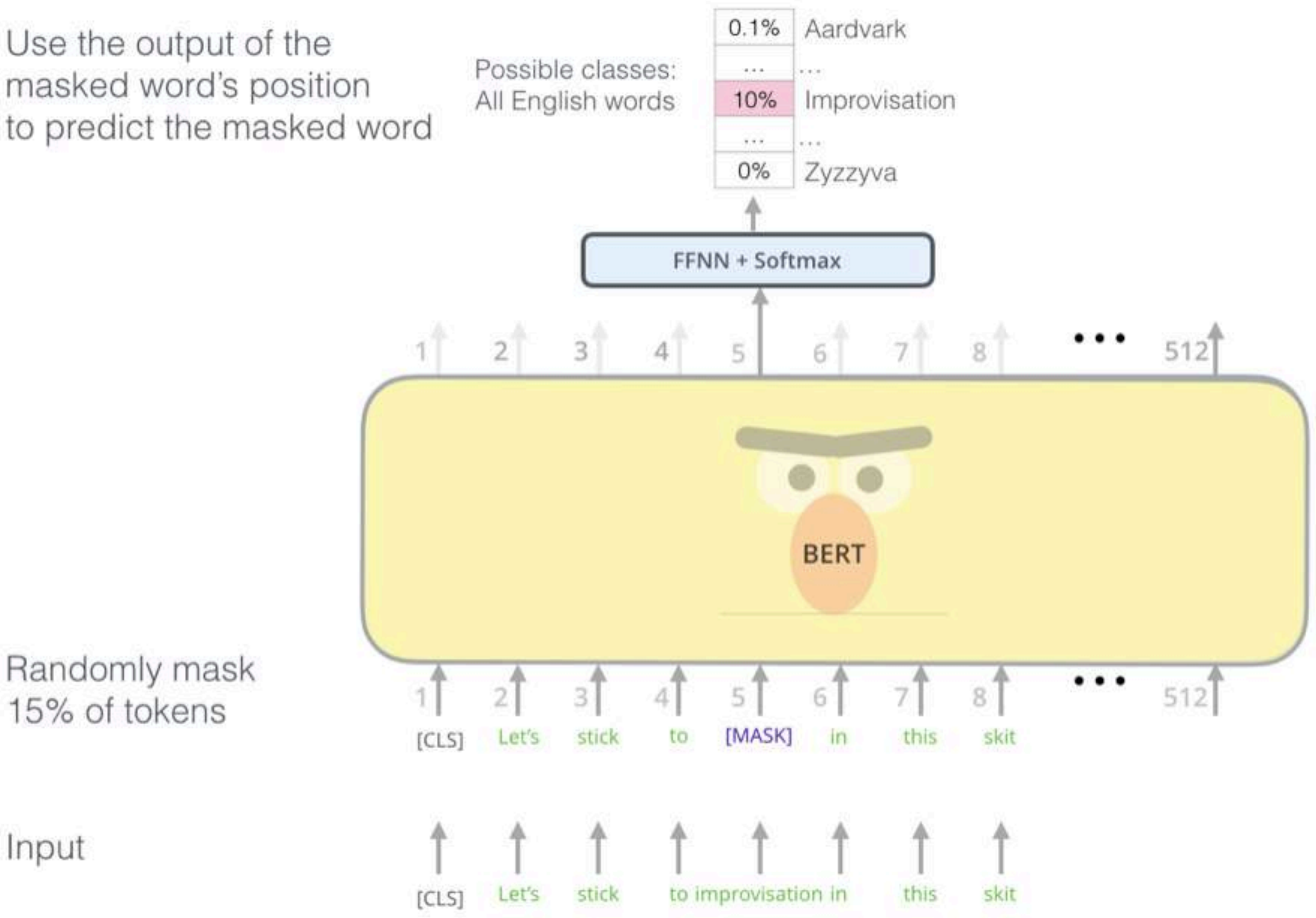
Can Unconditional Language Models Recover Arbitrary Sentences?

