Disentangled Representations for Natural Language Processing
Ph.D. Thesis Proposal

Michal Auersperger
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
auersperger@ufal.mff.cuni.cz

Abstract

We introduce the problem of disentangled representation learning and provide an overview of the relevant literature both in and outside of NLP. We find that unsupervised learning of such representations in NLP is underdeveloped. We motivate the research and specify experiments meant to address the issue.

1 Introduction

Reflecting on the stormy development of machine learning research in the recent years, we can observe reactions demonstrating either of two somewhat opposing attitudes. The first, *sky-is-the-limit* attitude focuses on the successes of the systems in many tasks across different domains, with systems often achieving human parity or beyond.

The second, more sobering approach stresses the shortcomings of the current paradigm, sometimes accenting differences to biological systems. The concept of disentangled representations is closer to the latter category.

Unfortunately, there is not yet a broadly accepted formal definition of disentanglement, so we start with the following intuition:

> A disentangled representation should capture and **isolate** all potentially relevant characteristics of the data.

This vague starting point should be made more precise in the following section (2), where we also argue why a disentangled representation is desirable, i.e. what problems is it supposed to solve. Section (3) then presents the main approaches towards learning disentangled representations from the literature. The proposed areas of future work are discussed in Section (4), followed by a short conclusion (Section 5)

2 Disentanglement

Disentangled representation learning is typically discussed in the context of generative modelling. Observations $\mathbf{x}$ are assumed to be realizations of some hidden, generative factors $\mathbf{z}$. Thus the data can be described as originating by following a two step procedure: 1) sampling the generative factors from a prior distribution and 2) sampling a concrete observation from the conditional distribution (likelihood):

$$\mathbf{z} \sim p(\mathbf{z}), \mathbf{x} \sim p(\mathbf{x}|\mathbf{z}). \qquad (1)$$

For example in a room with a single cube, the individual factors in $\mathbf{z}$ could contain the position, color, size and rotation of the cube as well as position and type of the source of light (and many other things). The observed $\mathbf{x}$ could be a photograph of the room. Ideally, given an observation, one would be able to reconstruct the generative factors as these would be the perfect representation of the real world state not distorted by, say, our perceptual system. In practice, this ideal is not to be expected, but if one learns to represent $\mathbf{x}$ by a code $\mathbf{c}$ such that when a single generative factor from $\mathbf{z}$ changes, only a small part of $\mathbf{c}$ changes, then the representation is said to be disentangled. In the example above, moving the source of light (z_i, z_j, z_k) could affect the values of all the pixels in a photograph $(\mathbf{x})$, but the change of a disentangled representation of the scene would be very local (c_a, c_b, c_c).

2.1 Motivation

Deep learning has been able to beat the best players in different board (Silver et al., 2016) or computer (Mnih et al., 2015) games, surpass human judges on various classification tasks (He et al., 2015), drive cars in real traffic (Buehler et al., 2009), and translate between languages achieving comparable results to human translators (Barrault et al., 2019).

Despite the impressive ever-growing track record some shortcomings of current deep learning approaches attract the attention.

The models are known to be extremely **data-hungry** (Lake et al., 2015, 2017; Gu et al., 2016; Higgins et al., 2018; Achille et al., 2018). In fact, *AlphaGo* has learned from more than 100 million games while its opponent Lee Sedol is estimated to play about $50,000$ games in his entire life (Lake et al., 2017). Lake et al. (2015) show that based on a single example, people can learn (to recognize and generate) a new handwritten character and be much more successful than a neural network classifier (trained on many alphabets). Moreover, the classifier is limited to the recognition task only.

Another related problem with current deep learning is poor **generalization** (Garnelo et al., 2016; Higgins et al., 2017a; Achille et al., 2018). The models often overfit to a task by learning only features that can be directly utilized (Achille et al., 2018). When a model encounters a new task it learns a new set of features from scratch instead of reusing previously learned information.

Not only is there little improvement on the performance on the new task but also the models tend to struggle to retain the previously learned information. By learning a new set of features, a model can overwrite the previous ones which leads to a drop of performance on the initial task (Rusu et al., 2016). This effect is known as *catastrophic forgetting* (French, 1999). Although natural cognitive systems exhibit some forgetting, they rarely overwrite all previous information.

Comparing human performance with the performance of artificial neural networks often points to an important role of prior knowledge. People carry their experience (be it individual or ancestral) to any new task while artificial neural networks start usually almost from scratch.[1] The importance of early inductive biases in people is well documented in cognitive literature (see Lake et al. (2017) for an overview). But even in the context of artificial neural networks, embedding certain assumptions about the structure of the data into the architecture proved successful as demonstrated by convolutional or recurrent NNs processing images and sequential data, respectively. Requiring learned representations to be disentangled can be seen as another type of inductive bias.

