Supplemental files of "Hierarchical Topic-Aware Contextualized Transformers"

Ruiying Lu, Bo Chen, *Senior Member, IEEE,* Dandan Guo, Dongsheng Wang, and Mingyuan Zhou

I. Optimization and Inference

Here, we briefly describe the sampling methods of multi-layer representations and weight matrices for fast inference.

(i) *Sample the multi-layer representations $\boldsymbol{\theta}_{1:K}^{1:T}$ for all segments:* All segments of the target corpus are treated as BoW vectors $(\boldsymbol{d}_1, ..., \boldsymbol{d}_K)$, ignoring word order. We introduce a Weibull hybrid autoencoding inference (WHAI) network (encoder) (1) for PGBN (decoder). Denoting $Q = \prod_{t=1}^{T}\prod_{k=1}^{K} q(\boldsymbol{\theta}_k^t \mid \boldsymbol{d}_k)$, the negative ELBO of PGBN be expressed as follows:

$$L_{TM} = -\sum_{k=1}^{K}\mathbb{E}_Q\left[\ln P\left(\boldsymbol{d}_k \mid \boldsymbol{\Phi}^1\boldsymbol{\theta}_k^1\right)\right] + \sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}_Q\left[\ln\frac{q\left(\boldsymbol{\theta}_k^t \mid \boldsymbol{d}_k\right)}{P\left(\boldsymbol{\theta}_k^t \mid \boldsymbol{\Phi}^{t+1}\boldsymbol{\theta}_k^{t+1}, \boldsymbol{\tau}_k^{t+1}\right)}\right], \tag{1}$$

Similar to (1), we define $q(\boldsymbol{\theta}_k^t \mid \boldsymbol{d}_k) = \text{Weibull}(\boldsymbol{\beta}_k^t, \boldsymbol{\lambda}_k^t)$, a random sample from which can be obtained by transforming standard uniform noises $\boldsymbol{\epsilon}_k^t$ as

$$\boldsymbol{\theta}_k^t = \boldsymbol{\lambda}_k^t\big(-\ln(1-\boldsymbol{\epsilon}_k^t)\big)^{1/\boldsymbol{\kappa}_k^t}, \tag{2}$$

where $\boldsymbol{\beta}_k^t$ and $\boldsymbol{\lambda}_k^{(t)}$ are the parameters of $\boldsymbol{\theta}_k^t$ that are nonlinearly transformed from the hidden units $\boldsymbol{h}_k^t$ as

$$\boldsymbol{\beta}_k^{(t)} = \ln[1+\exp(\mathbf{W}_{h\boldsymbol{k}}^{(t)}\boldsymbol{h}_k^{(t)} + \boldsymbol{b}_1^{(t)})], \tag{3}$$

$$\boldsymbol{\lambda}_k^{(t)} = \ln[1+\exp(\mathbf{W}_{h\lambda}^{(t)}\boldsymbol{h}_k^{(t)} + \boldsymbol{b}_2^{(t)})], \tag{4}$$

where $\boldsymbol{h}_k^{(t)}$ are deterministically nonlinearly transformed from $\boldsymbol{d}_k$. All parameters in the encoder network are denoted as $\mathbf{W}_I$, which can be learned via SGD with negative ELBO expressed in (1).

(ii) *Sample the hierarchical connection weight matrices $\{\boldsymbol{\Phi}^t\}_{t=1}^T$:* For $\boldsymbol{\phi}_m^t$, the m-th column of the loading matrix $\boldsymbol{\Phi}^t$ of layer t, its sampling can be efficiently realized as

$$\left(\boldsymbol{\phi}_m^t\right)_{i+1} = \big[\left(\boldsymbol{\phi}_m^t\right)_i + \frac{\varepsilon_i}{P_m^t}\Big[\Big(\rho\widetilde{\boldsymbol{A}}_{:m.}^t + \eta_0^t\Big) - \Big(\rho\widetilde{A}_{\cdot m.}^t + M_{t-1}\eta_0^t\Big)\left(\boldsymbol{\phi}_m^t\right)_i \tag{5}$$

$$+\mathcal{N}\left(0, \frac{2\varepsilon_n}{P_m^t}\left[\text{diag}(\boldsymbol{\phi}_m^t)_i - (\boldsymbol{\phi}_m^t)_i(\boldsymbol{\phi}_m^t)_i^T\right]\right)\Big]_{\angle}, \tag{6}$$

Ruiying Lu, Bo Chen, Dandan Guo and Dongsheng Wang are with National Laboratory of Radar Signal Processing, Collaborative Innovation Center of Information Sensing and Understanding, Xidian University, Xi'an 710071, China. E-mail: ruiyinglu_xidian@163.com; bchen@mail.xidian.edu.cn; gdd_xidian@126.com; wds_dana@163.com;