Nishant Subramani, Sam Bowman, Kyunghyun Cho
New York University

Hareesh Bahuleyan
Borealis AI

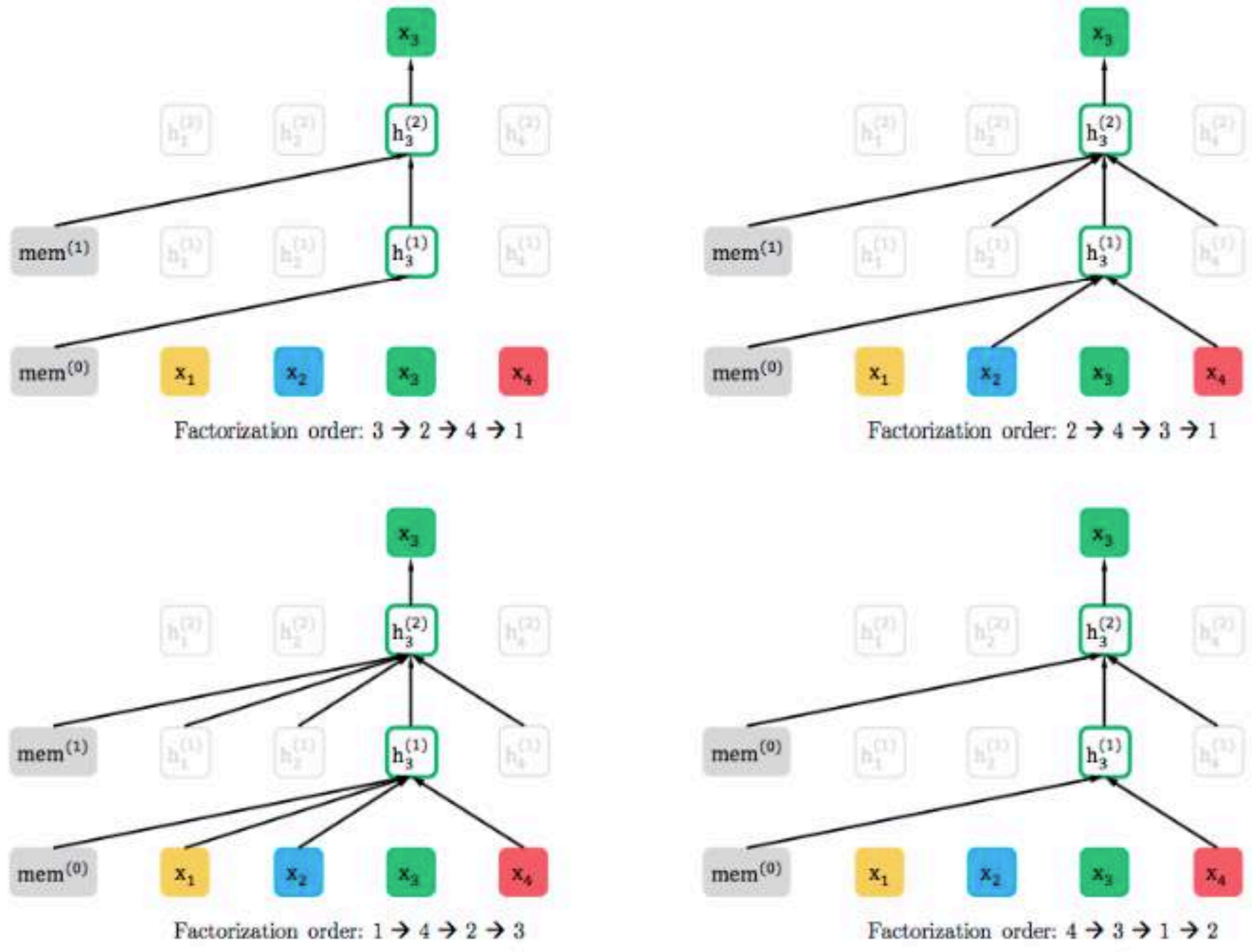
Background - Language Model Pre-training



BERT

(Devlin et al., 2018)

Image Source: https://jalammar.github.io/illustrated-bert/



XLNet

(Yang et al., 2019)

What are such models useful for ?

- learn contextualized representation of words

- can work effectively as **general purpose sentence encoders** in text classification with or without further fine-tuning

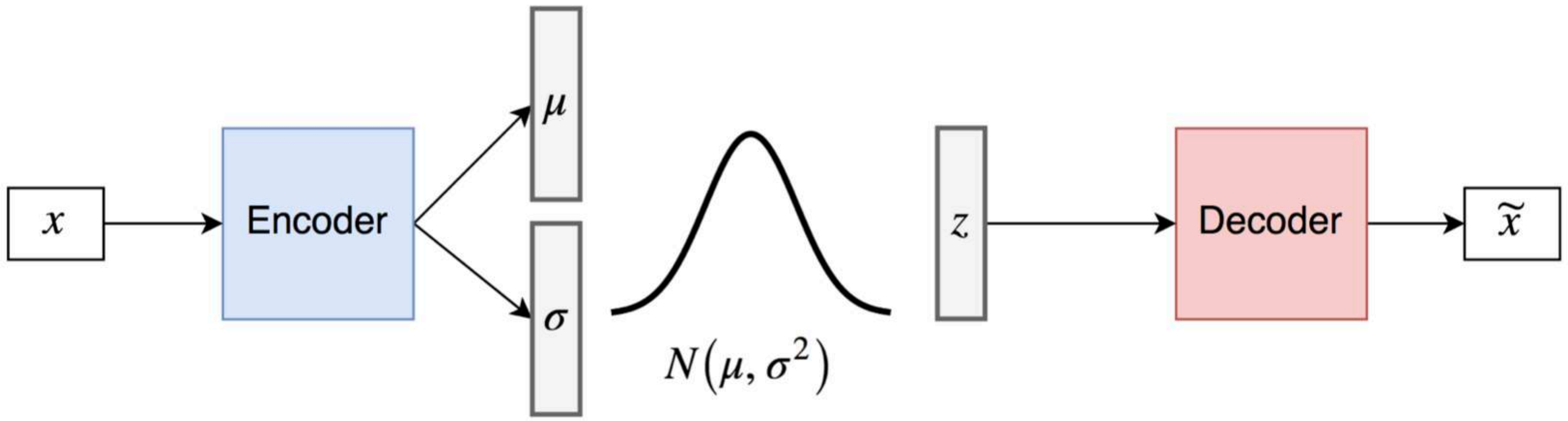
Is it possible to use a pre-trained language model as a **general-purpose decoder** in a similar fashion ?

