Variational Auto-encoding

Advanced NLP: Summer 2023

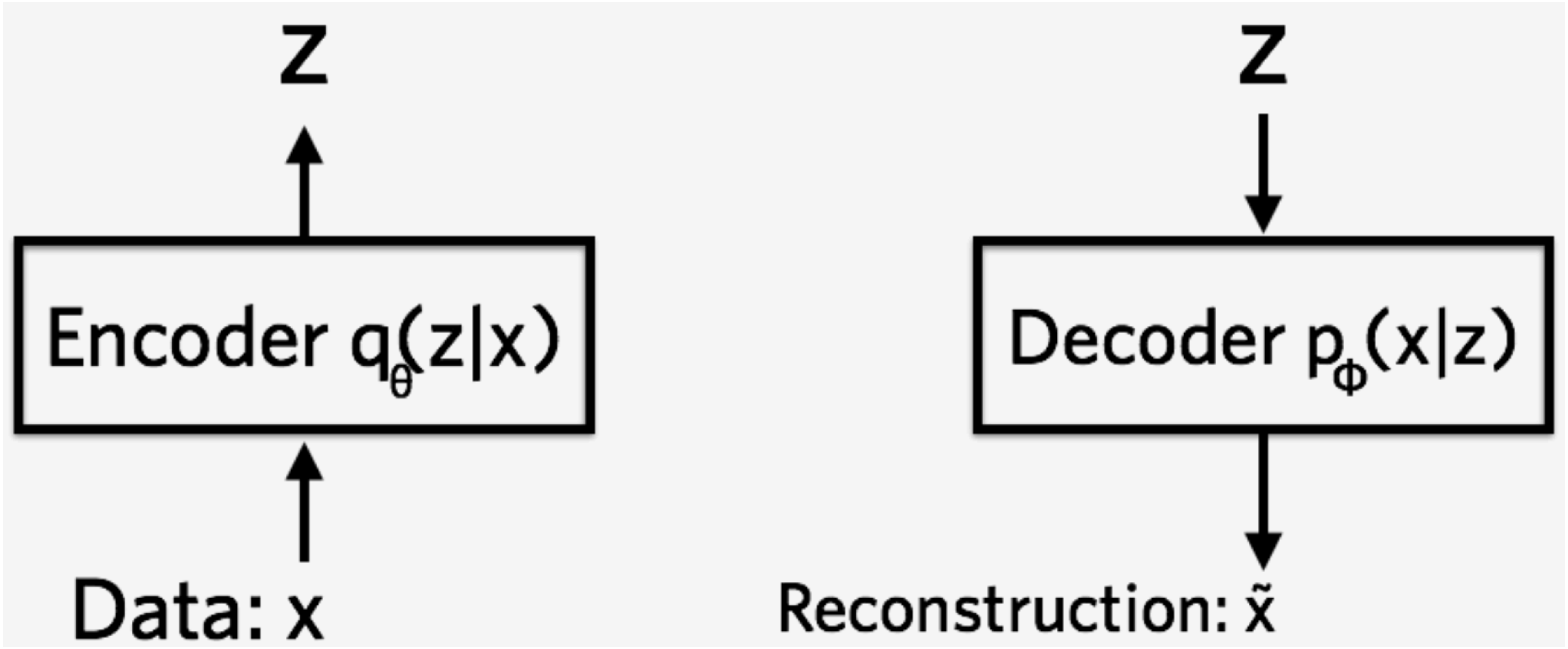
Anoop Sarkar

Encoder-Decoder neural nets

- An encoder q takes an input x and encodes it into a hidden representation z using some parameters θ. Encoder: $q_\theta(z \mid x)$

- An decoder p takes a hidden layer z and decodes it into an output $\tilde{x}$ using some parameters ϕ. Decoder: $\tilde{x} \sim p_\phi(x \mid z)$

- The output $\tilde{x}$ should be similar to but not necessarily identical to the true x



