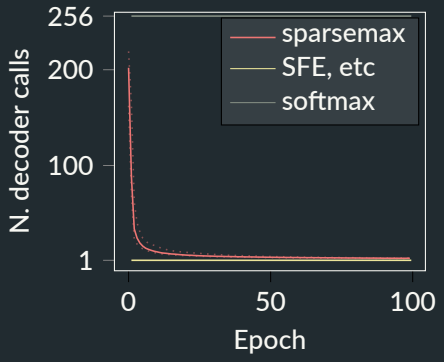
Method	success (%)	Dec. calls
Monte Carlo		
SFE	33.05 ±2.84	1
SFE+	44.32 ±2.72	2
NVIL	37.04 ±1.61	1
Gumbel	23.51 ±16.19	1
ST Gumbel	27.42 ±13.36	1
Marginalization		
Dense	93.37 ±0.42	256
Sparse	93.35 ±0.50	3.13 ±0.48

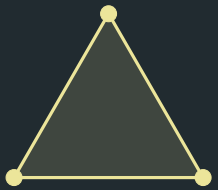
(Lazaridou et al., 2017; Correia et al., 2020)

Emergent Communication

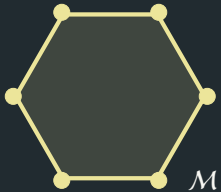
... but make it harder: $|\mathcal{Z}| = 256$, $|\mathcal{V}| = 16$

Method	success (%)	Dec. calls
Monte Carlo		
SFE	33.05 ±2.84	1
SFE+	44.32 ±2.72	2
NVIL	37.04 ±1.61	1
Gumbel	23.51 ±16.19	1
ST Gumbel	27.42 ±13.36	1
Marginalization		
Dense	93.37 ±0.42	256
Sparse	93.35 ±0.50	3.13 ±0.48



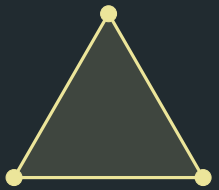


Δ

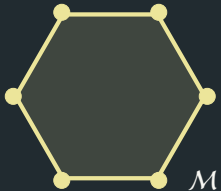


$\mathcal{M}$

$$\begin{aligned}\mathcal{M} &:= \operatorname{conv}\left\{\boldsymbol{a}_z : z \in \mathcal{Z}\right\} \\ &= \left\{\boldsymbol{A}\boldsymbol{p} : \boldsymbol{p} \in \Delta\right\} \\ &= \left\{\mathbb{E}_{Z\sim\boldsymbol{p}}\, \boldsymbol{a}_Z : \boldsymbol{p} \in \Delta\right\}\end{aligned}$$

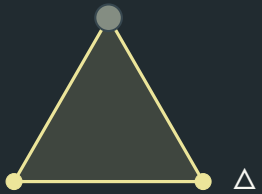


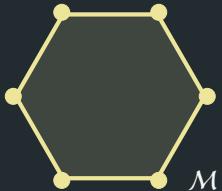
Δ



$\mathcal{M}$

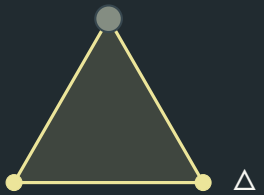
- **argmax** $\arg\max_{p \in \Delta} p^\top s$

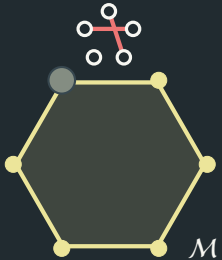




argmax $\arg\max_{\boldsymbol{p} \in \triangle} \boldsymbol{p}^\top \boldsymbol{s}$

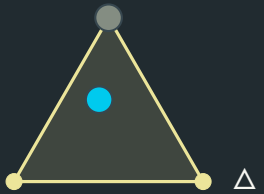
MAP $\arg\max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{t}$

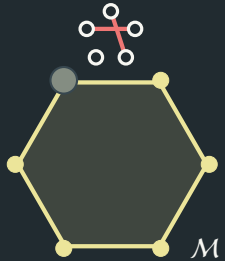




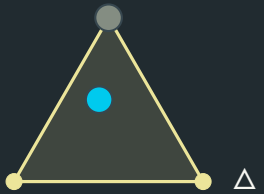
- **argmax** $\arg\max_{\boldsymbol{p} \in \Delta} \boldsymbol{p}^\top \boldsymbol{s}$
- **softmax** $\arg\max_{\boldsymbol{p} \in \Delta} \boldsymbol{p}^\top \boldsymbol{s} + \mathrm{H}(\boldsymbol{p})$