Is there some continuous representation that can be passed to the LM to cause it to reproduce a desired sentence ?

Recap - Deep Generative Models

Variational Auto-Encoders (Kingma and Welling, 2013)

- Encoder, decoder, latent space.
- At generation time, discard encoder and
  - generate new samples
  - linear interpolation
  - arithmetic operation on z-vectors



$$\mathscr{L} = -\mathop{\mathbb{E}}_{\mathtt{z}^{(\mathtt{n})} \sim \mathtt{q}} \left[\log \mathtt{p}\left(\mathtt{x}^{(\mathtt{n})} \,|\, \mathtt{z}^{(\mathtt{n})}\right) \right] + \lambda \cdot \mathtt{KL}\left(\mathtt{q}\left(\mathtt{z}^{(\mathtt{n})} \,|\, \mathtt{x}^{(\mathtt{n})}\right) \| \mathtt{p}(\mathtt{z}) \right)$$

Generative Adversarial Networks (Goodfellow et. al, 2014)

- Generator and Discriminator
- Adversarial training
- Trained till equilibrium
- At generation time, discard discriminator
  - generate new samples
  - linear interpolation
  - arithmetic operation on z-vectors

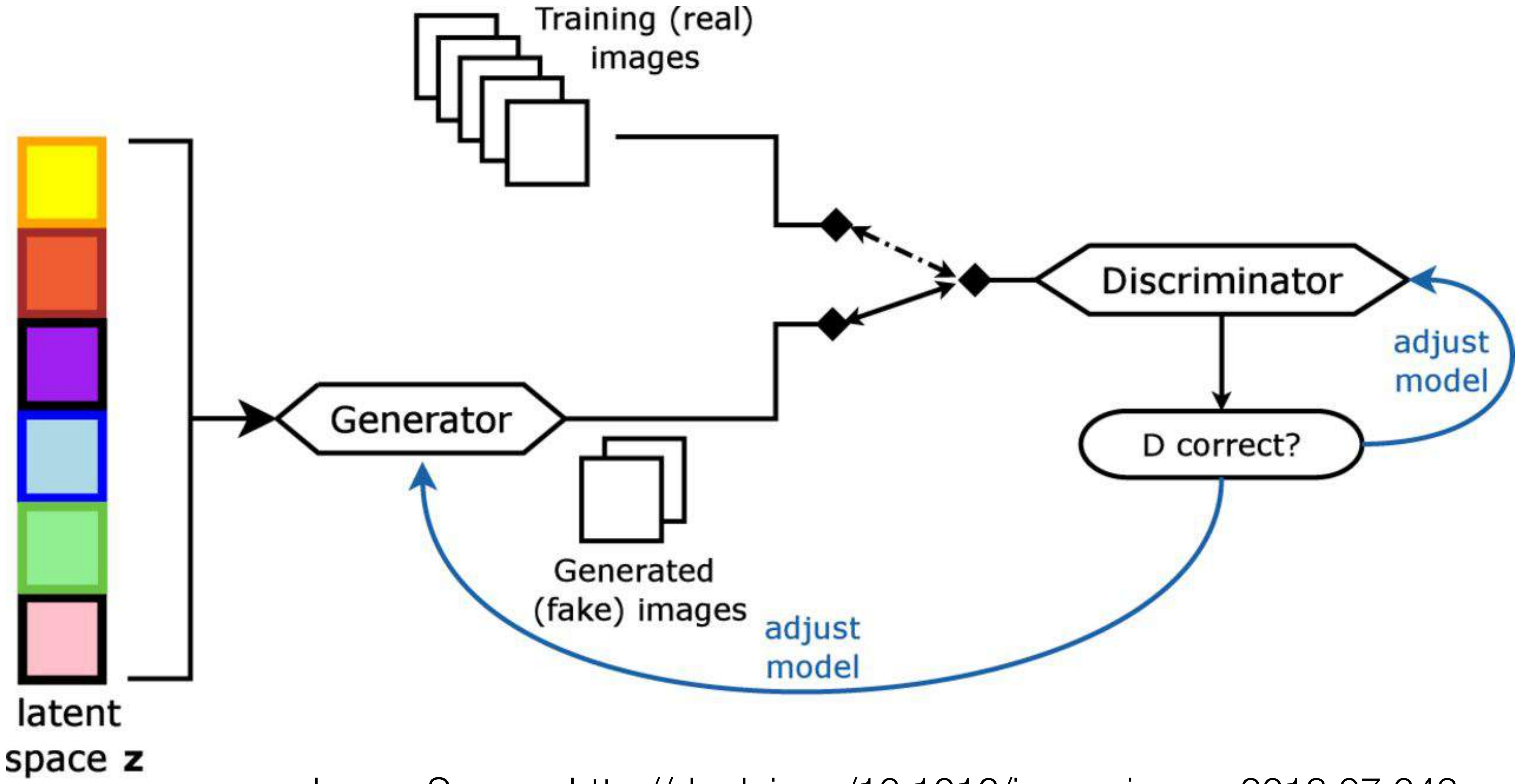
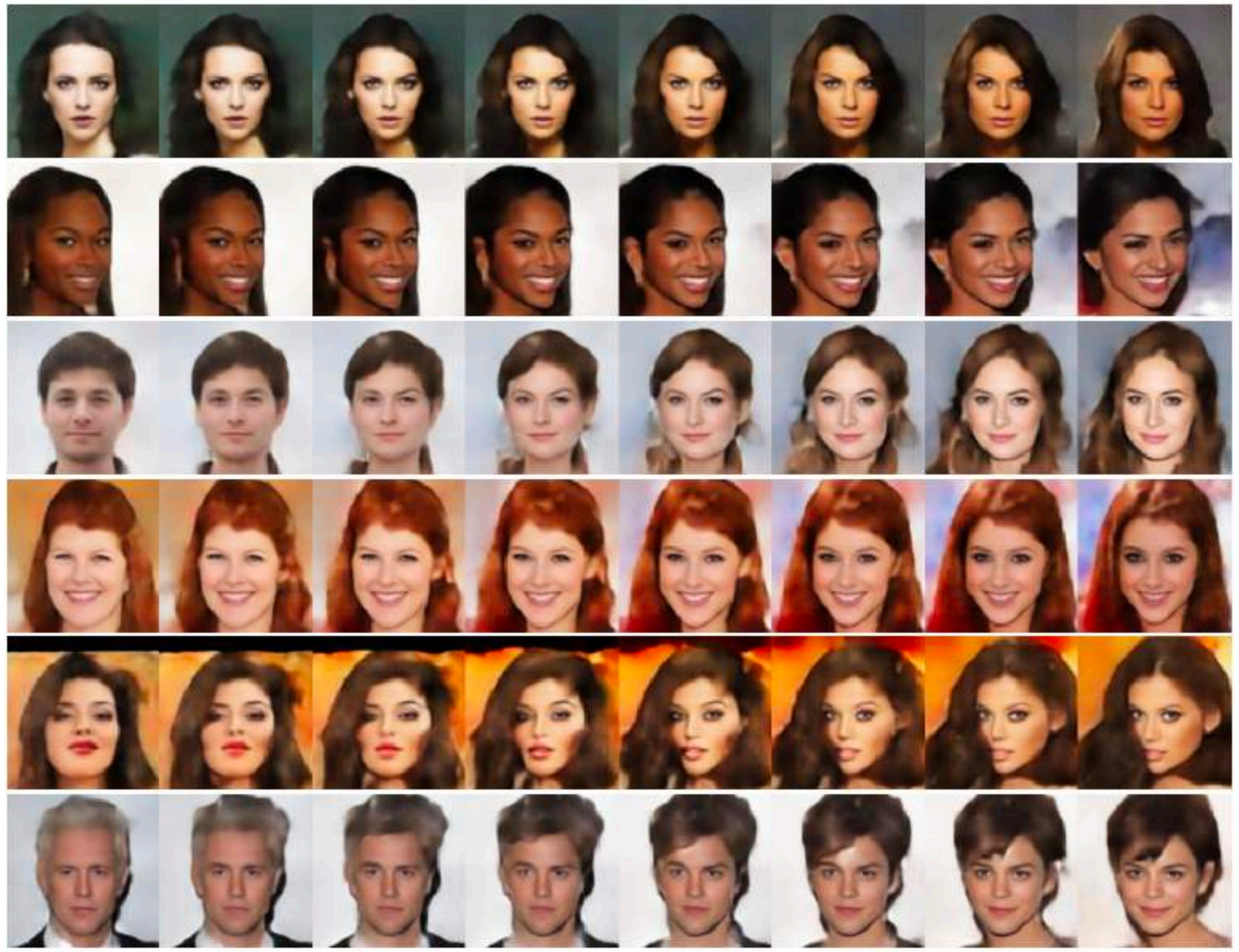


