Why Bigger is Not Better: The Asymptotic Behavior of VAE Latent Space

Shayan K. Azmoodeh

Department of Computer Science
University of Illinois at Urbana-Champaign
shayana3@illinois.edu

***Abstract*—We study the effect of the dimensionality of the latent space of variational autoencoders (VAE) on the model's ability to effectively reconstruct input data (encoding/decoding quality) as well as generate new images. We provide a source coding interpretation of the VAE, reframing the minimization of the ELBO loss as the creation of a minimal length source code. Under this framework, we prove that in the case where the posterior latent distribution is modelled as a normal distribution with i.i.d. components, increasing the latent dimension results in poor recovery of details in input data when considering the model's reconstruction abilities and provide an explanation for the diminishing quality of generated data. We empirically show that these results hold beyond just the case of i.i.d. latent components.**

I. Introduction

The field of machine learning has seen considerable progress in the area of generative modeling in recent years with advancements in language, image, and video modelling and generation. Despite these recent achievements, variational autoencoders (VAEs) [1] are still widely used for modelling data distributions and generation of synthetic data. A notable advantage of the VAE is the trainable latent space that correlates with the input data distribution, providing a "deep" alternative to standard statistical signal decomposition methods such as principal component analysis or the non-negative matrix factorization. There have been many works expanding upon the original VAE to improve the interpretability of the latent space [2] or maximize the model's utilization of the latents to improve the learned distribution [3]. In this paper we study the effect of the dimensionality of the latent space on the performance of the VAE and the asymptotic behavior as the latent dimension grows infinitely large.

The VAE looks to learn a distribution for the data $p_\theta(\boldsymbol{x})$ parameterized by θ through the use of latent variables $\boldsymbol{z} \sim \mathcal{N}(0, I)$. This also allows for generation of images by sampling $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ Thus, following the method of maximum likelihood estimation (MLE), we would like to find the optimal θ such that $\mathbb{E}_{\boldsymbol{x}\sim p_\mathcal{D}}\left[\log p_\theta(\boldsymbol{x})\right]$ is maximized. Note that this is equivalent to minimizing the entropy of the learned distribution. The latent variables are introduced by marginalizing this likelihood,

$$p_\theta(\boldsymbol{z}) = \int p_\theta(\boldsymbol{z}) p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) \, \mathrm{d}\boldsymbol{z} \,.$$

An *encoder* $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ is introduced, parameterized by ϕ, which we would like to train to be similar to $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$, to enable more efficient Monte Carlo estimation of the integral above; the majority of latent values sampled from the prior will have very small posterior $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ and thus will contribute very little to the final estimate, so sampling from $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ will provide latent values with significant posterior probabilities [1]. However, in many cases the integral above is intractable, thus the following variational lower bound (ELBO) is optimized instead, which also allow joint optimization of θ and ϕ,

$$\log p_\theta(\boldsymbol{x}) \geq \mathcal{L}(\theta, \phi, \boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\big] - D_{KL}\big(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) || p(z)\big), \tag{1}$$

where $p(\boldsymbol{z})$ denotes the probability of $\boldsymbol{z}$ under its prior.

We begin by showing the VAE under the ELBO objective is equivalent to optimizing transmission rates of a source coding with side-information problem in Section II. Section III uses this source-coding framework to mathematically explain the decreasing input reconstruction and image generation abilities of the VAE as latent dimension size increases. Section IV verifies the result of the previous section experimentally.

II. VAE as an Efficient Source Code

We extend the source coding analogy given in [4], formulating the VAE as a lossy source coding problem with side-information problem similar to that of Wyner & Ziv [5]. We suppose the VAE is available globally and use it within the encoder, which we detail below.

The source code operates as follows and is illustrated in figure 1. **Encoder**: Given an input $\boldsymbol{x}$, we compute $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ and sample a $\boldsymbol{z}$ from this posterior on the latent space. We then compute $p(\boldsymbol{z})$ (the probability under the latent prior) and $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ using the sampled $\boldsymbol{z}$ and the initial input $\boldsymbol{x}$. $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ is sent as *side information*. The encoder then then transmits the quantities $p(\boldsymbol{z})$ and the remainder of $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ that is not contained in the available side information $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$. **Decoder**: The decoder uses $p(\boldsymbol{z})$ to pick a $\hat{\boldsymbol{z}}$ that has the same probability (i.e., $p(\boldsymbol{z}) = \hat{\boldsymbol{z}}$). Then $\hat{\boldsymbol{x}}$ is picked such that $p_\theta(\hat{\boldsymbol{x}} \mid \hat{\boldsymbol{z}}) = p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$. In the case where there are multiple possible values are possible to pick, one of these potential values is picked at random.