MAP $\arg\max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{t}$

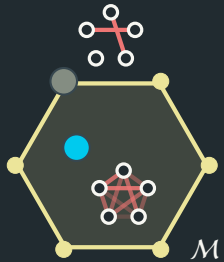




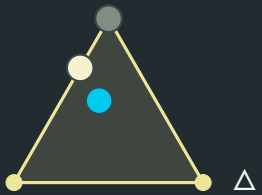
- **argmax** $\arg\max_{\boldsymbol{p} \in \triangle} \boldsymbol{p}^\top \boldsymbol{s}$
- **softmax** $\arg\max_{\boldsymbol{p} \in \triangle} \boldsymbol{p}^\top \boldsymbol{s} + \mathrm{H}(\boldsymbol{p})$



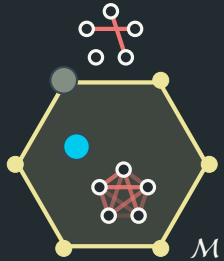
- **MAP** $\arg\max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{t}$
- **marginals** $\arg\max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{t} + \widetilde{\mathrm{H}}(\boldsymbol{\mu})$



- **argmax** $\arg\max_{p \in \triangle} p^\top s$
- **softmax** $\arg\max_{p \in \triangle} p^\top s + \mathrm{H}(p)$
- **sparsemax** $\arg\max_{p \in \triangle} p^\top s - \frac{1}{2}\|p\|^2$



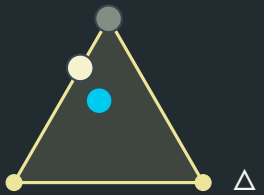
- **MAP** $\arg\max_{\mu \in \mathcal{M}} \mu^\top t$
- **marginals** $\arg\max_{\mu \in \mathcal{M}} \mu^\top t + \widetilde{\mathrm{H}}(\mu)$

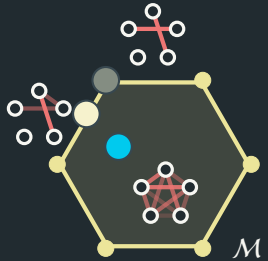


(Niculae, Martins, Blondel, and Cardie, 2018)

- **argmax** $\arg\max_{p \in \triangle} p^\top s$
- **softmax** $\arg\max_{p \in \triangle} p^\top s + \mathrm{H}(p)$
- **sparsemax** $\arg\max_{p \in \triangle} p^\top s - \frac{1}{2}\|p\|^2$

- **MAP** $\arg\max_{\mu \in \mathcal{M}} \mu^\top t$
- **marginals** $\arg\max_{\mu \in \mathcal{M}} \mu^\top t + \widetilde{\mathrm{H}}(\mu)$
- **SparseMAP** $\arg\max_{\mu \in \mathcal{M}} \mu^\top t - \frac{1}{2}\|\mu\|^2$

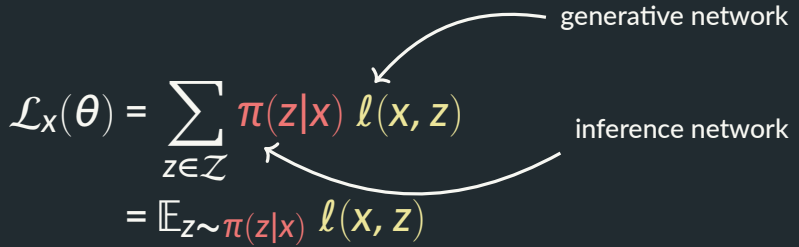




Bit-vector VAE

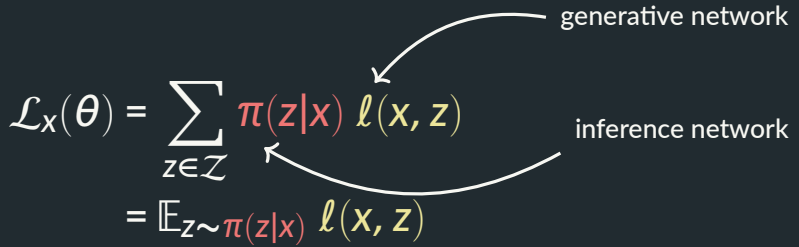
$$\mathcal{L}_x(\theta) = \sum_{z \in \mathcal{Z}} \pi(z|x)\, \ell(x, z)$$
$$= \mathbb{E}_{z \sim \pi(z|x)}\, \ell(x, z)$$