Autoencoder

Autoencoder loss

- How much information is lost by going from x to z and then back to $\tilde{x}$?
- We measure the information loss by representing using z using reconstruction log-likelihood
- $\log p_\phi(x \mid z)$ measured in nats (bits are base 2, nats are base e)
- The loss function for an *variational* autoencoder is the negative log likelihood with a regularizer
- For single data point x_i we compute the above loss l_i.
- Total loss for the dataset: $\sum_i l_i$

Variational autoencoder loss

- Loss function l_i for datapoint x_i is
- $l_i(\theta, \phi) = -E_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i \mid z)] + KL(q_\theta(z \mid x_i) || p(z))$
- First term is the expected negative log-likelihood of the data point x_i
- We want to place the most probability mass on the true output x_i
- Second term is the regularizer: the Kullback-Leibler divergence between the encoder distribution $q_\theta(z \mid x)$ and $p(z)$
- $p(z)$ is used to reward "good" values of the hidden representation that are efficient, can be sampled from easily and do not memorize the dataset.
- $z = \mu + \sigma \circ \epsilon$ where $\epsilon \sim$ Normal$(0,1)$ and $\circ$ is element wise multiplication

Reparametrize z

- We want to use gradient descent to learn $q_\theta(z \mid x)$
- Need to take derivative of $p(z)$ wrt θ
- We reparametrize z
- $z = \mu + \sigma \circ \epsilon$ where $\epsilon \sim$ Normal$(0,1)$ and $\circ$ is element wise multiplication
- Now we can take derivatives of $p(z)$ wrt μ and σ
- Output of $q_\theta(z \mid x)$ is a vector of μ's and σ's

Variational autoencoder loss

- The regularizer term keeps the representation of z sufficiently diverse
- Without the regularizer, given large enough z the encoder-decoder would simply memorize the entire dataset
- Two different x_i and x_j that are actually very close to each other would end up learning very different z_i and z_j which defeats the purpose of modeling similarity between inputs.
- The regularizer would make sure z_i and z_j cannot get too far from each other unless x_i is very different from x_j
- The variational autoencoder (vae) is trained using gradient descent

Variational autoencoder loss

- Unfortunately, gradient descent requires computing distribution $q_\theta(z \mid x)$
- This is exponential because it is over all configurations of latent variable z
- Variational inference approximates this using a distribution $q_\lambda(z \mid x)$
- λ is the variational parameter which indexes a family of distributions
- If q is a normal distribution then λ_{x_i} would be the mean μ and variance σ^2 for each data point x_i
- $\lambda_{x_i} = (\mu_{x_i}, \sigma^2_{x_i})$

Tractable variational inference

- We want to measure how well does the variational distribution $q_\lambda(z \mid x)$ approximate the true distribution $q(z \mid x)$
- We use the KL divergence again: $\mathrm{KL}(q_\lambda(z \mid x) \| q(z \mid x))$
- The optimal approximate distribution involves finding the optimal variational parameters λ
- $q^*_\lambda(z \mid x) = \arg\min_\lambda KL(q_\lambda(z \mid x) \| q(z \mid x))$
- Unfortunately, this is still intractable

Tractable variational inference

- Define ELBO(λ) the Evidence Lower BOund of λ
- ELBO(λ) = $E_{z\sim q_\lambda}[\log p(x \mid z)] - E_{z\sim q_\lambda}[\log q_\lambda(z \mid x)]$
- Minimizing KL$(q_\lambda \| q)$ wrt λ is equivalent to maximizing ELBO(λ)
- For each data point x_i
- ELBO$_i(\lambda) = E_{z\sim q_\lambda(z|x_i)}[\log p_\phi(x_i \mid z)]$ – KL$(q_\lambda(z \mid x_i) \| p(z)]$
- Maximizing ELBO$_i(\lambda)$ is equivalent to minimizing $l_i(\theta, \phi) = - E_{z\sim q_\theta(z|x_i)}[\log p_\phi(x_i \mid z)] + KL(q_\theta(z \mid x_i) \| p(z))$

