Tumor Growth Prediction as Stochastic Video Prediction

Neil Yeung
University of Rochester
Department of Computer Science
nyeung@u.rochester.edu

Abstract

The task of tumor growth prediction is beneficial for both the treatment and analysis of tumors. Accurately predicting whether a tumor will emerge and what a CT scan will look like given a set of initial observations allows doctors to more accurately diagnose cancer and suggest treatment. Traditional methods of predicting tumor growth involve complex systems of partial differential equations that contain limited parameters that fail to take advantage of group-wide patterns. We propose that tumor growth prediction can be formulated as a video prediction problem. To this end, we implement a set of state-of-the-art stochastic video prediction methods trained on a set of mice leg bone CT scans modified for medical data. Ultimately, we show promising results on a variety of both quantitative and qualitative metrics.

1. Introduction

According to the American Cancer Society, around 3600 people will be diagnosed with primary bone sarcoma, or cancer of the bone and joints, in 2020[1]. Although bone cancer makes up a relatively small 0.2% of cancers, malignant bone tumors are a common location for metastasis, where cancers originating from other places within the body spread onto the bone. Modeling and predicting tumor growth yields benefits to healthcare practitioners.

Traditionally, tumor growth modeling efforts involve systems of ordinary partial differential equations and partial differential equations to model tumor spread [14]. There are a few key issues in this approach. First, these systems utilize a limited set of parameters which have a difficult time generalizing to different types, shapes, and aggressiveness of tumors [6]. Second, these models only look at patient-level patterns; that is, they can only model the dynamics within a specific patient's tumor growth, failing to take into account group-wide patterns [6]. Finally, the models also rely heavily on hand-selected segmentation and features of tissue.

With the advent of deep learning, the shortcomings of traditional mathematical approaches can be addressed. Deep learning is able to utilize a larger set of parameters to model the complex dynamics of tumor growth as well as leverage group-level patterns.

We propose that tumor growth prediction can be formulated as a video prediction problem. The video prediction problem is given a set of T past frames, output K future frames [9]. The literature surrounding the mathematical modeling of tumor growth suggests that stochastic models appear to model tumor growth dynamics more accurately [14]. Furthermore, in the video prediction literature in general, assuming stochasticity leads to clearer and higher quality predictions even in deterministic environments [9, 4]. Thus, we assume that tumor growth is a stochastic process.

To the knowledge of the author, no prior work has treated tumor growth as a stochastic video prediction problem. In summary, our primary contribution is: we conduct experiments on whether state-of-the-art stochastic video prediction methods generalize to sparse sparse medical data and the tumor growth prediction task without the need for explicit segmentation or features.

2. Related Work

Reaction Diffusion Model The vast majority of mathematical models of tumor growth prediction build upon the Fisher-Kolmogorov model for reaction-diffusion processes [14, 5]. It is given by

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} + AC - BC^2 \qquad (1)$$

for $x \in \mathbb{R}$ and where A, B denotes the reaction coefficents and C denotes the concentration of molecules with a diffusion constant D [1]. The stochastic case can be considered by assuming C fluctuates.

Machine learning for Tumor Prediction Tumor prediction that involves machine learning may still rely on explicit

[1] https://www.cancer.org/cancer/bone-cancer/about/key-statistics.html

paramaters or segmentation. For example, Ref. [3] predicts 3D kidney tumors by feeding the meshes of the kidneys into a partial differential equation. However, this method still relies on explicitly defining different diffusion constants for the various types of tumors. Another approach uses convolution long term short memory (ConvLSTMs) for pancreatic tumor prediction [15]. The architecture is a spatio-temporal ConvLSTM that take advantage of the 4D nature of medical data (height, width, depth, and time). The algorithm is also able to incorporate modalities outside of the four dimensions, such as patient history, to better predict patient outcomes.

Deep generative models appear to be a promising avenue of machine learning research for tumor growth prediction; GP-GAN is one example of such an approach [6]. It uses stacked 3D Generative Adversarial Network (3D-GAN) architecture to predict the growth of glioma, or brain tumors. The paper introduces a novel loss function that combines l_1 and *Dice* loss. The data utilized in the paper studies predictions over three time points. That is, it attempts prediction of the frame 2 and 3 based of frame 1, and frame 3 based off frames 1 and 2. However, one issue with the method outlined in GP-GAN is that it still requires expert annotations and segmentation, which may be costly to acquire.