Bit-vector VAE



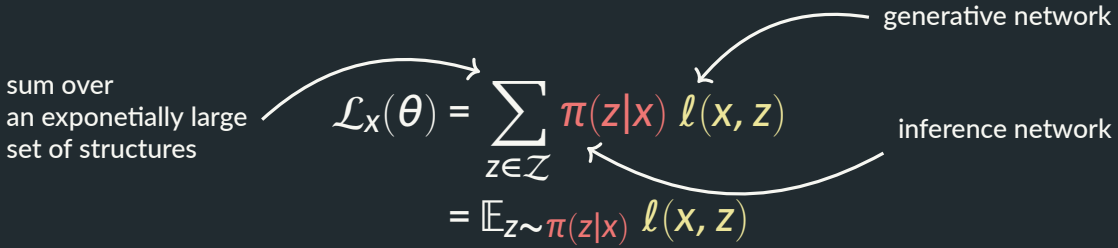
- VAE where z is a collection of D bits

Bit-vector VAE



- VAE where z is a collection of D bits
- Minimize the negative ELBO

Bit-vector VAE



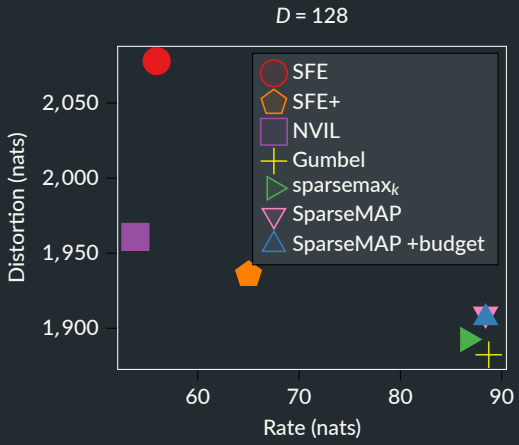
- VAE where z is a collection of D bits
- Minimize the negative ELBO

Bit-vector VAE

Method	$D = 32$	$D = 128$
Monte Carlo		
SFE	3.74	3.77
SFE+	3.61	3.59
NVIL	3.65	3.60
Gumbel	3.57	3.49
Marginalization		
Top-k sparsemax	3.62	3.61
SparseMAP	3.72	3.67
SparseMAP (w/ budget)	3.64	3.66

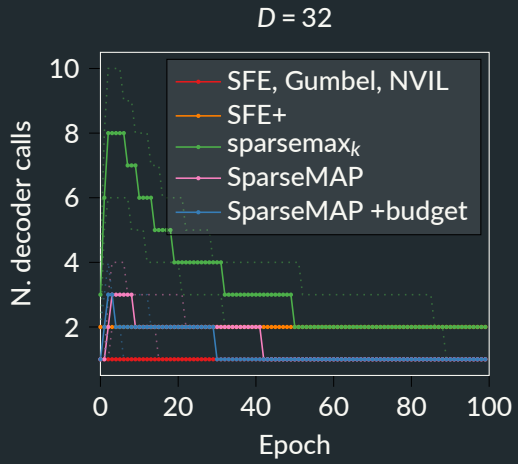
Bit-vector VAE

Method	$D = 32$	$D = 128$
Monte Carlo		
SFE	3.74	3.77
SFE+	3.61	3.59
NVIL	3.65	3.60
Gumbel	3.57	3.49
Marginalization		
Top-k sparsemax	3.62	3.61
SparseMAP	3.72	3.67
SparseMAP (w/ budget)	3.64	3.66



Bit-vector VAE

Method	$D = 32$	$D = 128$
Monte Carlo		
SFE	3.74	3.77
SFE+	3.61	3.59
NVIL	3.65	3.60
Gumbel	3.57	3.49
Marginalization		
Top-k sparsemax	3.62	3.61
SparseMAP	3.72	3.67
SparseMAP (w/ budget)	3.64	3.66



Bit-vector VAE

Method	$D = 32$	$D = 128$
Monte Carlo		
SFE	3.74	3.77
SFE+	3.61	3.59
NVIL	3.65	3.60
Gumbel	3.57	3.49
Marginalization		
Top-k sparsemax	3.62	3.61
SparseMAP	3.72	3.67
SparseMAP (w/ budget)	3.64	3.66