Augmentation Strategies Optimization for Natural Language Understanding

Chang-Ting Chu* Mahdin Rohmatillah* Ching-hsien Lee†
Jen-Tzung Chien*

*Dept of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Taiwan
†Computational Intelligence Technology Center, Industrial Technology Research Institute, Taiwan

IEEE ICASSP 2022

Success of AutoML

- AutoML has successfully introduced automated search process for augmentation strategy to improve model performance in CV tasks.
- Unfortunately, this augmentation strategy method requires high computational cost.

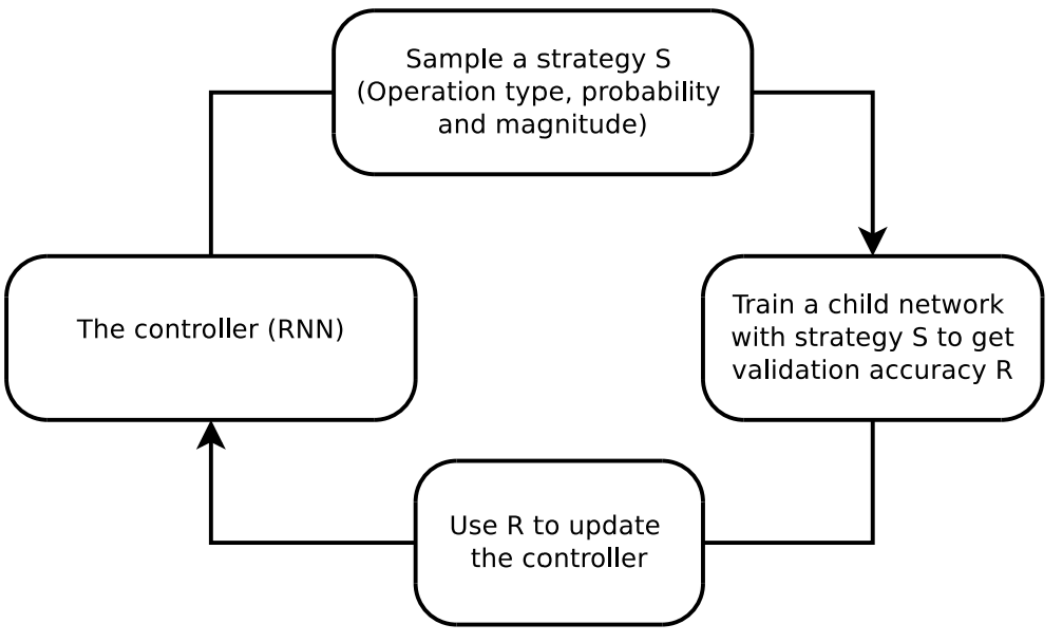


Figure: AutoAugment learning procedure.

The Urgency of Augmentation Strategy in NLP

- Lack of extensive research of augmentation strategy in NLP domain. Most of the previous approaches only apply single augmentation method for whole dataset.
- Simple noise affects the model's output significantly.

Delete

Document	Label	Pred.
–LRB– Taymor –RRB– utilizes the idea of making Kahlo's art a living, breathing part of the movie, often catapulting the artist into her own work.	5	4
–LRB– Taymor –RRB– utilizes the idea of making Kahlo's art a living, breathing part of ~~the~~ movie, often catapulting the artist into her own work.		1
There is no denying the power of Polanski's film...	4	4
There is no denying ~~the~~ power of Polanski's film...		1

Inconsistent Prediction - The Findings

Swap

Document	Label	Pred.
A sensitive and astute first feature by Anne–Sophie Birot.	**4**	4
A sensitive and **first astute** feature by Anne–Sophie Birot.		**1**
The concept behind Kung Pow: Enter the Fist is hilarious.	**4**	4
The concept behind **Pow Kung**: Enter the Fist is hilarious.		**1**