Notably, there does not seem to be machine learning methods for tumor prediction for primary bones sarcoma.

Stochastic Video Prediction When determinism is assumed but the events are at least partially stochastic, predictions can end up blurry as the prediction becomes the average all plausible futures. For video prediction to generalize to data outside of standardized datasets such as [11], assuming that at least some amount of randomness is present proves beneficial. Video prediction is an important computer vision (CV) task as the ability to predict future events is arguably a core part of intelligence.

State of the art stochastic video prediction architectures by far use variational auto-encoders (VAE) most often [9]. The variational auto-encoder does utilize LSTM networks, but still follow the VAE formalism. Indeed, VAE methods are common in state of the art results [4, 7, 12]. Variations of the standard VAE approach include implementing optical flow modifying the standard VAE loss function such as incoporating adverserial loss [7] or geodesic loss [2].

3. Methods

3.1. Variational Auto-encoder

The idea behind the variational auto-encoder is to extend and modify the standard auto-encoder architecture in a few key ways. First, instead of encoding the entire input vector, a latent distribution is encoded instead. The distribution is forced to be continuous and Gaussian through a regularization term. Specifically, the regularization term utilizes the Kullback-Leibler divergence [10] denoted by

$$D_{\mathrm{KL}}(p\|q) = \sum_i p_i \log_2 \frac{p_i}{q_i} \tag{2}$$

where p denotes a probability distribution and q denotes a "target" distribution. In practice, q is a standard Gaussian distribution. During generation time, we sample from the latent distribution and pass it through the decoder to generate the output.

SVG with LP Ref. [4] uses the VAE formalism. The finding is that the Gaussian prior, or the fixed prior, can be replaced by a learned prior primed on the data. This provides better results on the task of stochastic video prediction as it avoids the issue of randomly sampling from the fixes Gaussian distribution thereby ignoring temporal dependencies. The model is trained based on optimizing for the evidence lower bound (ELBO)

$$\begin{aligned}\mathcal{L}_{\theta,\phi,\psi}\left(\mathbf{x}_{1:T}\right) = \sum_{t=1}^{T} &\left[\mathbb{E}_{q_\phi(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})} \log p_\theta\left(\mathbf{x}_t \mid \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t}\right)\right. \\ &\left.-\beta D_{KL}\left(q_\phi\left(\mathbf{z}_t \mid \mathbf{x}_{1:t}\right) \| p_\psi\left(\mathbf{z}_t \mid \mathbf{x}_{1:t-1}\right)\right)\right]\end{aligned}$$

where the learned prior is given by $p_\psi\left(\mathbf{z}_t \mid \mathbf{x}_{1:t-1}\right)$. During testing, the frame at timestep t is generated by sampling the latent vector z_t from the learned prior and then feeding it into a prediction network with a long short term memory (LSTM) architecture. The recurrence for the generation of a frame is given by

$$\begin{aligned}\mu_\phi(t), \sigma_\phi(t) &= LSTM_\phi\left(h_t(Enc(\mathbf{x}_t)\right) \\ \mathbf{z}_t &\sim \mathcal{N}\left(\mu_\phi(t), \sigma_\phi(t)\right) \\ g_t &= L\,\mathrm{STM}_\theta\left(h_{t-1}, \mathbf{z}_t\right) \quad h_{t-1} = \mathrm{Enc}\left(\mathbf{x}_{t-1}\right) \\ \mu_\theta(t) &= \mathrm{Dec}\left(g_t\right)\end{aligned}$$

In this way, a frame at time t is generated based off of the frames from 1 to $t-1$ [2].

SVG' Ref [12] investigates whether minimizing inductive bias without a specialized architecture can yield state of the art video prediction results. The paper shows that scaling a modified SVG architecture without any specialized computations (e.g. adverserial loss [7]), one can still obtain state of the art video prediction results. It optimizes the same variational lower bound as SVG.