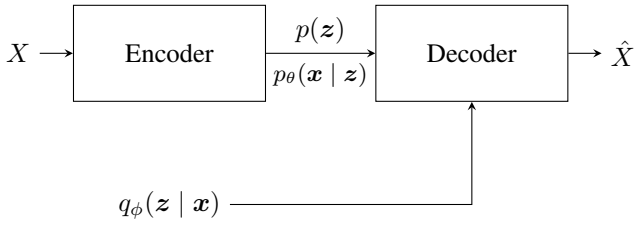

Fig. 1. Block diagram of source coding setup of VAE, equating transmission rate of encoder to the ELBO loss.

Analyzing the the transmission rate of this source code, without the side information the code has a rate as the expected value of $-\log p(\boldsymbol{z}) - \log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$. However, the availability of side information reduces the number of bits needed to be transmitted for $p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$ to give us a rate of,

$$\begin{aligned} R_1 &= \mathbb{E}_{\substack{\boldsymbol{x}\sim p_{\mathcal{D}} \\ \boldsymbol{z}\sim q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}} \Big[-\log p(\boldsymbol{z}) - \log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) + \log q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \Big] \\ &= \mathbb{E}_{\boldsymbol{x}\sim p_{\mathcal{D}}} \big[-\mathcal{L}(\theta, \phi, \boldsymbol{x}) \big] \end{aligned} \tag{2}$$

following the definition of the ELBO in (1). Thus, finding ϕ and θ to get an efficient source code is equivalent to maximizing the ELBO loss. Moreover, this formulation allows studying the VAE independent of the training process, as we can consider the distortion while varying factors (such as the dimensions of the latent space) at a fixed rate (since the rate is analogous to the ELBO loss).

The use of the side information to reduce the number of bits needed for the encoding also gives a more intuitive understanding of how the distributions p_θ and q_ϕ behave as the model reaches its global optimum. Namely, one expects the side information $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ to resemble the posterior $p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ to minimize the number of bits needed to transmit. This is captured by an equivalent formulation of the ELBO loss given in [3],

$$\begin{aligned} \mathcal{L}(\theta, \phi, \boldsymbol{x}) = &-D_{KL}\big(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p(\boldsymbol{z})\big) \\ &- \mathbb{E}_{q_\phi(z)}\Big[D_{KL}\big(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) || p_\theta(\boldsymbol{z} || \boldsymbol{x})\big)\Big]. \end{aligned} \tag{3}$$

III. Mathematical Results

A. Preliminaries

In the following section we assume all random variables are real-valued and operate on the probability space $(\Omega, \mathcal{F}, P)$ and we denote the distribution of a random variable X with density f as $\mu : \mathcal{B} \to \mathbb{R}$, $\mu(A) = \int_A f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$, where $\mathcal{B}$ denotes the Borel σ-algebra.

Using the source coding framework developed in section II, we show that the VAE decoding performance and its ability to generate new data from randomly sampled latent values decreases as the size of the latent space grows infinitely large.

We first define the notion of the typical set for continuous random variables.

Definition 1 (Typical Set)**.** *Let X be a random variable with alphabet $\mathcal{X}$ with density f and let $\epsilon > 0 \in \mathbb{R}$. The* ϵ-typical set *of length n, denoted $A_\epsilon^{(n)} \subseteq \mathcal{X}^n$, is the set,*

$$\left\{ (x_1, x_2, \ldots, x_n) : \left| -\frac{1}{n} \log f(x_1, \ldots, x_n) - h(X) \right| < \epsilon \right\}$$

where $h(X)$ denotes the differential entropy of X.

The following is a well-known result about elements of the typical set and follows immediately from the weak law of large numbers.

Lemma 1 (AEP)**.** *Suppose $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables with density f. Then,*

$$-\frac{1}{n} \log f(X_1, \ldots, X_n) \xrightarrow{n\to\infty} h(X)$$

in probability.

We also make use of the following properties of the typical set which follow from the AEP [6].

Lemma 2 (Properties of Typical Set)**.** *Let X be a random variable with alphabet $\mathcal{X}$ and let $\epsilon > 0$ and $n \in \mathbb{N}$. Then,*

(i) $2^{-n(h(X)+\epsilon)} \le f(x_1, \ldots, x_n) \le 2^{-n(h(X)-\epsilon)}$

(ii) $\mu\big(A_\epsilon^{(n)}\big) > 1 - \epsilon$

(iii) $(1-\epsilon)2^{n(h(x)-\epsilon)} \le \left| A_\epsilon^{(n)} \right| \le 2^{n(h(X)+\epsilon)}$.