Inconsistent Prediction - The Findings

Insert

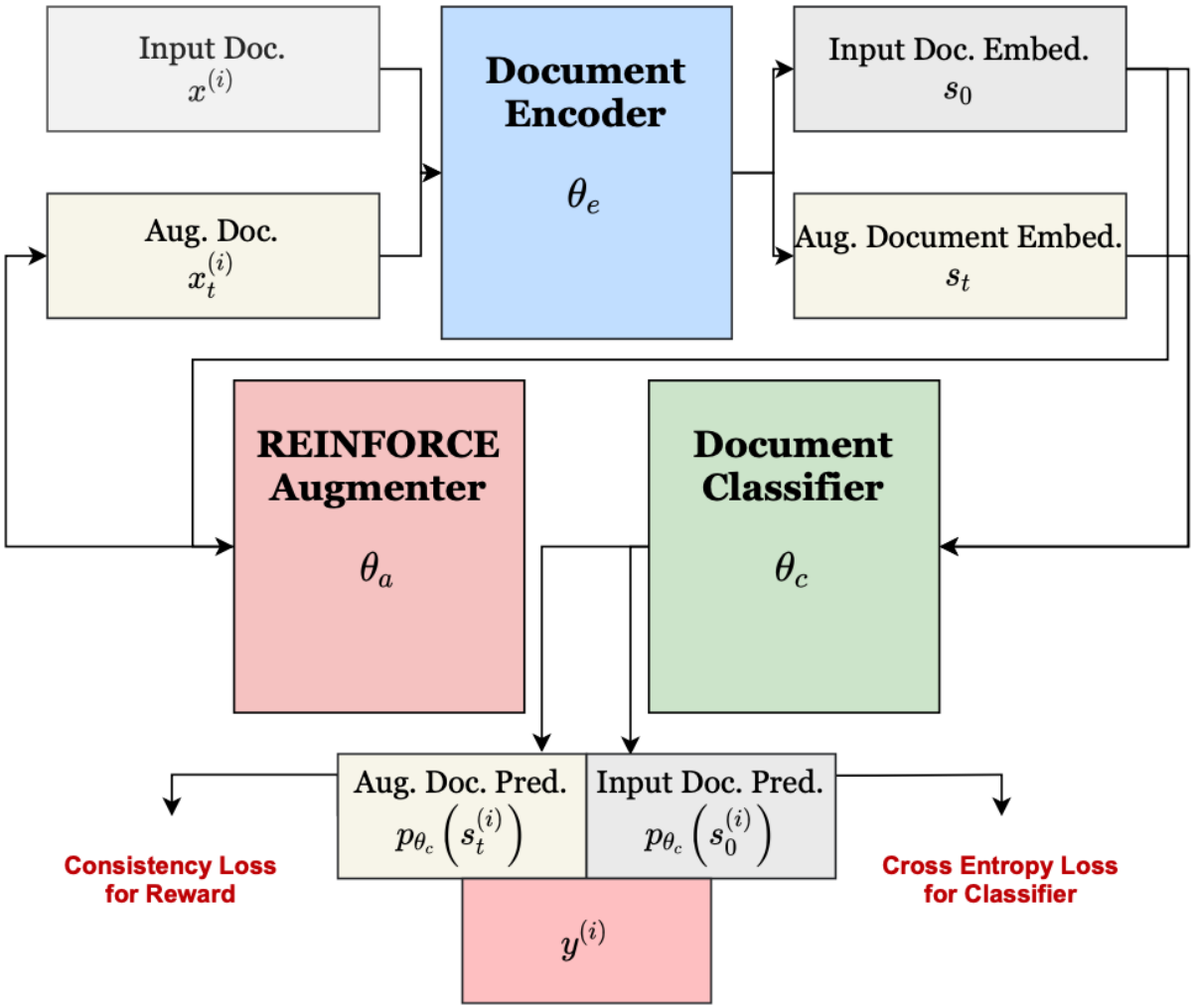
Document	Label	Pred.
A brutal and funny work.	**5**	4
A **cruel** brutal and **mirthful** funny work.		**1**
If you ignore the cliches and concentrate on City by the Sea's interpersonal drama, it ain't half–bad.	**3**	3
If you ignore the cliches and concentrate on City by the **ocean** Sea's interpersonal drama, it ain't half–bad.		**1**