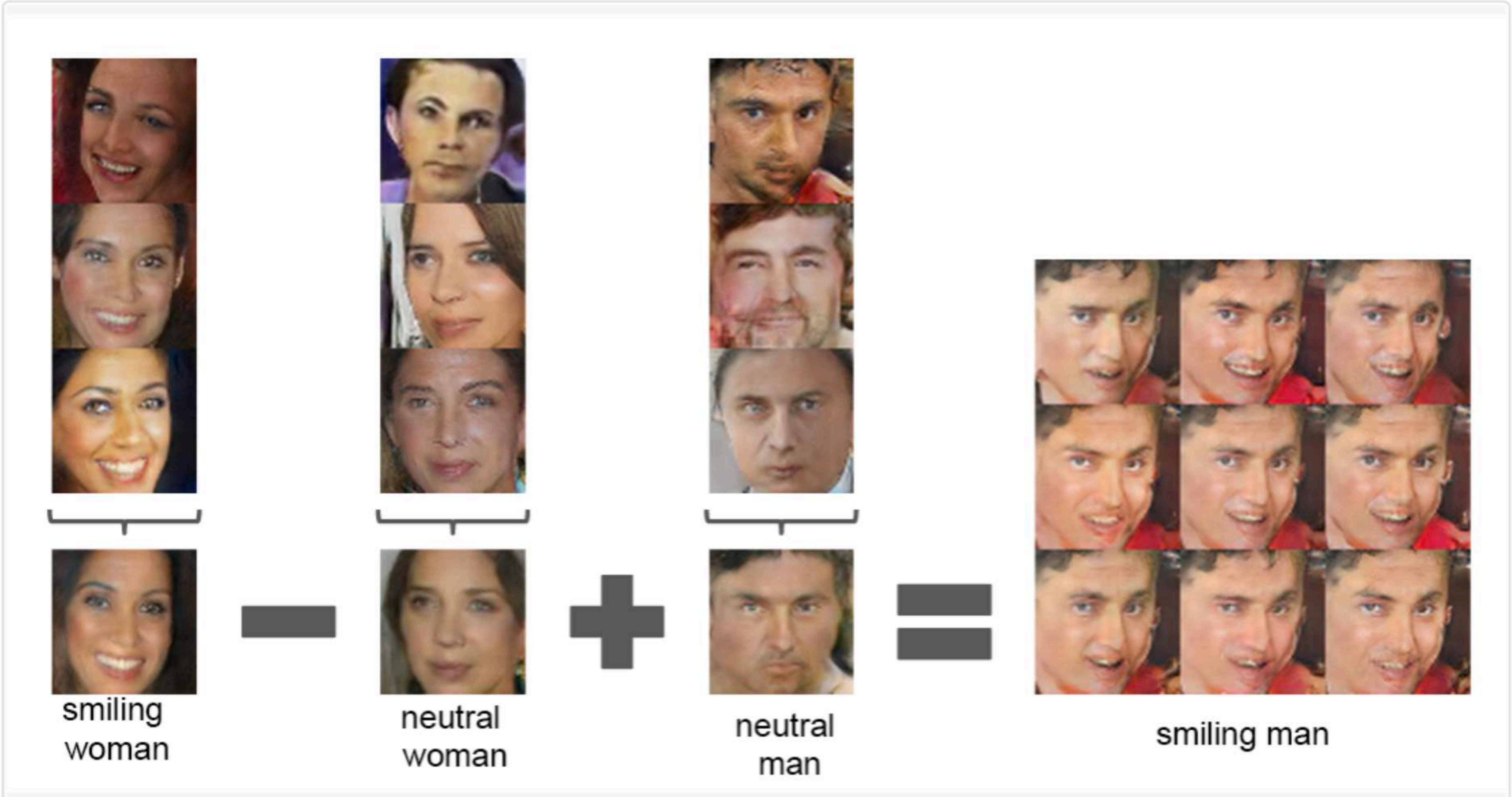
Image Source: http://dx.doi.org/10.1016/j.neuroimage.2018.07.043