2.2 Purported benefits

Having access to explicit factors of variation in the data is believed to be desirable for many reasons. An information tied to a specific factor could be **transferred** to other tasks (Bengio et al., 2013; Lake et al., 2017; Higgins et al., 2017a; Achille et al., 2018). A new task would require the model to only acquire knowledge about the newly relevant factors potentially increasing **data-efficiency** (Lake et al., 2017; Higgins et al., 2017a; Achille et al., 2018; van Steenkiste et al., 2019). Performance at a given task could be explicitly connected to the relevant factors which would make the model more **robust** (less sensitive to irrelevant changes in other variables) (Bengio et al., 2013; van Steenkiste et al., 2019). Knowing probabilities of isolated factors, one could easily detect **anomalies** (Barlow et al., 1989; Schmidhuber, 1992), infer **causality** (Barlow et al., 1989) resulting in further benefits (Lake et al., 2017), or represent data more compactly (Barlow et al., 1989; Schmidhuber, 1992) Moreover, disentangled representations would be **interpretable** (Kulkarni et al., 2015; Chen et al., 2016; Higgins et al., 2017a) and would lend themselves to performing **interventions and counterfactuals** (Locatello et al., 2019; Peters et al., 2017). It is also believed that the representations should be learnable by **unsupervised** learning (Bengio et al., 2013).

2.3 Defining disentanglement

As stated above, a single definition of disentanglement has not yet been universally accepted. However recently, two research groups have independently come up with three identical criteria for judging the degree of disentanglement in a representation: *modularity, compactness and explicitness* (Eastwood and Williams, 2018; Ridgeway and Mozer, 2018; terminology from Ridgeway and Mozer, 2018). Modularity and compactness describe the mapping between the elements of the representation (or code) $\mathbf{c}$ and the ground-truth generative factors $\mathbf{z}$.

A *modular* representation requires each element of $\mathbf{c}$ to describe a single generative factor in $\mathbf{z}$ (leaving open the possibility of a larger piece of code describing a single generative factor).

Analogously, a *compact* representation requires each element of $\mathbf{z}$ to to be described by a single dimension in the code $\mathbf{c}$ (with the option of a single element of the code capturing more generative

[1] Pre-training is an important phase of current neural models, but the pre-training task is usually connected closely to the task at hand.

factors). A representation both modular and compact would have a one-to-one mapping between the elements of $\mathbf{c}$ and $\mathbf{z}$.

Explicitness relates to how much information about $\mathbf{z}$ is contained in $\mathbf{c}$ and can be formalized as an error a regressor would make when predicting the values of $\mathbf{z}$ from $\mathbf{c}$. The type of the regressor also plays the role: the simpler the regressor is (e.g. linear vs non-linear), the more informative a representation is deemed to be.

In general, modularity is required in all the reviewed papers trying to formalize or measure the notion of disentanglement (Eastwood and Williams, 2018; Ridgeway and Mozer, 2018; Higgins et al., 2017a; Kumar et al., 2018, . . .). Compactness is more problematic: some factors might be more natural to represent by multiple code dimensions (rotation) and moreover, a neural network discovering a redundant code can be easier to optimize (Ridgeway and Mozer, 2018). The degree of explicitness is dependent on the subsequent processing of the code. Ridgeway and Mozer (2018) argue for linear separability, Eastwood and Williams (2018) use a non-linear regressor.

Trying to formalize the notion of the underlying generative factors, Higgins et al. (2018) approaches the definition from the angle of group theory. The core observation is that the state of the world undergoes transformations whose effects are very localized, leaving the majority of the state unchanged (Bengio et al., 2013). A group containing such transformations (symmetries) can be decomposed into subgroups and a vector representation is disentangled if it can be decomposed into independent subspaces each reflecting actions of a separate subgroup. We believe the main benefit of this approach is that it tries to facilitate thinking about the underlying factors of variation in non-trivial scenarios (e.g. 3D rotation should be encoded in a single subspace and not disentangled Higgins et al., 2018)

2.4 Disentanglement in NLP

In the context of NLP, disentanglement is sometimes understood in a more restricted sense (Lample et al., 2019). Rather than discovering unspecified sources of variation in the data, researchers aim at disentangling some pre-defined set of features from the representation. In style transfer, the representation of a piece of text is desired not to include specific features corresponding to the style of text. The generator processing such *content-only* representation then takes the value of the attributes as additional input and produces a text that contains roughly the same information, but presented in a different way (e.g. positive vs. negative review) (Fu et al., 2017; Shen et al., 2017; Hu et al., 2017). Interpreting the language of an utterance as its *style*, one can similarly arrive at a method of unsupervised machine translation (Lample et al., 2018).

The same approach can be observed in other domains: Lample et al. (2017) disentangle some attributes (has glasses, male/female, . . .) in a labelled dataset of faces and by controlling the attributes modify existing images; Wang et al. (2019) use the approach to disentangle the speaker's identity from the content of the utterance to get better speaker representation.

All these methods can be classified as unsupervised from the point of view of the target task (e.g. the unsupervised machine translation system of Lample et al. (2018) is not trained on parallel sentences). At the same time, they are supervised from the point of view of disentanglement, as the values of the attributes in question are used during training of the models.

2.5 Evaluation

The leading approaches to unsupervised disentangled representation learning (Section 3) have been developed in the image domain by using generative models. In such cases, one possible approach to evaluation is visual inspection of the effects of manipulating different latent variables. One can encode an image into a code $\mathbf{c}$ and generate images from the code by gradually changing its value at a given position c_i. A successful model generates a sequence of images differing in a single interpretable attribute, such as rotation or width of an object (Chen et al., 2016; Higgins et al., 2017a). With rising interest in the topic the need to compare different methods in a more principled way arose.