Inconsistent Prediction - The Findings

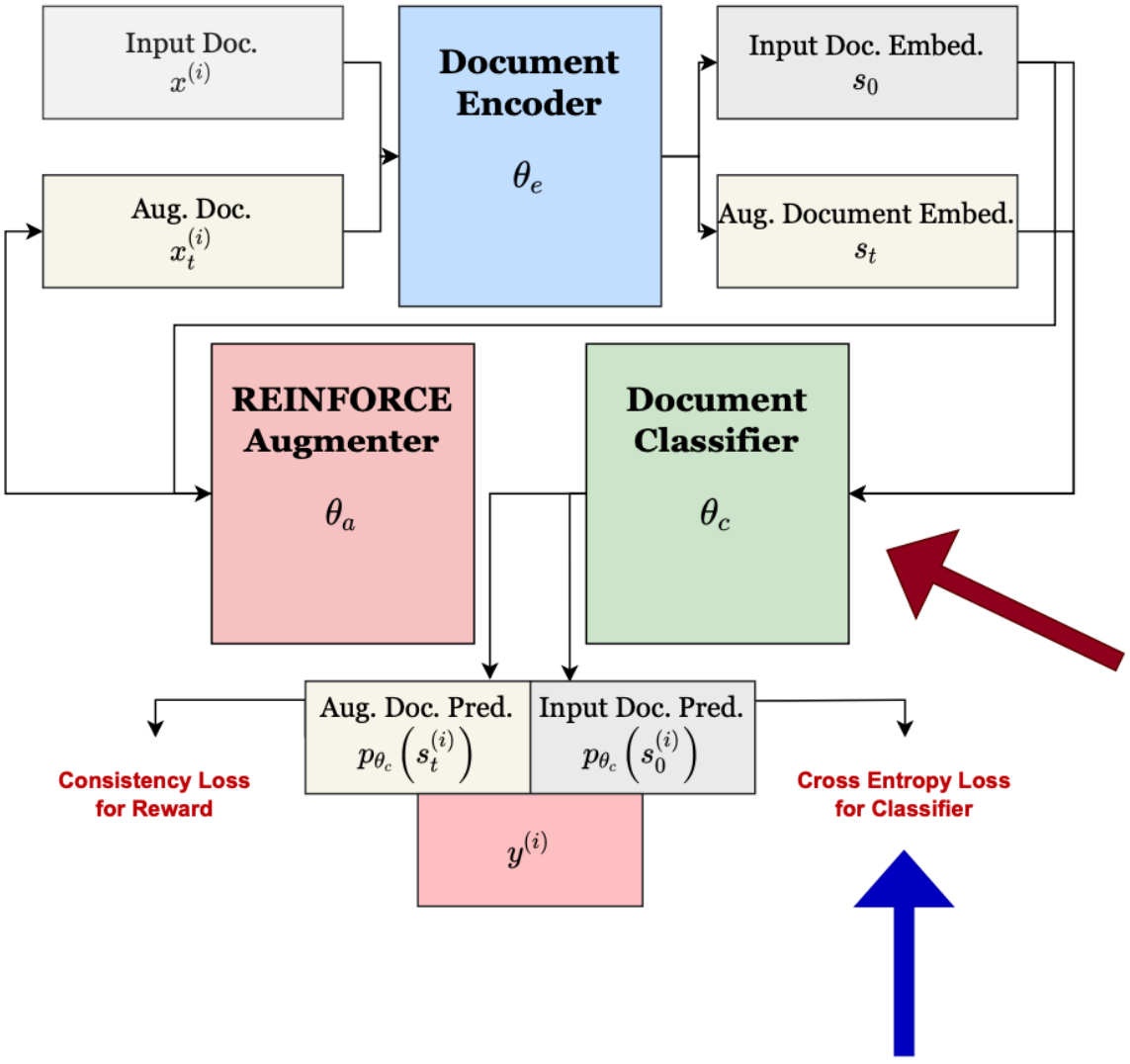
Replace

Document	Label	Pred.
One of the smarter offerings the horror genre has produced in recent memory, even if it's far tamer than advertised.	4	4
One of the smarter offerings the horror genre has produced in recent **retentiveness**, even if it's far tamer than advertised.		1
You don't need to be a hip–hop fan to appreciate Scratch, and that's the mark of a documentary that works.	4	4
You don't need to be a hip–hop fan to **apprise** Scratch, and that's the mark of a documentary that works.		1