$$\min_{G} \max_{D} \mathbb{E}_{x \sim P} \log d(x) + \mathbb{E}_{z \sim Q}(1 - \log d(g(z)))$$

Latent Space Properties

Linear interpolations in the noise space into semantically meaningful non-linear interpolations in the image space





- Linear arithmetic in the noise space
- because of the ability to disentangle factors of variation

Generative Latent Optimization (GLO, Bojanowski et al, 2018)

- Is it possible achieve desirable properties of GANs without adversarial training ?

- An auto-encoder where the latent representation is not produced by a parametric encoder, but learned freely in a non-parametric manner

Training Objective : $$\min_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N}\left[\min_{z_i\in\mathcal{Z}} \ell\left(g_\theta(z_i), x_i\right)\right]$$

- Jointly optimizes the z_i and the model parameters θ with stochastic gradient descent

- Demonstrate similar levels of latent space properties

Recurrent Language Models

$$p\left(x_1, \ldots, x_T\right) = \prod_{t=1}^{T} p\left(x_t \mid x_1, \ldots, x_{t-1}\right)$$

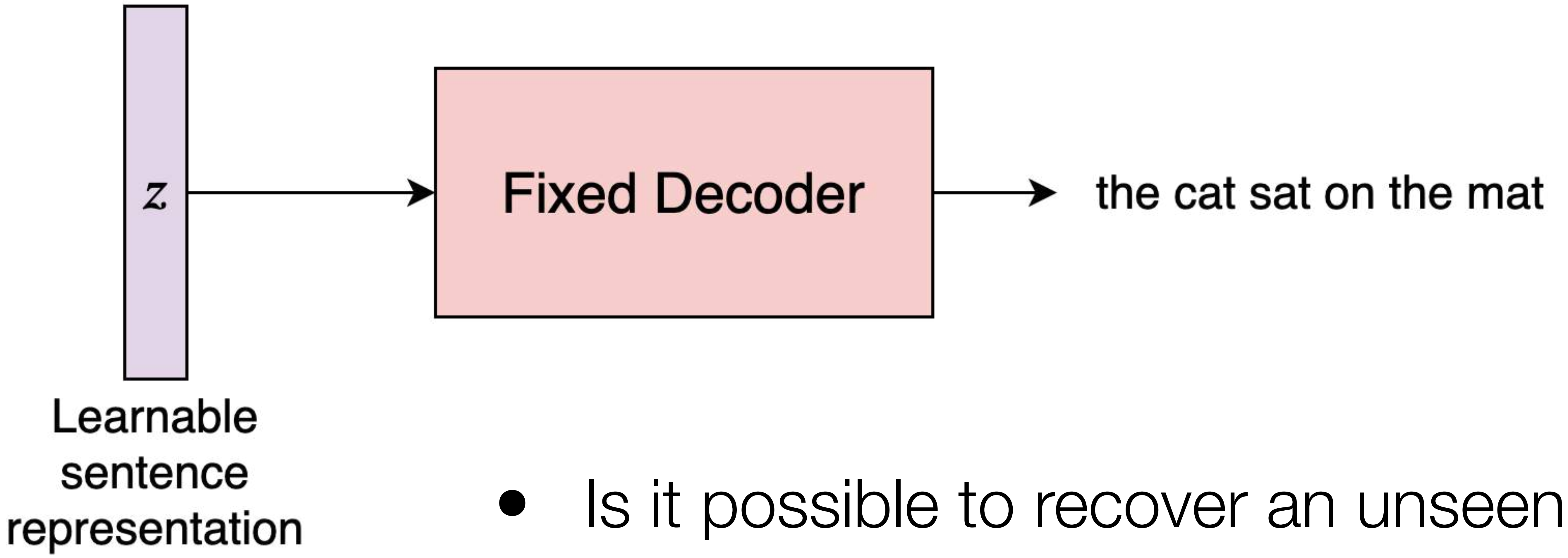
- standard autoregressive RNN based training
- stochastic gradient decent with negative log likelihood loss
- Once learning is complete, a LM can be used in two ways:

1. To score - compute the log-probability of a newly observed sentence
2. To generate a new sentence, conditioned on a few tokens (either greedy or beam search)

Defining the Sentence Space

- Train the RNN-LM on a large text corpus

- Fix the weights of the LM:

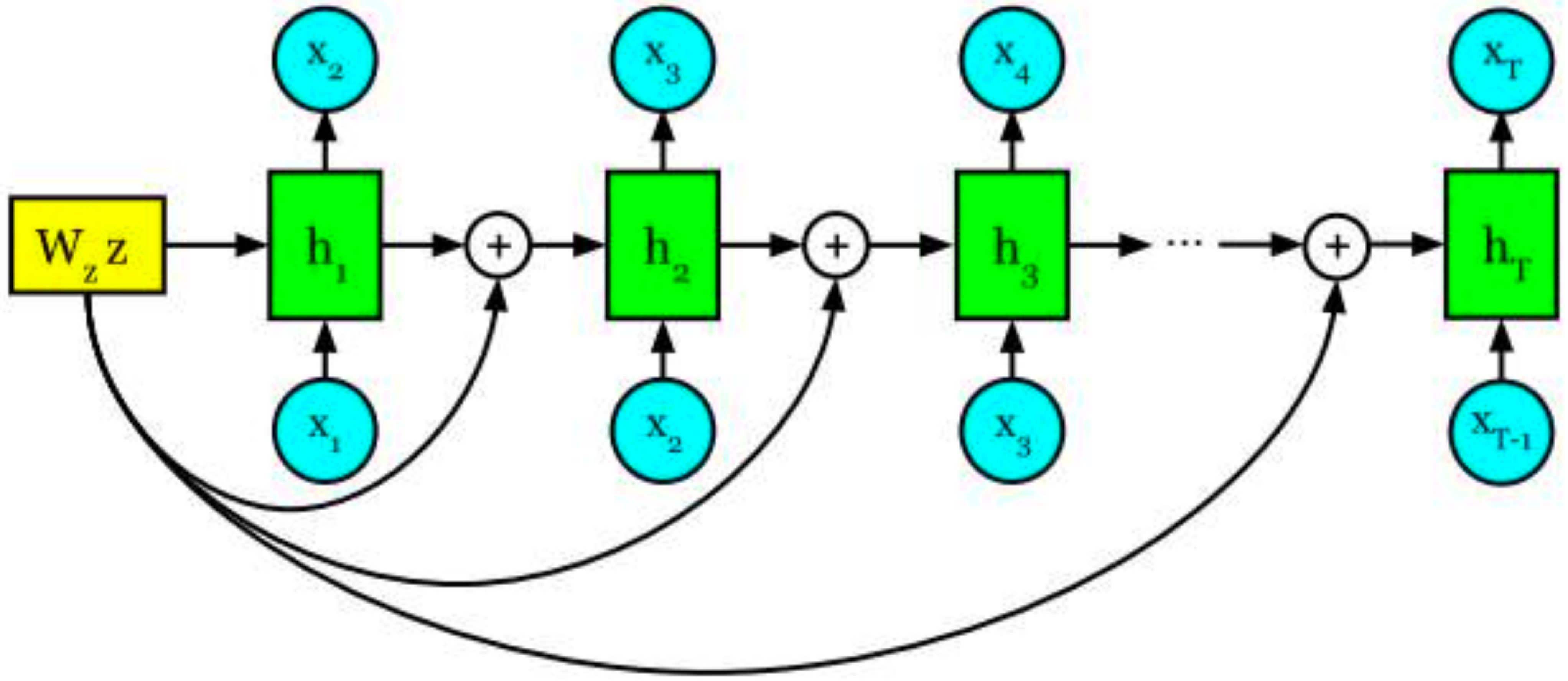


- Is it possible to recover an unseen sentence ?