3D GAN An alternative to the VAE formalism with LSTM is to utilize a 3D-GAN and encode temporal information as the third dimension [13]. In essence, the data is treated as a cube with height, width, and time as the three dimensions.

[2] For more details, please see [4]

4. Experiments

4.1. Dataset

The main data set consists of CT image scans of the leg bones of mice. Each data point is a stack of images of dimension 400 x 400 x 300 where 300 is the number of slices in the region of interest. The data is processed by SITK [8] which registers the CT scan and incorporates spatial information rather than channels. Thus, we can treat each 2D slice as a gray-scale, 1 channel image. There are 251 mice and approximately 100 of the mice have no tumor. The ratio of the tumor to non-tumor mice is explicitly adjusted in the experiments. Each mouse has 3-6 weeks worth of 3D stacks and the scans are done around once a week. We only consider mice samples with 4 weeks of data points to ensure temporal consistency. In total, from 1000 3D Scans we expect around 1000 representative 2D images. The data is loaded dynamically, where the seed represents the data point i.d. The data set is sourced from the University of Denver.

4.2. Evaluation

We utilize a train-test split of 70:30 and present a few samples of the test predictions. We measure the quantitative metrics of Peak signal to noise ratio (PSNR) and structural similarity index measurement (SSIM) for the predicted fourth frame against the ground truth fourth frame. For the evaluation methodology, we follow [4], where the highest PSNR and SSIM sample prediction for each batch. We also measure another task of interest, tumor break emergence. Tumor “breaks”, depicted as holes or gaps between the white parts of CT scans, are a good indicator for tumor growth. Therefore, it follows that one of the most important things for the generation of the image to get correct is prediction of the emergence of “breaks” in the tumor in addition to other metrics such as PSNR and SSIM. Thus, we measure the accuracy for this task as well by drawing the same 100 sample points and assesing whether the generated image reflects the “break” status of the ground truth frame. For an example of a bone break, see the 4th row of figure 1 and the fourth frame.

4.3. Architecture Details

We implement three models: SVG with LP [4], SVG’ [12], and 3D-GAN [13].

SVG with LP SVG uses a DCGAN discriminator and generator architecture for encoder and decoder respectively. The decoder has a sigmoid output layer. The dimensionality of the sample from the latent distribution is $|z_t| = 10$ and the dimensionality of the output vector of the encoder is given by $g_{dim} = 128$. The loss function uses MSE loss which is equivalent to the l_2 norm. We run SVG with a batch size of 100, for 300 epochs, with an epoch size of 600. We use the Adam optimizer with momentum $\beta_1 = 0.9$ and a regularization term of $\beta = 0.0001$ on the KL term on the prior.

SVG’ The SVG’ implementation builds upon SVG and differs from it in a few ways. First, instead of optimizing for MSE, SVG’ optimizes for l_1 loss, which has been demonstrated to generate sharper images by [7]. Second, SVG’ uses a shallower encoder-decoder architecture (functionally, we interpret this to mean one less layer). The architecture also introduces two hyper parameters K and M which denote scaling factors for the neurons of the encoder-decoder and LSTM respectively. For the sake of a fair comparison, we set K and M to be 1 when comparing PSNR and SSIM in Table 1. We investigate the effect of increasing K and M in Table 2 using the methodology given by [12]. K is only able to be scaled to 2 while M is able to be scaled to 3 due to device limits.

3D-GAN We implement a standard 3D-GAN. The data x is encoded as a cube with H x W x T drawn from $\mathbb{R}$ as the three dimensions.

4.4. Results and Analysis

Qualatative The 3D-GAN predicted results appear to be blurrier. Due to working on larger images, the inner dynamics of each of the bone scans appears to be modeled better for the 3D-GAN, the little specks of white are clearer. On the SVG and SVG’ side, optimizing for l_1 loss rather than MSE does seem to improve the clarity of the images. The white components of the SVG’ predicted outputs are clearer.

Another thing to note is that the predicted frames appear very close to the ground truth frames. This could mean a variety of things. It could mean that the algorithm is overfitting to the data. However, this is not probable as proper train-test splits are implemented as well as the fact that the dataset is generated dynamically with the seed as the index. Another reason may be the fact that given the small number of frames to prime on as well as the small number of frames to predict (i.e. 3 and 1, respectively), the algorithm finds it optimizes by predicting a frame similar to the training data it had access to during the prior. This does not seem feasible as well as the algorithm sometimes predicts the emergence of bone breaks wrong, which implies that it is learning both the representations correctly and the temporal dependency between the representations.