Applications: Image generation

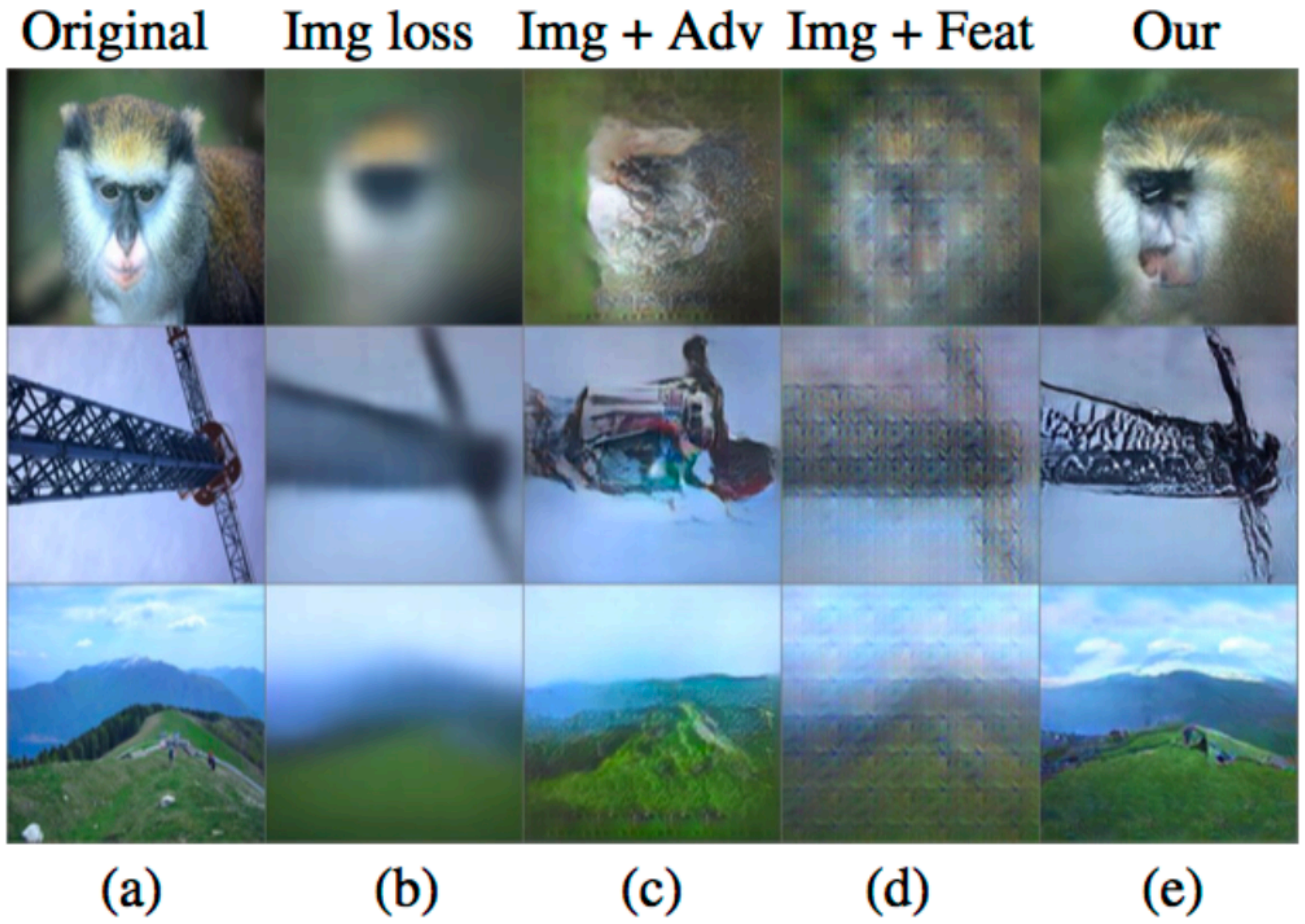
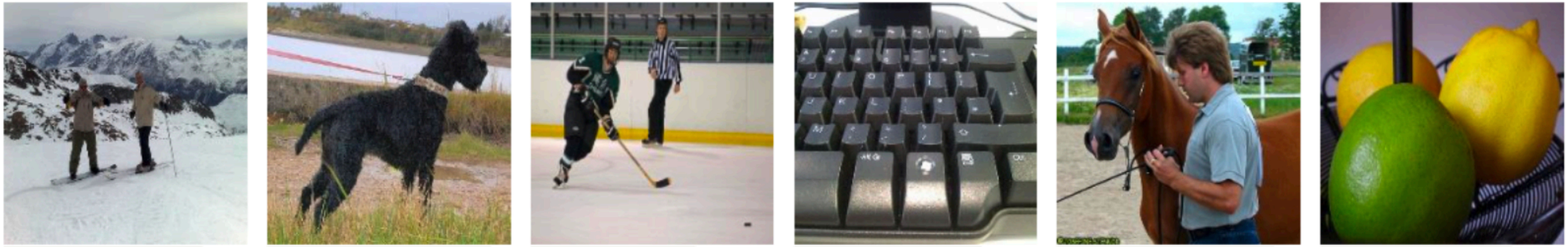