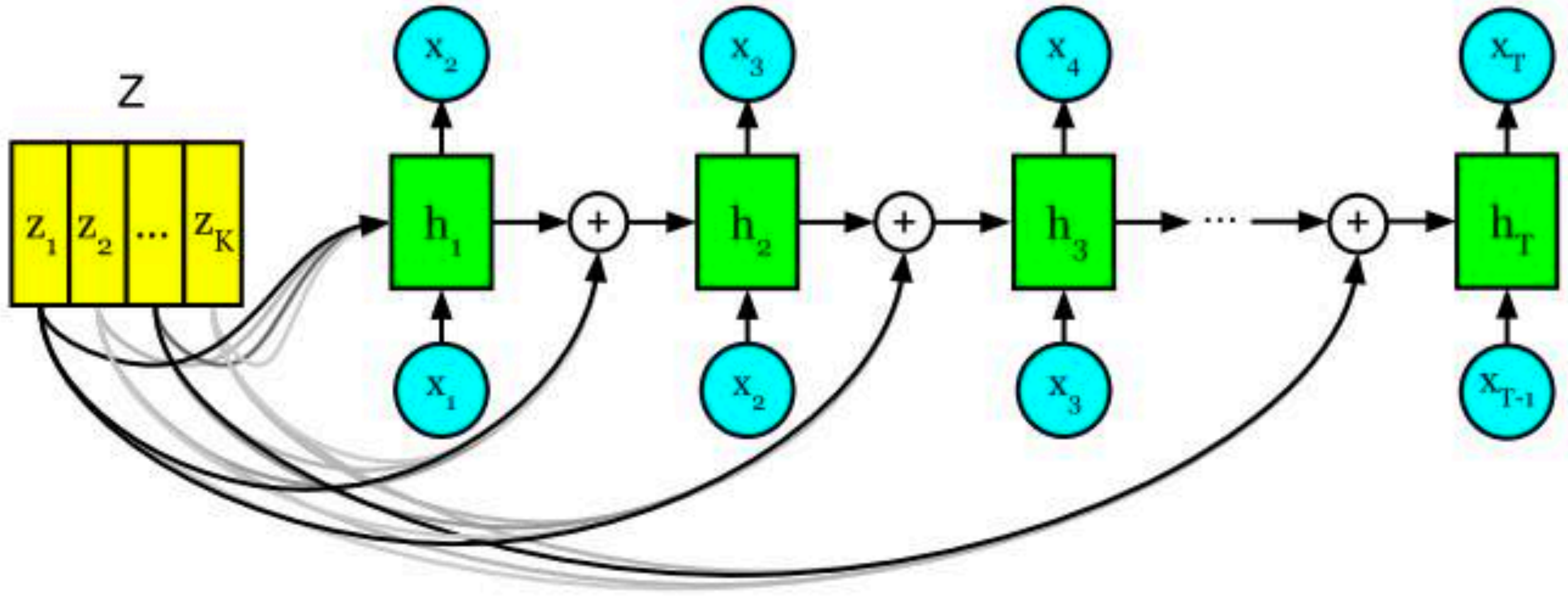
- Does there exist a representation z, that has all information corresponding the above sentence ?

Feeding z to the Decoder

- Add bias terms to the previous hidden and cell state at each time step (*d-dimensional* LSTM)



Case I - dim(z) < d



Case II - dim(z) > d

$$h_{t-1} = f_\theta(h_{t-2} + \underline{z'}, x_{t-1})$$

$$z' = \begin{cases} W_z z, & \text{if } \dim(z) \leq d^* \\ \text{softmax}(h_{t-2}^\top Z) Z^\top, & \text{if } \dim(z) > d^* \end{cases}$$

Using the sentence space

Forward Estimation (X → Z):

- estimate z by maximizing the log-probability of the given sentence under this modified model, while fixing the original parameters θ

$$\hat{z} = \operatorname*{argmax}_{z \in \mathcal{Z}} \sum_{t=1}^{T} \log p\left(x_t \,|\, x_{<t}, z\right)$$

- highly non-convex, potentially leading to multiple approximately optimal z's.
- use nonlinear conjugate gradient method with a limit of 10k iterations

Using the sentence space

Backward Estimation (Z → X):

- Given a vector z, estimate the most plausible sentence: (x1, . . . , xT)
- This is a combinatorial optimization problem and cannot be solved easily!
- Instead use beam search approximation

    - to choose the best sentence after decoding multiple of them

Evaluation

- **Recoverability:**
  - how much information about the original sentence x = (x1, . . . , xT) ∈ X is preserved in the re-parameterized sentence space Z
  - First <u>forward-estimate</u> the sentence vector z ∈ Z given x ∈ X
  - Then, reconstruct the sentence $\hat{x}$ from the estimated z via <u>backward estimation</u>

Evaluation

- Compare the original sentence to the reconstructed sentence

  1. Exact Match (EM)

$$\sum_{t=1}^{T} \mathbb{I}\left(x_t = \hat{x}_t\right)/T$$

  2. Prefix Match (PM): longest consecutive sequence of tokens that are perfectly recovered from the beginning of the sentence divided by the sentence length.