We now prove the following theorem, which provides insight into the declining performance of the VAE as the dimension of the latent space increases.

Theorem 3. *Let $Z_1, Z_2 \sim \mathcal{N}(\mu\mathbf{1}, \sigma^2 I) \in \mathbb{R}^n$ be n-dimensional random variables with density f_n, where $\mathbf{1} \in \mathbb{R}^n$ denotes a vector of 1s and $I \in \mathbb{R}^{n\times n}$ is the identity. Then $f_n(Z_1) - f_n(Z_2) \to 0$ in probability as $n \to \infty$.*

Proof. Take $\mu, \sigma \in \mathbb{R}$ and $n \in \mathbb{N}$. Let $Z_1, Z_2 \sim \mathcal{N}(\mu\mathbf{1}, \sigma I) \in \mathbb{R}^n$. Since the covariance is diagonal, each component in Z_i is independent and has distribution $\mathcal{N}(\mu, \sigma)$. Thus the components are Z_1 and Z_2 are i.i.d so we can apply properties of typicality to them.

Let $f = f_1$ be the density of $\mathcal{N}(\mu, \sigma^2)$. Take $\epsilon > 0$. Since Z_1 and Z_2 are independent,

$$\begin{aligned} P\big(Z_1 \in A_\epsilon^{(n)}, Z_2 \in A_\epsilon^{(n)}\big) \\ = P\big(Z_1 \in A_\epsilon^{(n)}\big) P\big(Z_2 \in A_\epsilon^{(n)}\big) > (1-\epsilon)^2, \end{aligned} \tag{4}$$

where the bound follows from lemma 2. Note that when both Z_1 and Z_2 are typical, again by lemma 2 we have ($i = 1, 2$),

$$2^{-n(h(f)+\epsilon)} \le f_n(Z_i) \le 2^{-n(h(f)-\epsilon)}.$$

This implies,

$$\begin{aligned} |f_n(Z_1) - f_n(Z_2)| &\le 2^{-n(h(f)-\epsilon)} - 2^{-n(h(f)+\epsilon)} \\ &= 2^{-n(h(f)-\epsilon)} \left(1 - 2^{-2n\epsilon}\right) \qquad (5) \\ &\xrightarrow{n\to\infty} 0. \qquad (6) \end{aligned}$$

Hence,

$$
\begin{aligned}
&P\left(|f_n(Z_1) - f_n(Z_2)| \leq 2^{-n(h(f)-\epsilon)}\left(1-2^{-2n\epsilon}\right)\right) \\
\geq &P\left(2^{-n(h(f)+\epsilon)} \leq f_n(Z_i) \leq 2^{-n(h(f)-\epsilon)},\ i=1,2\right) \\
\geq &P\Big(Z_1 \in A_\epsilon^{(n)}, Z_2 \in A_\epsilon^{(n)}\Big) \\
> &(1-\epsilon)^2. \qquad (7)
\end{aligned}
$$

Now take $\delta > 0$. By the limit in (6), we can take n sufficiently large such that the bound in (5) is $< \delta$. Therefore,

$$
\begin{aligned}
&\lim_{n\to\infty} P\left(|f_n(Z_1) - f_n(Z_2)| \leq 2^{-n(h(f)-\epsilon)}\left(1-2^{-2n\epsilon}\right)\right) \\
\leq &\lim_{n\to\infty} P\Big(|f_n(Z_1) - f_n(Z_2)| < \delta\Big).
\end{aligned}
$$

This holds for all $\delta > 0$ so we have,

$$
\begin{aligned}
&\lim_{n\to\infty} P\Big(|f_n(Z_1) - f_n(Z_2)| = 0\Big) \\
\geq &\lim_{n\to\infty} P\left(|f_n(Z_1) - f_n(Z_2)| \leq 2^{-n(h(f)-\epsilon)}\left(1-2^{-2n\epsilon}\right)\right) \\
> &(1-\epsilon)^2,
\end{aligned}
$$

where the last inequality follows from (7). Moreover, this bound holds for all $\epsilon > 0$, so we must have,