Typically, disentanglement is evaluated along (one or more of) the three axes introduced in Section 2.3: *modularity, compactness and expliciteness*. However, to do that one needs access to the ground-truth factors of variation. For processing images, there are some datasets of choice: either simple pictures of objects artificially generated according to some generative factors: 3D Chairs (Aubry et al., 2014), 3D Faces (Paysan et al., 2009); or pictures manually annotated with a given set of attributes: celebA (Liu et al., 2015)

Couple of metrics have been proposed. Higgins et al. (2017a) use a linear classifier to predict the index of a single generative factor that has been fixed when generating a batch of samples. This is slightly modified in (Kim and Mnih, 2018). Two metrics have been designed based on pointwise mutual information between pairs of code dimensions and factors: Mutual Information Gap (Chen et al., 2018) and the Modularity metric (Ridgeway and Mozer, 2018). Further metrics include the disentanglement score by Eastwood and Williams (2018) and the SAP score by Kumar et al. (2018). Locatello et al. (2019) show the correlation between these metrics on various datasets.

A less direct way of evaluation is also possible. One of the purported benefits of disentanglement discussed earlier (Section 2.2) can be measured with and without enforcing disentanglement: Eastwood and Williams (2018) evaluate sample-efficiency of disentangled representations on a visual abstract reasoning task, Higgins et al. (2017b) focus on zero-shot domain transfer in the reinforcement learning context.

2.6 Connections to other areas of research

Disentangled representation learning is related to other, better known, areas of research. Some have already been mentioned when discussing potential benefits of isolating sources of variation in the data (Section 2.2). Here, we try to extend the list and clarify potential confusion arising from where there is too much overlap.

Transfer learning and related concepts of **zero-/one-/few-shot learning** describe the ability of models to learn in a new setting (data distribution) while reusing information learned in the previous one. When the task remains the same and only the source data distribution changes, one can speak about **domain adaptation**. **Multi-task learning** corresponds to training models under different objectives simultaneously. In all these cases disentanglement is believed to be beneficial, but most current approaches do not explicitly enforce it on their representations.

Causal inference tries to discover the underlying directed acyclic graph representing causation between variables. In most cases, these variables are presupposed, i.e. disentangled a priori. Dealing with real-world perceptual data (images, text), one needs to discover the latent variables first. Recently, Bengio et al. (2020) try to infer causality and learn disentangled representations jointly, albeit in a very restricted settings of two latent and two observational variables, the later being generated by a rotation of the former.

A truly disentangled representation is supposed to be interpretable (as it would correspond to the ground-truth generative factors), but the **interpretability** literature often works with entangled vectors, e.g. self-explaining models (Ribeiro et al., 2016). Moreover, getting fully disentangled codes for real-world data seems not realistic. Even if it was, the process of getting such a representation might still be obscured. Thus, we see a place for both lines of research.

Nonlinear ICA (independence component analysis) tries to find independent sources of variation. This is not required from disentangled representations, although in practice independence is often assumed for practical reasons (Section 3).

In the context of neural networks, **sparsity** is mostly thought of as a means of data compression (Gale et al., 2019) or (potentially) computational efficiency (Zhou et al., 2019). However, Bengio (2017) proposes sparse representations for representing higher-level (language-related) concepts. The interplay between disentanglement and sparsity is also to be expected in **continual learning**, where the space for new information encountered later needs to be preserved (Achille et al., 2018).

3 Methods

There are two main competing approaches of unsupervised learning of disentangled representations. Each of them is based on a different method of generative modelling with neural networks. These are discussed first (from Section 3.1 up to 3.4). A third, unrelated method uses meta-learning approach and is discussed in Section 3.5. Finally, we provide a review of empirical results achieved with disentangled representations (Section 3.6) and discuss the application of generative models in the domain of NLP (Section 3.7).

3.1 The marginal likelihood problem

Given the generative process described by Equation 1, we can express the desired latent representation $\mathbf{z}$ using Bayes' theorem:

$$p_\theta(\mathbf{z}|\mathbf{x}) = \frac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})}{p_\theta(\mathbf{x})} \tag{2}$$

Unfortunately, trying to optimize the model parameters θ by maximizing the likelihood of training data is not practical since finding the $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{x})d\mathbf{z}$ is usually intractable.

3.2 GANs

One way of dealing with the problem of intractable marginal likelihood is using implicit models that avoid specifying probabilities. A popular example are Generative adversial networks (GANs) introduced by Goodfellow et al. (2014).

GAN's main component is a deterministic function G_{θ_G} (generator) that transforms a sample drawn from a specified prior distribution $\mathbf{z} \sim p(\mathbf{z})$ to the data space. The goal is to train the parameters of the generator θ_G so that sampling $\mathbf{z}$ and transforming it by $G_{\theta_G}(\mathbf{z})$ produces examples matching the true data distribution $p(\mathbf{x})$. To do that an adversial discriminator D_{θ_D} is introduced, whose aim is to distinguish between the true data samples and the outputs of the generator. Both G and D are implemented as multi-layer perceptrons and trained by playing a minimax game:

$$\begin{aligned}\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[\log D(\mathbf{x})] \\ + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))].\end{aligned} \tag{3}$$

$D_{\theta_D}(\mathbf{x})$ represents the probability of $\mathbf{x}$ coming from the data distribution.

3.2.1 Information Maximizing GANs

Chen et al. (2016) follow with separating a subspace of the latent variable vector $\mathbf{z}$ and designating it as a latent code $\mathbf{c}$. The purpose of the code is to capture the disentangled attributes of the data, while z keeps its function of '*incompressible noise*' (Chen et al., 2016, p. 3).