Quantative The SVG methods had better performance on PSNR and SSIM than the 3D-GAN. SVG’ had the best performance on both PSNR and SSIM. One reason for why the SVG and SVG’ methods work better is because encoding temporal information as a third dimension is not a viable method. Time should not be treated as a third dimension like depth; rather the LSTMs utilized in SVG and SVG’ take better advantage of the temporal dependencies in the

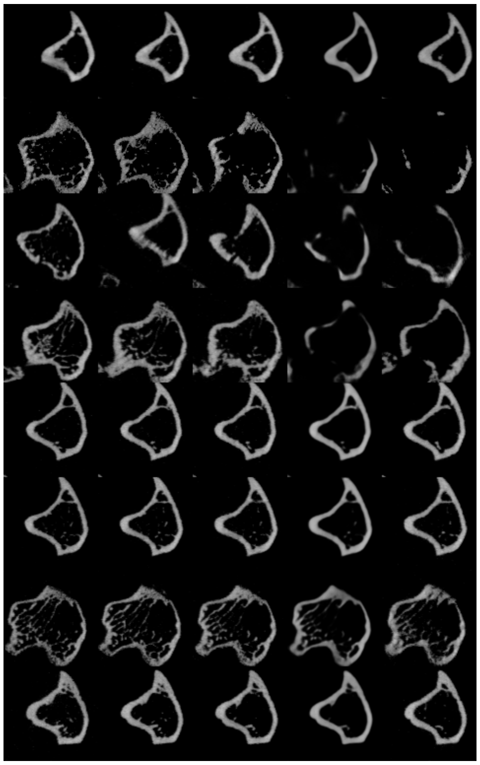

Figure 1. 3D-GAN prediction, fourth frame is predicted and the last frame is actual ground truth frame

Method	PSNR*	SSIM*
SVG	24.27	0.860
SVG'	**24.52**	**0.873**
3D-GAN'	23.20	0.813

Table 1. $*$ = higher is better

data. SVG' and SVG appear to perform the best for tumor break emergence as well given by its precision and recall scores. It is interesting to note that the minor architectural differences did not seem to have an effect on the precision and recall between SVG' and SVG. The effects of scaling K and M seems to support the hypothesis given by [12]; that larger networks while minimizing bias appear to yield state of the art results. However, this effect appears to be modest for this particular data.

5. Conclusion

In conclusion, we demonstrate that treating tumor growth prediction as a stochastic video prediction task appears to be a viable path for future research and yields promising results. The CT scans generated are qualitatively

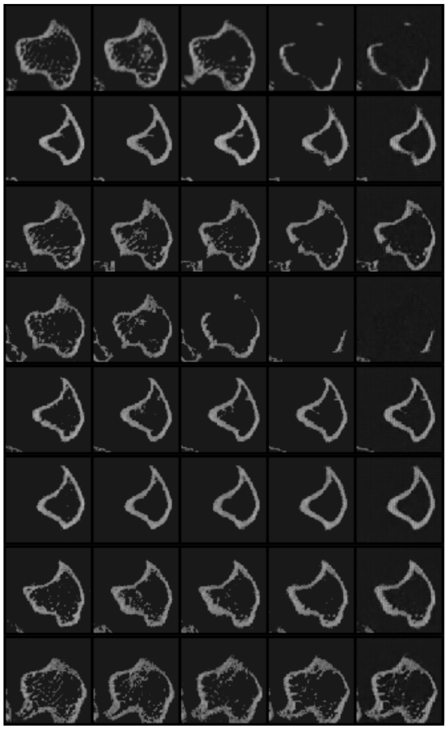

Figure 2. SVG prediction samples, fourth frame is actual ground truth and last frame is predicted