Mingyuan Zhou is with McCombs School of Business, The University of Texas at Austin, Austin, TX 78712, USA. E-mail: mingyuan.zhou@mccombs.utexas.edu;

where $[.]_{\angle}$ denotes the simplex constraint that $\boldsymbol{\phi}_{m,k}^{(l)} \geqslant 0$ and $\sum_{m=1}^{M_t} \boldsymbol{\phi}_{m,k}^{(t)} = 1$, P_m^t is calculated using the estimated FIM, ε_i denotes the learning rate at the i-th iteration, both $\widetilde{\boldsymbol{A}}_{:m.}^{t}$ and $\widetilde{A}_{:m.}^{(t)}$ come from the augmented latent counts A^t, and η_0^t denote the prior of $\boldsymbol{\phi}_m^t$. More details of TLASGR-MCMC can be found in (2).

II. Generation Comparison

In order to compare the generation results of HiTac GPT-2 and original GPT-2, we utilize the same preceding context as input.As shown in Fig. 1 and Fig. 2, we compare the generated context of our proposed HiTac GPT-2 with the original pre-trained GPT-2, whose generated result is directly cited from the original paper (3). We find that our model can remember the semantic information with longer dependencies beyond the current segment, and generate semantically consistent paragraphs. Our fine-tuned model based on GPT-2 is powerful for generating long text paragraphs, whose generated text is highly semantic-related to the preceding context. Note that with the topic information extracted from the preceding context, our proposed model can generate both the semantically meaningful words that stay on the same topic, and the novel content which may not exist in the original corpus.

Preceding context :

Corporal Michael P. Goeldin was an unskilled laborer from Ireland when he enlisted in Company A in November 1860. Goldein survived the war. …… On 10 June, Artificer William Collins was promoted to corporal, probably to fill a combat leadership void for the crossing of the James River. Collins's service record does not reflect the qualities he demonstrated to earn this promotion, but he had obviously overcome some serious problems. Born in Sacketts Harbor, New York, Collins enlisted in the company in December 1853 at the age of twenty-two, and reenlisted in December 1858. Just a month before the war began in April 1861, Collins went "over the hill" and was not caught until three years later. Returned to the company on 22 March 1864, he was tried

GPT-2:

in June of that year. Sentenced to be shot for desertion, he remained on active duty. Captain James J. Dall was born in Virginia in 1829 during the Revolution and joined his father in the Army in 1836. He served with the 2d Virginia Infantry until 1845, when he went to the Western Reserve Regiment; and, on his discharge, he reenlisted with Company I, 25th New York Volunteers (which in 1845 numbered about 4,500 men). One of his brothers, James, served with his company in the same capacity in the same brigade. While the regiment remained in Connecticut through 1846, Captain Dall served with Company I at various times before his discharge. A veteran of the Mexican War, he had served two years on the staff of Brigadier General John J. Sullivan at San Antonio. During the Peninsula Campaign of 1863, he commanded a company of twenty-three and was in charge of the battery of fifty-four heavy mortars and the gunboat Pinta of the Battery, a gunboat and several other boats. Captain Dall was on active duty with Company I in the fall of 1865. Two years later, he was promoted to second lieutenant.

HiTac GPT-2:

he was tried for treason and later on in the court of the D.J. He was sentenced on 17 April, and was executed on 9 May. He was also sentenced to life imprisonment in the camp at Brandy with an appeal on 10 June, with the company receiving the order to withdraw, on a condition that he would not return to the American frontier. He returned to the company and had five years of his service, and the regiment received one year of enlistment. After retiring to Brandy, he worked as a clerk for his uncle at Wilkes for six years. He was discharged in April of 1864 when he was arrested under the pretence of being a lieutenant of the American Legion but did not serve. He returned to the company, and was also assigned to the Company S, at Ponder Point, Ohio. The regiment received five years of enlistment, served in service the next year that he enlisted, and was returned to the New York Corps in 1884. He was sent to the West Company under the command of Company C in October 1875. As the new Confederate Army took all of its territory with the same General, the men were not able to fully complete his journey until it had already moved down the Mississippi, when he was received on command by General H. W. Davis. With the regiments was enlisted, a number of these on all four regiments in March. During the first day of training there that they were selected to serve on the North, under General H. E. Davis. During that week the American corps, and the Army of Northern Virginia (North Carolina, Army of the United States Confederate Veterans) was divided. By 1877, an active-ment and full-service corps were now the Army of Fort Jackson. In January the army was transferred to the East Company by a special command and the Army of the Northern Virginia (E.W. Hatton. New. Col, West Company. New. Col, and Fort. Jackson. All the other regiments had the same division, which at the end of this year was called the division of the United State Army. In April 1784 the troops was to be sent in to assist on the new battle. Captain C. William H. Brown. First. S. A.