Thus the generator is now a function of two vectors: $G_{\theta_G}(\mathbf{z}, \mathbf{c})$. To prevent the generator from ignoring $\mathbf{c}$, it is proposed to regularize the minimax game (Eq. 3) by the mutual information between the code and the generator output:

$$\min_G \max_D V_I(D,G) = V(D,G) + \lambda I(\mathbf{c}; G(\mathbf{z},\mathbf{c})). \tag{4}$$

The mutual information can be expressed as

$$I(c;x) = H(x) - H(c|x). \tag{5}$$

Intuitively, keeping I high during training forces the generator to incorporate $\mathbf{c}$ into the output, so that seeing the output gives us information about the code ($H(c|x)$ is low).

Unfortunately, having an implicit model means no access to explicit probabilities that would enable the calculation of $I(\mathbf{c}; G(\mathbf{z},\mathbf{c})$. However, a lower bound on I can be derived by approximating the true posterior $p(c|x)$ by another distribution $Q(c|x)$. Then it can be shown that:

$$\begin{aligned}L_I(G,Q) &= \mathbb{E}_{\mathbf{x}\sim G, \mathbf{c}\sim P(\mathbf{c})}[\log Q(\mathbf{c}|\mathbf{x}) + H(\mathbf{c})] \\ &\leq I(\mathbf{c}; G(\mathbf{z},\mathbf{c})).\end{aligned} \tag{6}$$

The approximating distribution is parametrized as another neural network, but it can share some layers with the discriminator reducing additional computational cost compared to the original GANs. Apart from providing the means of keeping mutual information between the generator output and the latent code, it enables inference, i.e. getting the disentangled representations of actual data.

3.3 Regularized variational auto-encoders

Unlike plain GANs, variational auto-encoders (VAEs) (Kingma and Welling, 2014) resolve the problem of intractable marginal likelihood by variational approximation to the true posterior: $q_\phi(z|x) \sim p_\theta(z|x)$. The marginal likelihood of the data can then be expressed as:

$$\log p_\theta(\mathbf{x}) = D_{KL}(q_\phi(z|x)||p_\theta(\mathbf{z}|\mathbf{x})) + \mathcal{L}(\theta, \phi; x). \tag{7}$$

As Kullback-Leibler divergence must be non-negative, $\mathcal{L}(\theta, \phi; x)$ is the lower bound on the evidence, i.e. $\log p_\theta(\mathbf{x})$, (marginal likelihood). Evidence lower bound (ELBO) corresponds to:

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[-\log q_\phi(\mathbf{z}|\mathbf{x}) + \log p_\theta(\mathbf{x}, \mathbf{z})] \tag{8}$$

and since $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$, it can be rewritten as:

$$\begin{aligned}\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}))] \\ - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})).\end{aligned} \tag{9}$$

Both conditional distributions $q_\phi(\mathbf{z}|\mathbf{x})$ and $p_\theta(\mathbf{x}|\mathbf{z})$ can be parametrized by neural networks, producing an auto-encoder architecture. The first term in Equation 9 corresponds to the reconstruction error while the second term is a regularizer keeping the posterior distribution close to the prior $p_\theta(\mathbf{z})$. Moreover, under certain assumptions, such as the prior and posterior being Gaussian, the regularizer term can be integrated analytically, leaving only the estimation of the reconstruction error for sampling.

Unlike in the traditional auto-encoders, the VAE encoder specifies the whole multivariate posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$, not just point estimates. This is done by predicting the distribution parameters, such as the mean and variance of individual Gaussian latent variables.

During training, the forward pass includes sampling a representation from a parametrized distribution. In order for the gradient to flow back through such a step into the encoder, $\mathbf{z}$ needs to be reparametrized by a differentiable transformation, such as the location-scale transformation:

$$z \sim \mathcal{N}(\mu, \sigma^2); \quad z = \mu + \sigma\epsilon; \quad \epsilon \sim \mathcal{N}(0,1)$$

3.3.1 β-VAE

VAEs lend themselves for disentangled representation learning particularly well due to the regularization term, which prevents the posterior from diverging from the prior. As the prior tends to be a factorized distribution ($\mathcal{N}(\mu, I)$) a push for disentanglement arises automatically. Higgins et al. (2017a) introduce a simple extension to VAEs by adding a hyperparameter β to control the strength of the regularization. The new training objective is thus a simple modification of the evidence lower bound (Eq. 9):

$$\begin{aligned}\mathcal{L}(\theta, \phi; x, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}))] \\ - \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})).\end{aligned} \tag{10}$$

By setting $\beta = 1$, we get the original VAE formulation.

Higgins et al. (2017a) demonstrate this simple change of increasing the strength of regularization has positive effect on the number of recovered features and quality of disentanglement. Interestingly, β-VAEs seem to be able to learn the number of disentagled features as it was shown to keep some latent variables close to the prior distribution (measured by D_{KL}).

A follow-up work (Burgess et al., 2017) proposes to gradually anneal the strength of the regularization. At the beginning of training, the regularization is strong, which forces the model to focus on the most important generative factors. With gradual lowering of the regularization strength, the model can focus more on faithful reconstruction of the input. The authors conclude this approach leads to better reconstructions while keeping the representations disentangled.