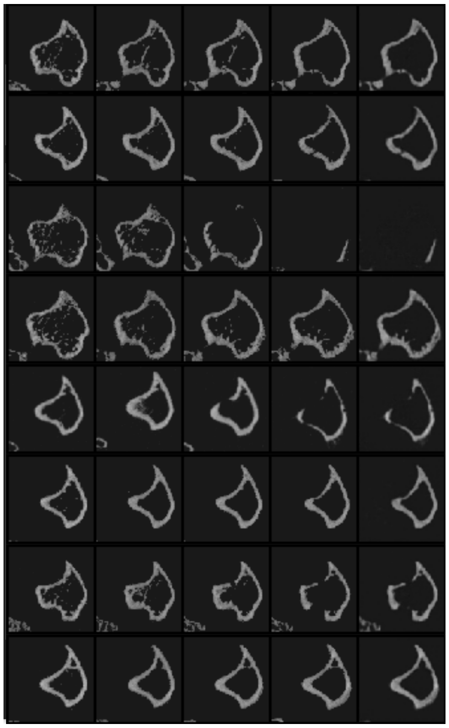

Figure 3. SVG' prediction samples, fourth frame is actual ground truth and last frame is predicted

K	M	PSNR	SSIM
1	1	24.52	0.873
2	2	24.53	0.877
2	3	**24.55**	**0.880**

Table 2. Scaling Hyperparameters K and M for SVG'

Method	Accuracy
SVG	94%
SVG'	94%
3D-GAN'	90%

Table 3. Accuracy for Break Prediction

clear, do well according to quantative metrics such as PSNR and SSIM, and predict tumor breaks to a reasonable accuracy.

Sparsity of medical data continues to be an issue for generative models which utilize many examples. By considering tumor growth prediction machine learning methods that do not rely on costly annotation, segmentation, and computation, which may require expert assistance, deep learning methods can be more widely accessible and greater assist all medical professionals. Future work may focus on producing more interpretable predictions, as the sensitive nature of medical data may require it.

References

[1] G. Adomian. Fisher-kolmogorov equation. *Applied Mathematics Letters*, 8(2):51 – 52, 1995.

[2] S. Bhagat, S. Uppal, Z. Yin, and N. Lim. Disentangling multiple features in video sequences using gaussian processes in variational autoencoders, 2020.

[3] X. Chen, R. Summers, and J. Yao. Fem-based 3-d tumor growth prediction for kidney tumor. *IEEE Transactions on Biomedical Engineering*, 58(3):463–467, 2011.

[4] E. Denton and R. Fergus. Stochastic video generation with a learned prior. *CoRR*, abs/1802.07687, 2018.

[5] A. Elazab, Y. M. Abdulazeem, A. M. Anter, Q. Hu, T. Wang, and B. Lei. Macroscopic cerebral tumor growth modeling from medical images: A review. *IEEE Access*, 6:30663–30679, 2018.

[6] A. Elazab, C. Wang, S. J. S. Gardezi, H. Bai, Q. Hu, T. Wang, C. Chang, and B. Lei. Gp-gan: Brain tumor growth prediction using stacked 3d generative adversarial networks from longitudinal mr images. *Neural Networks*, 132:321 – 332, 2020.

[7] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine. Stochastic adversarial video prediction, 2018.

[8] Y. Z. B. H. R;. Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research, Dec 2019.

[9] A. Rasouli. Deep learning for vision-based prediction: A survey. 2020.

[10] J. Shlens. Notes on kullback-leibler divergence and likelihood, 2014.

[11] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. *CoRR*, abs/1502.04681, 2015.

[12] R. Villegas, A. Pathak, H. Kannan, D. Erhan, Q. V. Le, and H. Lee. High fidelity video prediction with large stochastic recurrent neural networks, 2019.

[13] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90, 2016.

[14] A. Yin, D. J. A. Moes, J. G. van Hasselt, J. J. Swen, and H.-J. Guchelaar. A review of mathematical models for tumor dynamics and treatment resistance evolution of solid tumors. *CPT: pharmacometrics & systems pharmacology*, 8(10):720–737, 2019.

[15] L. Zhang, L. Lu, X. Wang, R. M. Zhu, M. Bagheri, R. M. Summers, and J. Yao. Spatio-temporal convolutional lstms for tumor growth prediction by learning 4d longitudinal patient data, 2019.