Fig. 1: Generated paragraphs given preceding context compared to GPT-2 (same example from the experiments of GPT-2).

Preceding context :
This pairing was the first outfit I thought of when I bought the shoes. It's like a summer version of this Jake Grantham outfit; in fact, my shoes are close to the colors of his Nike Racers! Instead of a heavy Harris Tweed jacket and denim shirt, I'm wearing a cotton DB jacket and and a linen shirt. Both fabrics (in these colors) are an absolute must for summer, as they go with both dark and and light pants! As you can see, they pair wonderfully with the dark jeans and shoes. It's a pseudo menswear/prep outfit. Overall, this is a very casual outfit which is why I paired my sneakers with it. I'm not about wearing a full wool suit with sneakers (as GQ shows a lot) but I'm definitely open to keeping things casual, like this cotton DB. Casual fabrics are key to pulling off your sneakers in a dressed down menswear outfit. I'd even suggest to wear these sneakers with a khaki chino suit or a white linen suit. Just be sure to ditch the tie or wear a tee or polo; wearing a tie with sneakers is a bit too much

GPT-2:
for me!
I know that some people might be opposed to wearing sneakers in a turtleneck, but I wanted to be true to myself, so I went with a slim fitting turtleneck in a color more similar to my favorite color of the day. Overall, it's a classic turtleneck, with nothing too flashy. I did swap my shoes for these ASICS Tiger Killshots and I wish I hadn't did that. The shoe itself is very comfortable, however, I found that wearing them in these sneakers made my feet look a bit larger. I'd wear them with denim, chinos, whatever, but would never recommend wearing them alone. There are just too many variables involved.

HiTac GPT-2:
if you want to wear your clothes casual.
I look pretty casual now, but the other jeans just don't always match.Here at first, I was pretty picky with my pants, but I'm now really in the top gear.This outfit is so flattering and so sexy for jeans and shorts, when you feel comfortable wearing. This summer outfit is also a good time to wear a skirt.It's actually a more formal shirt if only because it doesn't look like a top and this dress is a good for this kind of thing. And as is your choice. I like jeans and I'm wearing my dress, but the jeans just do not do it justice.One thing I wear now that would not be my current wear is a high street blouse.This outfit would make a great summer dress to outfit my jeans to give me casual outfits. I' t the only real choice when your dress is very dress shirt. My blouse is a great way to get a look I haven't worn for long, when I look over.You wouldn really want this pantsThere is also a high jumper, but the outfit I really like is that with the hat and gloves I have been wearing the entire time.This jacket has the usual button on back at its on the right side of the front for my suit.Now it is a good time to wear the shirt, if you want the look like that of a top.I always wear a paj.So to give my denim outfit another go, here comes all my best of a denim outfit, and I just don can you get into it. Here you should start wearing this hat so i‡s that it.t, with that tie I have a hat as well to give it one good go.For the trousers I had from my shirt, this shirt is my best summer clothing for my body.

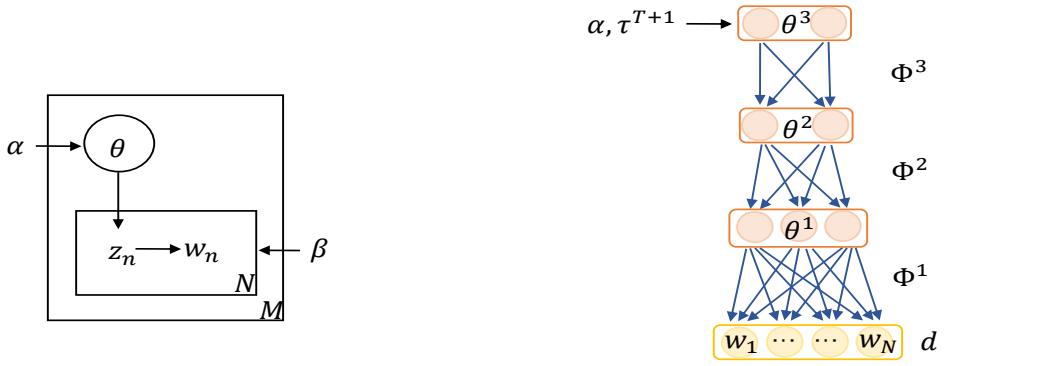
Fig. 2: Generated paragraphs given preceding context compared to GPT-2 (same example from the experiments of GPT-2).