3.3.2 Factor-VAE

Kim and Mnih (2018) note that the regularizer in β-VAEs involves penalizing $I(x; z)$, i.e. mutual information between the learned code and true data. Increasing β then leads to poor reconstruction performance. The proposed solution is to penalize total correlation:

$$TC = D_{KL}(q(\mathbf{z})||\bar{q}(\mathbf{z})), \tag{11}$$

where $\bar{q}(\mathbf{z}) = \prod_{i=1}^{d} q(z_i)$. To estimate the term, a discriminator is trained to recognize between samples from $q(z)$ and $\bar{q}(z)$.

3.3.3 β-TCVAE

Similar conclusion regarding the original β-VAE regularization, mutual information and total correlation was made by Chen et al. (2018). They decompose the D_{KL} term from Equation 10 into three terms and found enforcing TC (Eq. 11) gives the best results. While the objective is the same as in (Kim and Mnih, 2018), here, the TC is estimated by a Monte Carlo estimate with importance sampling. No additional model is required.

3.3.4 DIP-VAE

As yet another alternative to β-VAE, (Kumar et al., 2018) propose regularizing the covariance between $q(z)$ and $p(z)$ which under certain assumptions ($p(\mathbf{z}) \sim \mathcal{N}(0, I)$) can be decomposed into $\lambda_{od} \sum_{i \neq j} [\text{Cov}_{q_\phi(\mathbf{z})}[\mathbf{z}]]_{ij}^2 + \lambda_d \sum_i ([\text{Cov}_{q_\phi(\mathbf{z})}[\mathbf{z}]]_{ij} - 1)^2$ with λ_d and λ_{od} being two hyper-parameters controlling the strength of the diagonal and off-diagonal entries, respectively.

3.4 VAEs vs GANs

Compared to VAEs, approaches using GANs tend to be more difficult to optimize as the performance of both the generator and discriminator must be kept synchronized during training. In the context of disentangled representation learning, IngoGAN has been shown to perform worse than VAE-based methods (Higgins et al., 2017a; Chen et al., 2018). There are modifications to the standard GAN framework (Arjovsky et al., 2017) that aim to alleviate some of the instability problems, but even these have been shown to perform worse than regularized VAEs for disentanglement (Kim and Mnih, 2018).

On the other hand, since GANs do not need to pass the gradient through the latent variables, there is more room for putting assumptions on the prior distribution $p(\mathbf{z})$ such as a random variable being discrete. Yet, while it was initially claimed that InfoGAN can encode the identity of a digit in its discrete variable (Chen et al., 2016), Kim and Mnih (2018) struggled to replicate the results.

VAE-based methods learn continuous latent variables, although these can often encode binary attributes (has glasses, male/female, ...) (Higgins et al., 2017a; Lample et al., 2017). On the other hand, they easily allow for discrete observed variables as there is no need for the gradient to flow through them (unlike in the generator–discriminator interface in GANs).

By using the encoder to parametrize the approximation to the posterior, VAEs create the "amortized gap" between the true posterior and the approximation (Cremer et al., 2018). This is an additional source of error that compounds with the "approximation gap." Approximation gap appears when the considered parametric family of distributions does not include the true distribution. Amortization gap arises due to the inability to find the best distribution in the family. This is caused by the fact that the parameters are not optimised for each training example separately as was the case in stochastic variational inference (Hoffman et al., 2013). Instead, by using the encoder, the parameters and the cost of learning are *amortized* over the training data.

The amortization gap can be the main source of error especially if the data are complex (Cremer et al., 2018) which motivates the research for a 'middle ground' between the stochastic and amortized variational inference. Kim et al. (2018) perform encoding by the global encoder, followed by local refinement of the parameters.

However, the above dilemma is not unique for VAEs. Once, there is a need for inference using GANs, the posterior must also be approximated as discussed in Section 3.2.1.

Currently it seems the focus in the literature on disentangled representations favours VAE-based approaches, but no specific form of regularization has emerged as the method of choice yet. While some disadvantages of β-VAE have been shown, its main appeal is the ease of implementation, an aspect missing from the other models.

3.5 Disentangling by meta-learning

We also note the recent work on causality by Bengio et al. (2020). They infer the causal direction between two latent variables ($A \to B$, $B \to A$) together with a way of disentangling them from observed variables. The encoder and the parameter representing the belief about the causal structure ($P(A \to B)$) are updated in the 'outer loop' of a meta-learning objective. The parameters of the decomposed joint distributions under the two hypotheses are learned in the 'inner loop' ($P_{A\to B}(A)P_{A\to B}(B|A)$; $P_{B\to A}(B)P_{B\to A}(A|B)$).

While there is no special disentanglement-enforcing term in the cost function of the model, the inductive bias lies in the assumption of frequent sparse changes of the ground-truth distribution. The main insight is that knowing the correct causal model together with disentangled representations of the observed data would enable faster learning when there is a local change in the underlying distribution. (A follow-up work (Ke et al., 2019) modifies this approach for more complicated scenarios, but it works with the ground-truth variables directly.)

In a similar vein to Bengio et al. (2020), Javed and White (2019) propose to split a neural model into two parts: representation learning network (RLN$_\theta$) and prediction learning network (PLN$_W$) The output of the model is then given by $p(y|x) =$ PLN$_W$(RLN$_\theta(x)$). The two parts are learned separately, each on samples from a changed distribution. Intuitively, one first forms a hypothesis about the way to represent the data and trains the prediction network (i.e. θ is fixed while W is updated). Then the model is tested under new circumstances of a changed data distribution. The hypothesis about the way the data are represented is updated so that the knowledge learned in the previous step can be transferred more efficiently (i.e. W is fixed while θ is updated). These two steps are repeated. The authors demonstrate success of this approach on two continual learning tasks and show that the learned representations are very sparse.