$$
\lim_{n\to\infty} P\Big(|f_n(Z_1) - f_n(Z_2)| = 0\Big) = 1
$$

and thus $f_n(Z_1) - f_n(Z_2) \to 0$ in probability as $n \to \infty$. □

B. Application to VAEs

Theorem 3 tells us that in the case of a VAE with latent posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ of the form $\mathcal{N}(\mu_\phi(\boldsymbol{x})\mathbf{1}, \sigma_\phi(\boldsymbol{x})I)$ (where μ_ϕ and σ_ϕ are determined by the encoder q_ϕ), the density of sampled values collapses to a single value $\approx 2^{-nh(\mathcal{N}(\mu,\sigma))}$ *as the dimensions of the latent space grows large*. We assume that similar input data (i.e., same object or environment) will give similar posterior distributions on the latent space, and thus their probabilities will concentrate around a single value. Thus, in a dataset with multiple "similarity groups", elements within each group will have similar posterior latent probabilities and thus, in the source coding setup in section II, the decoder will be unable to pick the exact latent value that produced the probability it was given. This results in loss of detail in sampled decoded images as the decoder is only able to select an image from the data distribution that is "similar" to the input image based on the information it has available.

We apply a similar reasoning to sampling the latent space to generate new data. The latent prior has a standard normal distribution and thus, with high dimensionality, by theorem 3, sampled values will have approximately the same density. Hence, again in the source coding setup, given a sample from $p(\boldsymbol{z})$, the decoder is unable to uniquely identify a sample point to associate with the sampled latent. This results in the model either generating low-quality outputs that have only a faint resemblance to the original data or being biased towards outputting data with posterior latent probabilities close to the probability of typical sequences from the prior (which is in fact what we observe).

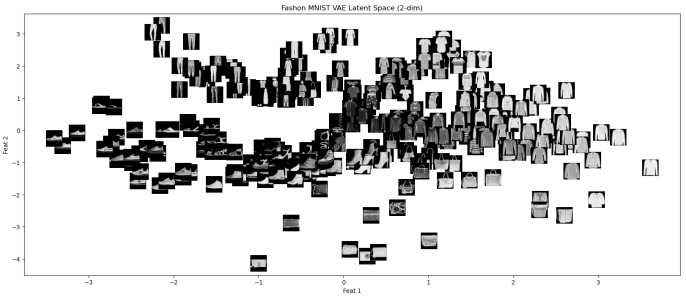

Fig. 2. Visualization of the latent space of the VAE with a latent dimension of 2. Upon inspection we see that similar articles of clothing are grouped near each other and that there is a gradient from clothes of one type (i.e., upper garments) to another (i.e., shoes).

IV. Empirical Results

In this section we provide empirical evidence for the mathematical results presented in section III. We run our experiments using the fashion MNIST dataset due to it's simple pixel representation (the data can be modelled as an 784-dimensional Bernoulli random variable) while still containing sufficient complexity (i.e., details in clothes/shoes) to gauge decoding and image generation quality.

A visualization of the two-dimensional latent space is given in figure 2. This matches our assumption from the previous section that data that are "similar" are likely to be mapped to latent values that are close by. As can be seen from the plot, clothing of the same kind form loose clusters in the latent space and there is a gradient from clothing of one type to another.

We train VAEs with latent dimensions of 2, 20, 200, and 800, where the final one is over-parameterized in that it has a more components than the input data. The decoder and encoder in each model consists of two fully-connected feed-forward layers with 256 neurons. The output of the decoder p_θ is the mean of a Bernoulli distribution and the output of the encoder q_ϕ is $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\sigma^2 \in \mathbb{R}$ for a normal distribution $\mathcal{N}(\boldsymbol{\mu}, \sigma^2 I)$ (where n is the latent dimension). Note here we use a vector mean for the distribution of the latent posterior to allow for better model performance, however we still observe the results discussed in the previous section. All models were trained for 100 epochs with a batch size of 256 and a learning rate of 10^{-4}.

We begin by noting that the probability of the latent values tend towards the entropy of $\mathcal{N}(0, \sigma_\phi(\boldsymbol{x})^2)$, $h = \frac{1}{2}\ln\big(2\pi e\sigma^2\big)$, as the size of the latent space increases (note that the entropy is independent of the mean of the distribution) as shown in figure 3. This is expected from the result of theorem 3.