Figure 1: Reconstructions from AlexNet FC6 with different components of the loss.

A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. *arXiv preprint arXiv*:1602.02644, 2016.

Applications: caption generation



a man with a snowboard next to a man with glasses

a big black dog standing on the grass

a player is holding a hockey stick

a desk with a keyboard

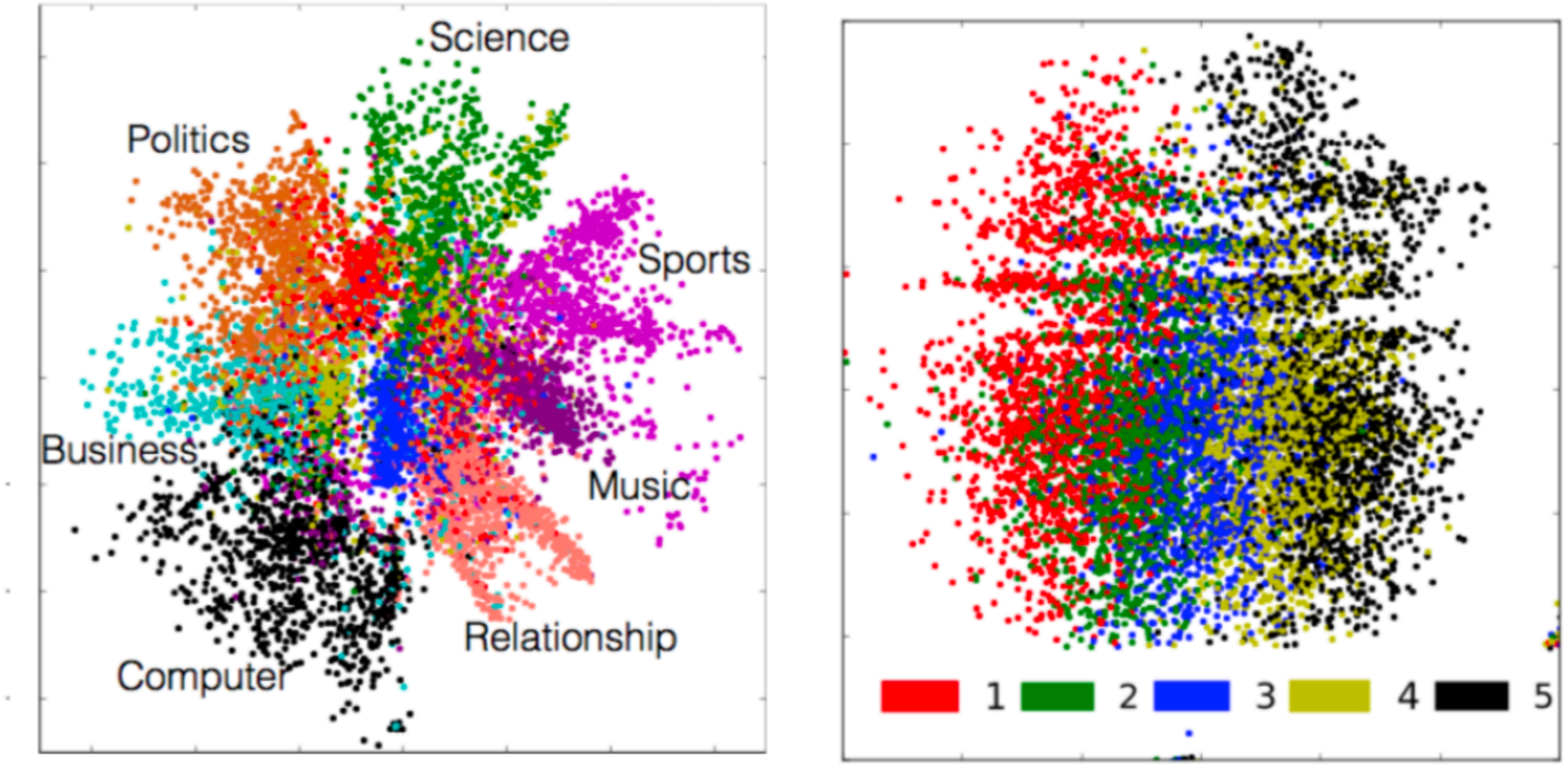
a man is standing next to a brown horse

a box full of apples and oranges

Figure 2: Examples of generated caption from unseen images on the validation dataset of ImageNet.

Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin. Variational autoencoder for deep learning of images, labels and captions. In NIPS, 2016.

Applications: document clustering



(a) Yahoo

(b) Yelp

Figure 3: Visualizations of learned latent representations.

Z. Yang, Z. Hu, R. Salakhutdinov, and T. Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of The 34rd International Conference on Machine Learning*, 2017.

Applications: sign clustering

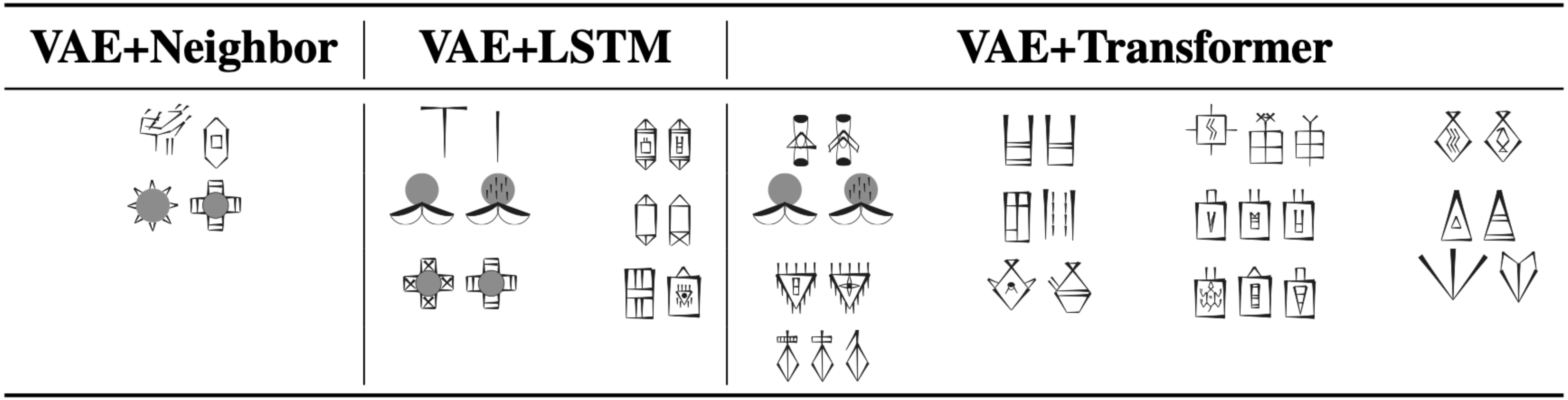


Table 3: Pairs/triplets of character images which have distinct labels in the working signlist, but which our models merge into single clusters.