3.6 Demonstrated benefits

The generative models have been evaluated with respect to some of the potential benefits presented in Section 2.2.

All the considered variants were developed using simple image datasets so their **interpretability** could be presented by various 'latent traversals', i.e. providing a sequence of images generated by modifying a single latent variable. Various attributes of the images have been disentangled in the learned representations, such as: object position, identity, color, presence of sunglasses, baldness, gender, etc. (Chen et al., 2016; Higgins et al., 2017a; Kim and Mnih, 2018; Chen et al., 2018; Kumar et al., 2018).

The ability to **transfer** the learned knowledge into a different scenario has also been demonstrated. van Steenkiste et al. (2019) trained a specialized model for an abstract visual reasoning task. Here, the goal is to complete a sequence of 3 pictures based on the pattern presented in two other completed sequences, e.g. color of the object and its position stays the same. The picture representation was pre-trained without supervision. The results show that during early stages of learning, the model benefits from representations having high disentanglement scores. (No type of regularization would consistently outperform the others).

Higgins et al. (2017b) propose the DARLA (Disentangled representation learning agent) model. It first learns, without supervision, to produce disentangled representations of its environment (β-VAE). This part of the model then remains fixed. In the second stage, it learns to act (various reinforcement learning algorithms are tested) in specific settings (picking certain objects in a room). Finally, its performance was evaluated in an unseen combination of factors (new room-objects combination). The authors report significant benefit of using disentanglement as measured by the agent's score under new conditions.

3.7 Generative models in NLP

In the context of Natural Language Processing, both basic types of deep generative models (GANs and VAEs) are used, but usually for reasons different from wanting to achieve unsupervised disentanglement. (A list of literature on disentangling hand-picked features from an entangled representation was given in section 2.4)

Each approach has its own specific challenges when applying it to text, but the basic solution to the issue of processing sequential data is the same, i.e. recurrent encoder/decoder.

A sentence (the basic unit in the majority of the literature) is encoded into a vector representation by a recurrent cell of choice. The cell processes the input word by word transforming its preceding state and the current word embedding into a new state (and output). When the cell reaches the end of the sentence, the corresponding hidden state is taken to be the desired representation of the whole.

During decoding, a recurrent cell sequentially generates output conditioned on its current state and the previously produced word, until it reaches a terminating symbol. A recurrent decoder can take its initial state from the output of an encoder producing the encoder-decoder architecture, which showed surprising results in neural machine translation (Sutskever et al., 2014; Bahdanau et al., 2015).

3.7.1 GANs

As mentioned earlier, using GANs in a discrete domain raises the problem of propagating the learning signal from the discriminator to the generator.

A common solution is to approximate the one-hot representation of the output with a continuous distribution (Zhang et al., 2016; Kusner and Hernández-Lobato, 2016). For example, the latter authors use the following approximation:

$$\begin{aligned} \mathbf{y} &= \text{one\_hot}(\arg\max_i(h_i + g_i)) \\ &\cong \text{softmax}\left(\frac{1}{\tau(\mathbf{h}+\mathbf{g})}\right), \end{aligned} \tag{12}$$

where each g_i is sampled from the Gumbel distribution, Gumbel$(0, 1)$. During training, the parameter τ is gradually decreased, which corresponds to making the approximation more accurate.

Another solution to the problem of non-differentiability is proposed by Yu et al. (2017). The generator is modelled as a stochastic policy where each action in a given state corresponds to generating a corresponding word. The policy is trained using the REINFORCE algorithm with the output of the (traditional) discriminator used as the reward signal.

3.7.2 VAEs

The interest in VAEs in NLP comes from their coverage of the latent space, which is useful when generating new samples.

Bowman et al. (2016) merge the VAE architecture with recurrent encoder and decoder to generate more coherent sentences. The final hidden state of the encoder is linearly transformed to get the parametrization of the estimated prior. They evaluate the approach on the task of missing words reconstruction and demonstrate the smooth nature of the learned latent space by providing examples of gradually changing sentences. These were generated from linear interpolation between two random points in the latent space.

Training such a model turned out to be problematic as the decoder often learned to ignore the encoded sentence representation. Bowman et al. (2016) point out that the decoder is a recurrent language model (RNNLM) that is known to be capable of learning to assign high probabilities to real sentences. Increasing the ELBO (Eq. 9) can thus be achieved by: 1) training a good language model (high data likelihood under the prior) and 2) encoding sentences so that the representations closely match the prior distribution (low D_{KL} term). This has became to be known as *posterior collapse* and it mirrors the problems of imbalance between the generator and discriminator known from GANs.

Two solutions were proposed by Bowman et al. (2016): annealing the cost of the D_{KL} term during training and dropping out (some) previously generated words. The identical annealing (in a different interval) was presented as an extension to β-VAE (section 3.3.1). The second solution effectively cripples the decoder by corrupting its expected input, which forces the model to utilize the latent variables more.

The posterior collapse became the target of much follow-up research. Yang et al. (2017) control the contextual capacity of the decoder by dilated convolutions and show first success of VAEs for language modelling. Qian and Cheung (2019) add a mutual information between the data and their latent representation (cf. Factor-VAE 3.3.2, β-TCVAE 3.3.3) into the objective and estimate it by a neural network and Kim et al. (2018) show combining the amortized and stochastic information is also helpful.