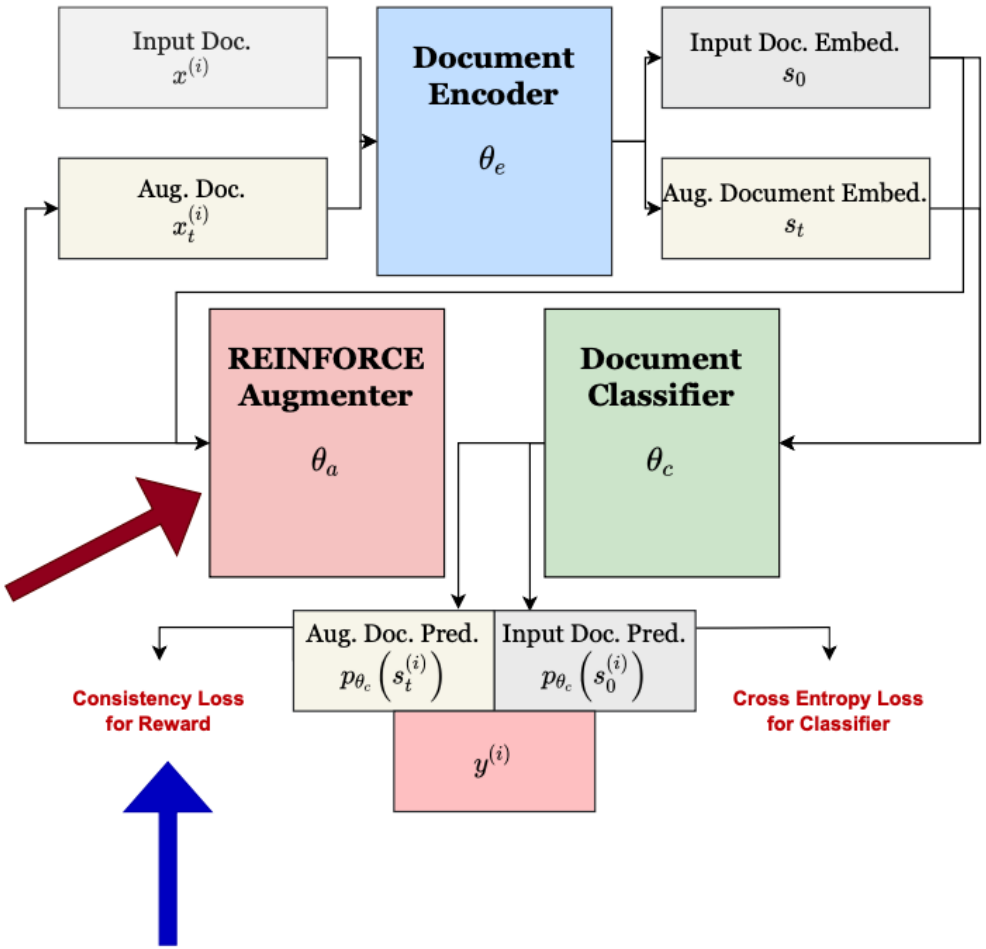
Overview Of Structure



Standard Text Classification Model Training



Stacked Data Augmentation



Stacked Data Augmentation

- State: embedded document, $s_t = f\left(x_t^{(i)}; \theta_e\right)$
- Action: five discrete actions

Label: Action	Sentence
0: **RD** (rand delete)	~~**Sparse**~~ only curiously compelling
1: **RS** (rand swap)	**compelling** only curiously **Sparse**
2: **SR** (syn replace)	Sparse only **oddly** compelling
3: **SI** (syn insertion)	Sparse only curiously **oddly** compelling
4: **Stop**	Sparse only curiously compelling