  3. BLEU: based on n-gram overlap

- What is the minimum dimension *d* of the LM needed to achieve a specified <u>recoverability τ</u> under the model θ ?

Experimental Setup

Corpus:
LM trained on 50M sentences from English Gigaword corpus, ~1.8M for validation and test

Model:
2-layer LSTM: 256d (Small) | 512d (Medium) | 1024d (Large)

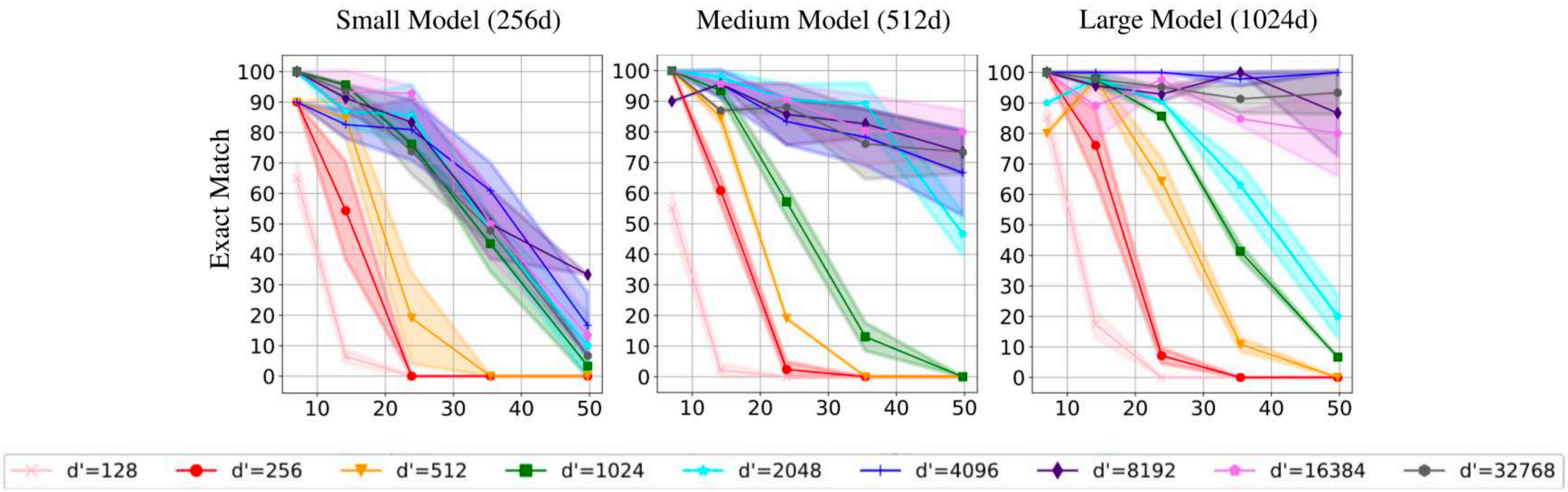
Table 1: Language modeling perplexities on English Gigaword for the models under study

		$\|\text{Train}\| = 10M$		$\|\text{Train}\| = 50M$	
Model	d	Dev Ppl.	Test Ppl.	Dev Ppl.	Test Ppl.
SMALL	256	122.9	125.2	77.2	79.2
MEDIUM	512	89.6	91.3	62.1	63.5
LARGE	1024	65.9	67.7	**47.4**	**48.9**

Sentence space:
- 128, 256, 512, 1024, 2048, 4096, 8192, 16384 and 32768 dimensions
- 10 random initializations of z, 10 random projection matrices for the optimization procedure

Results and Analysis

- Model capacity defined by $d* = 2dl$
- Recoverability increases as $d*$ increases, until $d' = d*$.
- Nearly perfect recoverability for the large model when $d' = 4096$ achieving EM ≥ 99
- LM trained with more data (50M vs 10M sentences), tends to have better recoverability



Results and Analysis

Effective Dimension of the Sentence Space (recoverability EM > 0.8):

- Large model : 4096d ; slight degradation when increasing beyond that
- Medium model : 2048d ; no real recoverability improvements when increasing beyond that
- Small model : 8192d which is much greater than $d* = 4096$

Negative correlation between recoverability and sentence length

Sources of Randomness

Two points of stochasticity in the proposed framework:

- z initialization and resulting non-convexity of the optimization procedure in forward estimation
- the sampling of a random projection matrix Wz
- small standard deviations => these sources of randomness have minimal impact on recoverability

Summary of Findings

- Able to generate held-out sentences near perfect recoverability (with sufficient model capacity)
- Recoverability increases with the dimension of the re-parametrized space until it reaches the model dimension.
- Recoverability improves with the size and quality of the language model
- Recoverability is more difficult for longer sentences.
- Choice of optimizer is crucial

Conclusions

- A frozen pre-trained language model with sufficient capacity can decode an arbitrary vector into a sentence
- Optimization based forward estimation
- Beam search approximation for backward estimation
- Recoverability metrics
- Future work: language models beyond plain LSTMs

Critical View

- Only very simple analyses of changing dimensions of model and sentence space
- Could have studied if the sentence latent space has useful properties

    - If certain dimensions correspond to certain attributes (disentanglement)

    - Interpolation between sentences

Thank You