4 Future Work

We perceive a gap in the literature on the topic of unsupervised learning of disentangled feature representation of language data, which we speculate can be caused by some of the following factors.

First, the goal might seem too ambitious considering the problem is far from being solved even on artificially generated image datasets. The existing approaches have been found useful (cf. Section 3.6), but they have also been shown to suffer from high variance and the disentanglement is never perfect. Moreover, some negative results on the downstream tasks have also been reported (Locatello et al., 2019).

Secondly, it is harder to think about what the underlying factors of variation actually are. In other words, we do not know what we are looking for. Using a 3d engine to render pictures of gradually rotating objects under different lightning conditions, one has a clear idea what to expect from a disentangled representation. However with language, this is harder to achieve.

A similar point is the difficulty of evaluation. It seems to be easier to get intuition about the learned representations by seeing images generated from a latent space traversal than by reading a sequence of generated sentences.

While acknowledging the above objections, we believe they may be addressed and that investigating language representation disentanglement has its own merits. As discussed earlier, successful disentanglement has been observed on the celebA dataset of human faces, with some of the discovered features quite abstract (gender, baldness, glasses, . . .). The problems with evaluation can be partially avoided by focusing on extrinsic evaluation (such as the speed of transfer). We also note that the need to solve specific language-related problems can lead to ideas applicable outside the original domain (cf. attention Bahdanau et al., 2015)

To make the proposed ideas more concrete two more or less arbitrary choices have been made. The first is the actual type of unit that should be represented. While the final ambition is to move beyond, we decided to start with words. This bottom-up approach seems natural, especially since there have not yet been relevant results on this ground level.

As mentioned above, comparing the speed of transfer of different representations can serve as a proxy for evaluating the level of their disentanglement. However, the choice of the transfer task actually matters as it should be related to the type of features that can be expected to be isolated in the representations. We believe the language model objective and part of speech tagging constitute a suitable pair for training and transfer evaluation.

Below, we propose three concrete experiments targeting unsupervised disentanglement of language representations.[2] We provide the working

[2] These are not intended as a comprehensive plan of the rest of the doctoral studies. Rather, they are to be understood

title, a summary in the form of a hypothesis or a research question and a short description.

4.1 Evaluating existing methods in the new settings

> *VAE-based disentangled word embeddings are beneficial for part of speech tagging*

Contextual word embeddings will be learned by using the traditional BERT-like masked language model objective (Devlin et al., 2019) both with and without disentanglement. Disentanglement will be enforced using a VAE-like approach, i.e. the output of the encoder will parametrize the approximation of the posterior $p(z|x)$ and will be pressured to match a factorised prior. The two types of representations will be used for training POS tagging models whose learning curves will be compared for evaluation. We expect the model with access to potentially disentangled features to learn faster.

Using the transformer architecture is assumed but not necessary. (Transformer as a VAE has recently been used by Wang and Wan (2019)). Using a recurrent encoder and decoder with LM objective is also possible. As the language model score is not of interest, we also point out the option of a non-autoregressive decoder (Gu et al., 2018) to encourage more information to be passed through the encoder.

4.2 Classification objective for disentanglement

> *Disentangled word embeddings can be learned by classification*

Analogously to VAEs, we propose to train an encoder parametrizing a distribution over the latent variables, but this time, followed by a classifier predicting part of speech for each word. The cross-entropy will still be regularized by the divergence of the learned latent distribution from a factorized Gaussian, as is the case in VAEs. The intuition behind this experiment is that the classification task could produce more 'focused' disentangled representations capturing e.g. grammatical categories. Moreover, by limiting the capacity of the classifier (e.g. linear), we hope to get representations that would lend themselves to manual inspection. (They should contain disentangled features whose linear transformation predicts parts of speech). Nevertheless, the representations could still be evaluated on a transfer task, such as training the masked language model.

4.3 Meta-learning disentanglement

> *Word embeddings can be disentangled by learning from sentences with gradually changing topic and complexity*

The idea of using non-stationarity as a training signal was described in Section 3.5. Firstly, it must be decided along which axes the data distribution should change. We propose the complexity of the text and/or its topic. The first can be approximated by number of words, the second by a topic model, such as LDA (M. Blei et al., 2003).

We would like to train an encoder/decoder architecture with a language modelling objective by repeating the following steps: 1) For a given state of the encoder, update the decoder by training it on samples from a particular distribution. 2) For the given decoder, update the encoder parameters by training it on samples from a changed distribution.

Evaluation can happen by a transfer to a new task such as POS tagging, or by a transfer to a new distribution (domain adaptation). The main question is again whether the proposed method of learning disentangled representations produces better contextual word embeddings than the baseline. *Better* stands for 'enabling faster transfer' and the *baseline* here stands for embeddings learned on the same data points but seen as a shuffled collection simulating i.i.d. samples.

5 Conclusion

Above, we have proposed and motivated a specific area of research within NLP, namely unsupervised learning of disentangled representations. We described disentanglement as a general concept of representation learning which is related to many other lines of research. Specific methods used to approach the problem were reviewed as well as their application in NLP in particular. We found the motivation for the use of the methods in NLP is usually different from wanting to achieve disentanglement. Alternatively, the disentanglement is understood in a limited sense of supervised isolation of preselected features. Therefore, we argue for filling in the gap and suggest three concrete experiments that can be used to start doing so.

as probes into the topic, based on which further directions will be considered. We believe this approach is justified by the lack of relevant literature.