Furthermore, we observe that, as predicted, the quality of generated images degrades significantly as the latent dimension increases. This can be seen more clearly in the jump from

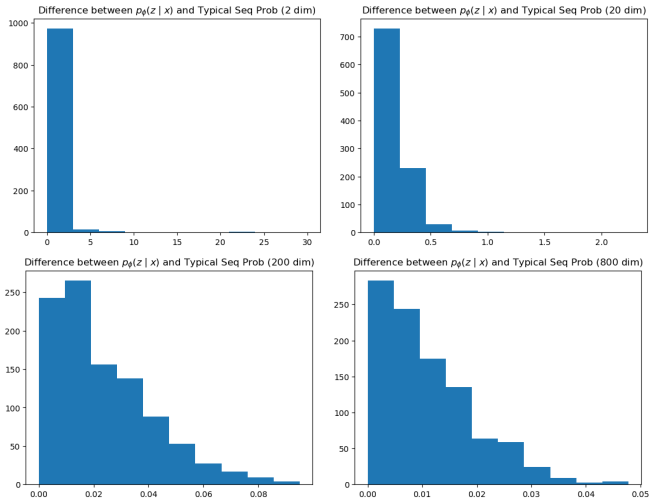

Fig. 3. Difference between density of sampled latent variables from the posterior distribution $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ and the probability of an n-dimensional typical sequence as a percent of the probability. Observe that as the latent dimension increases, the probability of sampled sequences get closer to the probability of a typical sequence.

a latent dimension of 20 to 200. With a larger latent space, the generated images are blurrier and appear to have less variety as hypothesized in the previous section. Examples of generated images from each of the models is given in figure 4.

We also observe that the decoding capabilities of all the models is approximately the same. However, as hypothesized, output images are blurry and only manage to capture the general idea of the input image without capturing any of its specific details. This lack of correlation between the size of the latent space and the decoding quality may be due to a lack of expressiveness in the models and may be resolved with more training. Examples are shown in figure 5.

Fig. 4. Images generated by VAEs with different latent dimension sizes. As the latent dimension increases, the quality and variety of the generated images decreases.

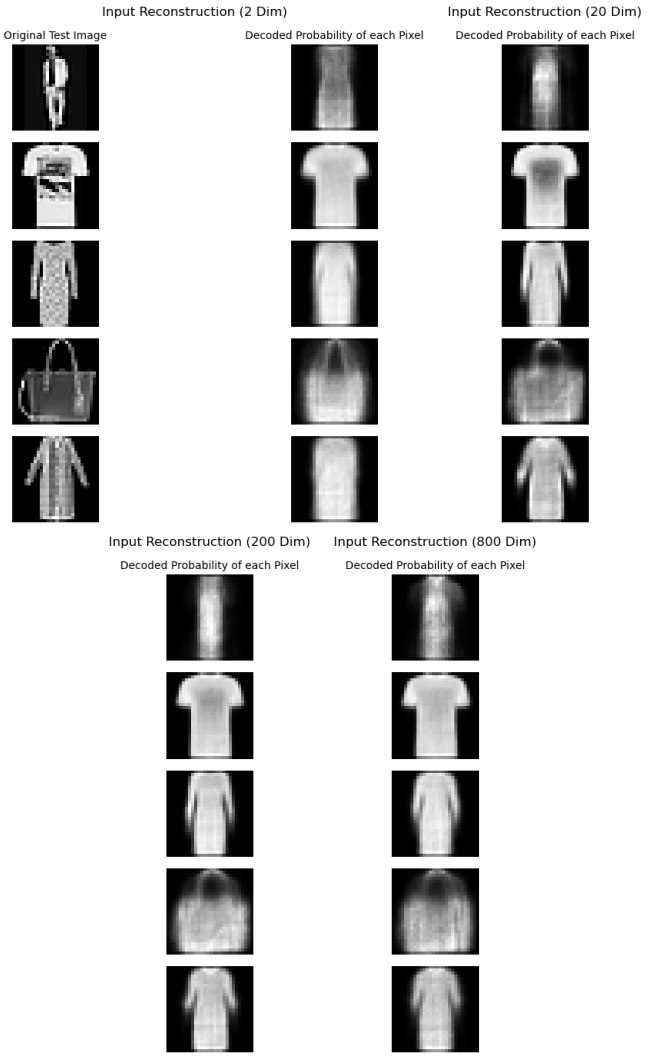

Fig. 5. Reconstruction of input images for VAE models with varying latent dimension size. Quality of reconstruction seems to be similar across all models, however all outputs are slightly blurry or distorted.

REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[2] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework.," *ICLR (Poster)*, vol. 3, 2017.

[3] S. Zhao, J. Song, and S. Ermon, "Infovae: Balancing learning and inference in variational autoencoders," in *Proceedings of the aaai conference on artificial intelligence*, vol. 33, pp. 5885–5892, 2019.

[4] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," *arXiv preprint arXiv:1611.02731*, 2016.

[5] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on information Theory*, vol. 22, no. 1, pp. 1–10, 1976.

[6] T. M. Cover and J. A. Thomas, *Asymptotic Equipartition Property*, ch. 3, pp. 57–69. John Wiley & Sons, Ltd, 2005.