- Reward:
  - $\mathrm{JS}\left(p_{\theta_c}\left(s_t\right), p_{\theta_c}\left(s_0\right)\right) = \frac{1}{2}\left(\mathrm{KL}\left(p_{\theta_c}\left(s_t\right) \| \mathcal{M}\right) + \mathrm{KL}\left(p_{\theta_c}\left(s_0\right) \| \mathcal{M}\right)\right)$
  - $\mathcal{M} = \frac{1}{2}\left(p_{\theta_c}\left(s_t\right) + p_{\theta_c}\left(s_0\right)\right)$
  - $r_t = \begin{cases} \varepsilon, & \text{if } \cos(s_t, s_0) < \alpha \\ \mathrm{JS}(p_{\theta_c}(s_t), p_{\theta_c}(s_0)), & \text{else.} \end{cases}$
  - $r_t \leftarrow r_t - \rho_t, \quad$ where $\rho_t = \frac{\bar{r}t}{T}$

Stacked Data Augmentation

Algorithm 1: Training for augmentation strategy

Require: $\mathcal{D}$ training dataset. η learning rate
T maximum number of steps
θ_e, θ_a, pars of encoder and augmenter

while θ_a *is not converged* **do**

  for *i=1, . . . ,* $|\mathcal{D}|$ **do**

    input $x_0^{(i)}$ as i^{th} document in $\mathcal{D}$

    $s_0 = f(x_0^{(i)}; \theta_e)$ as the embedding of $x_0^{(i)}$

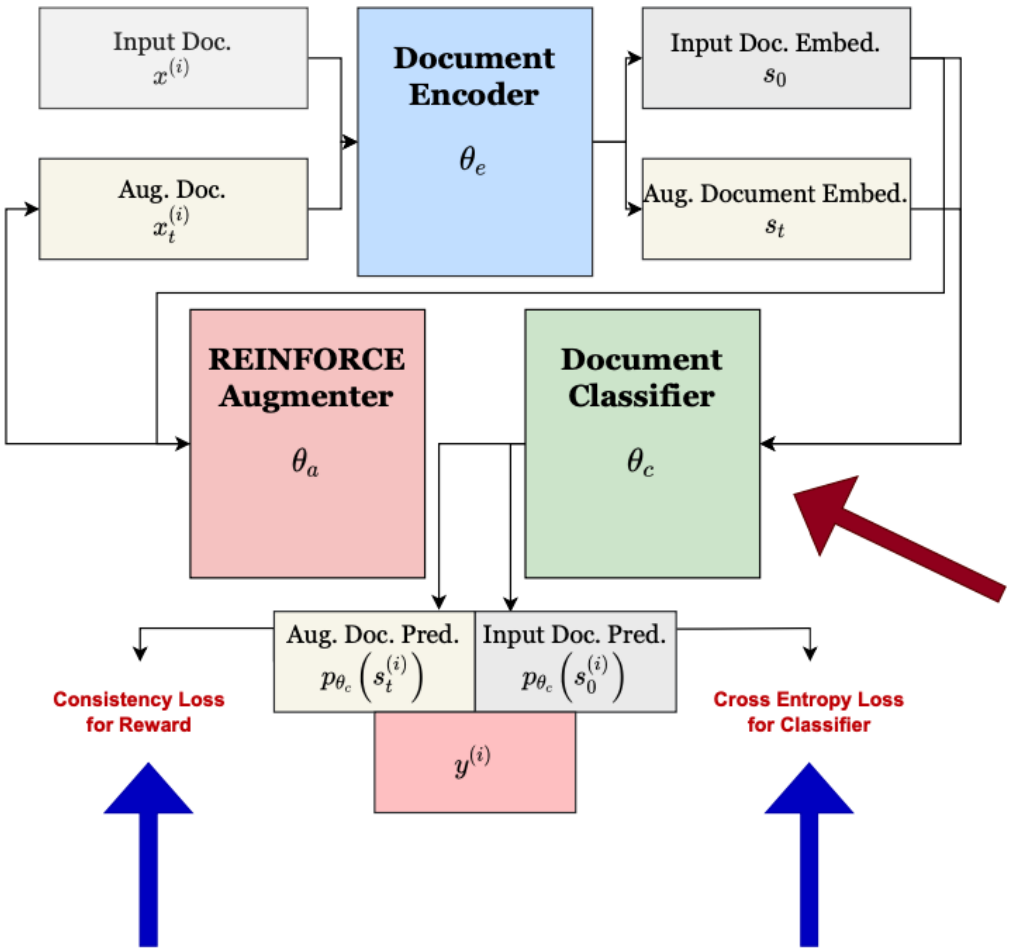
    $\{s_0, a_0, r_0, \ldots s_{T-1}, a_{T-1}, r_{T-1}\} \sim \pi_{\theta_a}(\tau)$