References

Alessandro Achille, Tom Eccles, Loic Matthey, Chris Burgess, Nicholas Watters, Alexander Lerchner, and Irina Higgins. 2018. Life-long disentangled representation learning with cross-domain latent homologies. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9873–9883. Curran Associates, Inc.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia. PMLR.

M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic. 2014. Seeing 3d chairs: Exemplar part-based 2d-3d alignment using a large dataset of cad models. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3769.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

H.B. Barlow, T.P. Kaushal, and G.J. Mitchison. 1989. Finding minimum entropy codes. *Neural Computation*, 1(3):412–423.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Yoshua Bengio. 2017. The consciousness prior. *CoRR*, abs/1709.08568.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. 2020. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations*.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Martin Buehler, Karl Iagnemma, and Sanjiv Singh, editors. 2009. *The DARPA Urban Challenge: Autonomous Vehicles in City Traffic, George Air Force Base, Victorville, California, USA*, volume 56 of *Springer Tracts in Advanced Robotics*. Springer.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. 2017. Understanding disentangling in -vae.

Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. 2018. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2610–2620. Curran Associates, Inc.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2172–2180. Curran Associates, Inc.

Chris Cremer, Xuechen Li, and David Duvenaud. 2018. Inference suboptimality in variational autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1078–1086, Stockholmsmässan, Stockholm Sweden. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Cian Eastwood and Christopher K. I. Williams. 2018. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128 – 135.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2017. Style transfer in text: Exploration and evaluation. *CoRR*, abs/1711.06861.

Trevor Gale, Erich Elsen, and Sara Hooker. 2019. The state of sparsity in deep neural networks. *CoRR*, abs/1902.09574.

Marta Garnelo, Kai Arulkumaran, and Murray Shanahan. 2016. Towards deep symbolic reinforcement learning. *CoRR*, abs/1609.05518.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. 2016. Continuous deep q-learning with model-based acceleration. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2829–2838, New York, New York, USA. PMLR.

K. He, X. Zhang, S. Ren, and J. Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.

Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo J. Rezende, and Alexander Lerchner. 2018. Towards a definition of disentangled representations. *CoRR*, abs/1812.02230.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017a. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. 2017b. DARLA: Improving zero-shot transfer in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1480–1490, International Convention Centre, Sydney, Australia. PMLR.

Matthew D. Hoffman, David M. Blei, Chong Wang, John Paisley, and Tommi Jaakkola. 2013. Stochastic variational inference. *Journal of Machine Learning Research*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596, International Convention Centre, Sydney, Australia. PMLR.

Khurram Javed and Martha White. 2019. Meta-learning representations for continual learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 1820–1830. Curran Associates, Inc.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Chris Pal, and Yoshua Bengio. 2019. Learning neural causal models from unknown interventions.

Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658, Stockholmsmässan, Stockholm Sweden. PMLR.

Yoon Kim, Sam Wiseman, Andrew Miller, David Sontag, and Alexander Rush. 2018. Semi-amortized variational autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2678–2687, Stockholmsmässan, Stockholm Sweden. PMLR.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Tejas D Kulkarni, William F. Whitney, Pushmeet Kohli, and Josh Tenenbaum. 2015. Deep convolutional inverse graphics network. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2539–2547. Curran Associates, Inc.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. 2018. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*.

Matt J. Kusner and José Miguel Hernández-Lobato. 2016. GANS for sequences of discrete elements with the gumbel-softmax distribution. *CoRR*, abs/1611.04051.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Fader networks:manipulating images by sliding attributes. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5967–5976. Curran Associates, Inc.

Z. Liu, P. Luo, X. Wang, and X. Tang. 2015. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124, Long Beach, California, USA. PMLR.

David M. Blei, Andrew Y. Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. 2009. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301.

J. Peters, D. Janzing, and B. Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA.

Dong Qian and William K. Cheung. 2019. Enhancing variational autoencoders with mutual information neural estimation for text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. New York, NY, USA. Association for Computing Machinery.

Karl Ridgeway and Michael C Mozer. 2018. Learning deep disentangled embeddings with the f-statistic loss. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 185–194. Curran Associates, Inc.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *CoRR*, abs/1606.04671.

Jürgen Schmidhuber. 1992. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6830–6841. Curran Associates, Inc.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. 2019. Are disentangled representations helpful for abstract visual reasoning? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14245–14258. Curran Associates, Inc.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Shuai Wang, A. Johan Rohdin, Lukáš Burget, Oldřich Plchot, Yanmin Qian, Kai Yu, and Jan Černocký. 2019. On the usage of phonetic information for text-independent speaker embedding extraction. In *Proceedings of Interspeech*, volume 2019, pages 1148–1152. International Speech Communication Association.

Tianming Wang and Xiaojun Wan. 2019. T-cvae: Transformer-based conditioned variational autoencoder for story completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5233–5239. International Joint Conferences on Artificial Intelligence Organization.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3881–3890, International Convention Centre, Sydney, Australia. PMLR.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*.

Yizhe Zhang, Zhe Gan, and Lawrence Carin. 2016. Generating text via adversarial training. In *NIPS Workshop on Adversarial Training*.

X. Zhou, Z. Du, S. Zhang, L. Zhang, H. Lan, S. Liu, L. Li, Q. Guo, T. Chen, and Y. Chen. 2019. Addressing sparsity in deep neural networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 38(10):1858–1871.