III. Latent Dirichlet Allocation (LDA) and Poisson Gamma Belief Network (PGBN)

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. A corpus is a collection of M documents denoted by $\mathbf{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_M\}$. A document $\boldsymbol{d}$ includes a sequence of N words, denoted by $\boldsymbol{d} = (\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_N)$, where $\boldsymbol{w}_n$ is the n^{th} word in the sequence. LDA assumes the following generative process for each document $\boldsymbol{d}$ in a corpus $\mathbf{D}$:

$$\begin{aligned} &N \sim \text{Poisson}(\xi), && \boldsymbol{\theta} \sim \text{Dir}(\alpha), \\ &\boldsymbol{w}_n \sim \text{Multinomial}(\boldsymbol{w}_n|\boldsymbol{z}_n, \beta), && \boldsymbol{z}_n \sim \text{Multinomial}(\boldsymbol{\theta}), \end{aligned} \tag{7}$$

where ξ and β are prior parameters, and N refers to the number of words in the document $\boldsymbol{d}$. LDA is a hierarchical Bayesian model, in which each item $\boldsymbol{w}_n$ of a document is modeled as a finite mixture over an underlying set of topics. Each topic $\boldsymbol{z}_n$ is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities $\boldsymbol{\theta}$. In the context of text modeling, the topic probabilities provide an explicit representation of a document.



(a) Latent Dirichlet Allocation (b) Poisson Gamma Belief Network

Fig. 3: The illustration of (a) graphical model representation of LDA, where the boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. (b) The hierarchical structure of Poisson Gamma Belief Network, where the topic information is embedded in multiple hidden units $\boldsymbol{\theta}^t$.

In our manuscript, the baseline model of the proposed HiTac Transformers is a deep version of Latent Dirichlet Allocation (LDA), i.e. Poisson Gamma Belief Network (PGBN). To capture the hierarchical topic representation, we use the Poisson gamma belief network, a deep probabilistic topic model, to extract semantically meaningful multi-layer representations from the text. We represent a segment as a bag-of-words (BoW) count vector $\boldsymbol{d} \in \mathbb{Z}_+^V$, the v-th element of which counts how many times that the v-th word in vocabulary appears at this segment. As shown in Fig. 3 (b), the generative topic model with T hidden layers, from top to bottom, can be expressed as:

$$\begin{aligned} &\boldsymbol{\theta}^T \sim \text{Gam}\left(\boldsymbol{r}, \boldsymbol{\tau}^{T+1}\right), ..., \boldsymbol{\theta}^t \sim \text{Gam}\left(\boldsymbol{\Phi}^{t+1}\boldsymbol{\theta}^{t+1}, \boldsymbol{\tau}^{t+1}\right), \\ &\boldsymbol{\theta}^1 \sim \text{Gam}\left(\boldsymbol{\Phi}^2\boldsymbol{\theta}^2, \boldsymbol{\tau}^2\right), \boldsymbol{d} \sim \text{Pois}(\boldsymbol{\Phi}^1\boldsymbol{\theta}^1), \end{aligned} \tag{8}$$

where $\boldsymbol{\theta}^t \in \mathbb{R}_+^{M_t}$ denotes the hidden units of layer t, $\boldsymbol{\Phi}^t \in \mathbb{R}_+^{M_{t-1} \times M_t}$ the connection weight matrix of layer t, $\boldsymbol{\tau}^t > 0$ the gamma scale parameter of layer t, $\boldsymbol{r}$ the gamma shape parameters at the top layer. M_t denotes the number of topics at layer t and $M_0 = V$. PGBN factorizes the observed multivariate count vector $\boldsymbol{d}$ into the product of $\boldsymbol{\Phi}^1$ and $\boldsymbol{\theta}^1$ under the Poisson likelihood. It further factorizes the shape parameter of the gamma distributed hidden units $\boldsymbol{\theta}^t$ at layer t into the product of connection weight matrix $\boldsymbol{\Phi}^{t+1}$ and hidden units $\boldsymbol{\theta}^{t+1}$ of the next layer, capturing the dependence between different layers. For scale consistency and ease of inference, Dirichlet priors are placed on each column of $\boldsymbol{\Phi}^t$, i.e. ϕ_m^t for $\{m \in M_t, t \in T\}$, which makes the elements of each column be non-negative and sum to one.

REFERENCES

[1] H. Zhang, B. Chen, D. Guo, and M. Zhou, "WHAI: Weibull hybrid autoencoding inference for deep topic modeling," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[2] Y. Cong, B. Chen, H. Liu, and M. Zhou, "Deep latent Dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC," in *Proc. Int. Conf. Mach. Learn.*, 2017.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.