    $G_t = \sum_{t'=t+1}^{T} \gamma^{t'-t-1} r_{t'}$

    $g_{\theta_a} \leftarrow \nabla_{\theta_a} \sum_{t=0}^{T-1} \log \pi_{\theta_a}(a_t|s_t) \cdot G_t$

    $\theta_a \leftarrow \theta_a + \eta \cdot \text{Adam}(\theta_a, g_{\theta_a})$

Diversity-Promoting Consistency Training



Example of Augmented Documents

Table 1: Illustration for the proposed stacked data augmentation (SDA) with five actions (0: random delete, 1: random swap, 2: random synonym replacement, 3: random synonym insertion, 4: stop operation). "x" denotes the failed action due to losing of original semantic meaning, indicated by the condition $\cos(s_t, s_0) < \alpha$. The order of actions and the received reward are shown.

Original Document	**Augmented Document**	**Action**	**Reward**
The name says it all.	The name pronounce it totally	2 2 0 4	0.0484
A lovely and beautifully photographed romance.	A take shoot photograph and lovely beautifully	3 1 2 0 3 4	0.0091
Rouge is less about a superficial midlife crisis than it is about the need to stay in touch with your own skin, at 18 or 80.	with skin. it than less about at or your the ain superficial is midlife crisis, stay need sense of touch touch contain is in about to Rouge vitamin a 18	1 2 2 3 2 3 0 2 3 1 x	0.0049

Performance Result with Action Distribution

Table 2: (left) Distribution of actions taken by the policy. **Sim.Thr.** stands for similarity threshold. **Stop** indicates the successfully augmented document without exceeding the max step T or violating the similarity threshold α. (right) Accuracy (%) on different classification tasks. The results from reference papers are shown. "-" denotes the missing results. Augmentation methods using EDA , and back-translation (Back) are compared with SDA. pre-trained model using RoBERTa is merged.

Sim.Thr.	0.7	0.8	0.9
Delete	8.3%	3.8%	3.7%
Swap	0.8%	4.9%	8.0%
Replace	39.8%	22.4%	**67.2%**
Insert	**51.1%**	**68.6%**	20.8%
Stop	7.5%	20.8%	29.1%

Model	**SST-2**	**SST-5**	**CR**	**MPQA**	**Subj**	**TREC**
EFL	**96.9**	-	92.5	90.8	97.1	-
byte mLSTM	91.7	54.6	90.6	88.8	94.7	90.4
BERT	93.1	55.5	-	-	**97.3**	96.8
RoBERTa	94.8	56.6	93.2	90.4	96.0	96.8
RoBERTa with EDA	94.6	56.9	93.3	90.0	95.3	96.6
RoBERTa with Back	95.0	57.3	94.1	90.9	96.9	**97.4**
RoBERTa with SDA	95.2	**58.6**	**94.7**	**91.4**	96.0	97.0

Conclusion

- A new method for searching the stacked distinct augmentation actions has been presented to employ in six text classification tasks.
- The results showed that the generalization of the model with strong language understanding module could be further improved with the proposed method
- The augmentation policy could generate some readable sentences and behaved diversely in different settings of REINFORCE augmenter