Table 2: Results of various pre-trained checkpoints on the validation set. We found statistically significant improvements in BLUE-N and EDS for some pre-trained checkpoints, but not all of them had such benefits (e.g., BERT and BART-large).

Checkpoint	BLEU-2	BLEU-3	BLEU-4	DIST-1	DIST-2	DIST-3	EDS
RND	44.4720.64	35.8818.12	28.8416.22	**10.614.32**	**28.4711.03**	**41.7915.24**	38.0213.10
BERT	21.7214.01	15.4410.61	10.477.90	8.715.49	19.0311.95	27.0917.03	24.8110.40
GPT-2	46.7621.23	38.2019.41	31.1418.14	10.394.39	27.8611.28	40.9315.70	39.9414.84
BART-base	**48.3421.47**	**39.8520.11**	**32.8219.22**	10.294.31	27.9811.01	41.2815.15	**40.7715.75**
BART-large	22.4114.28	15.9610.71	10.917.78	9.216.36	20.5413.28	29.0918.04	24.9810.27

In Fig. 1, we display the training and validation curves for the five pre-trained checkpoints mentioned before. Note that the vocabulary size of the BART decoder (50,265 tokens) is much larger than that of the other randomly initialised decoders (164 tokens), which leads to higher losses but does not necessarily mean that the generation quality of BART is worse. Regardless, Fig. 1 suggests that using pre-trained checkpoints to initialise the model does not guarantee a lower validation loss.

Because of the small amount of data, all models showed different degrees of overfitting. In particular, even though the number of parameters is approximately three times that of BART-base, the validation loss of BART-large is not lower. Intuitive ways to solve this problem are to collect more data, reduce the model size, or tune hyperparameters. However, due to the scarcity of symbolic music data, it is unlikely to find a human-annotated text-music dataset that is at least an order of magnitude larger (i.e., 1 million text-music pairs) than Textune for a long time. Thus, using smaller models or tuning hyperparameters are attainable solutions for now.

To verify the generation quality, we used all checkpoints with their lowest validation loss to generate tunes based on descriptions from the validation set, and using nucleus sampling with top-$p = 0.9$. We used the following metrics to evaluate the generated tunes from different models.

BLEU-N (?): An algorithm for evaluating the quality of text measures the proportion of N-grams in the reference text are reproduced by the candidate text. The higher the value, the closer the generated tunes are to ground truth. This is a common metric used in sequence-to-sequence tasks.

DIST-N (?): It evaluates the diversity of generated samples. A higher value of DIST-N means a higher proportion of distinct N-grams. We use this reference-free metric as text-to-music generation can be seen as conditional music generation, which is a creative task.

EDS: Edit **D**istance **S**imilarity is based on the Levenshtein distance $lev(a, b)$ to indicate how similar the generated tune b and the ground truth a are at the character level, ranging from 0 (no match at all) to 100 (exact match), which can be formalised as follows:

$$EDS(a, b) = (1 - \frac{lev(a, b)}{max(|a|, |b|)}) \times 100, \qquad (1)$$

where $|a|$ and $|b|$ are the length of two strings. As ABC tunes are nearly character-level sequences, EDS can effectively reflect the similarity between the generated tune and the ground truth.

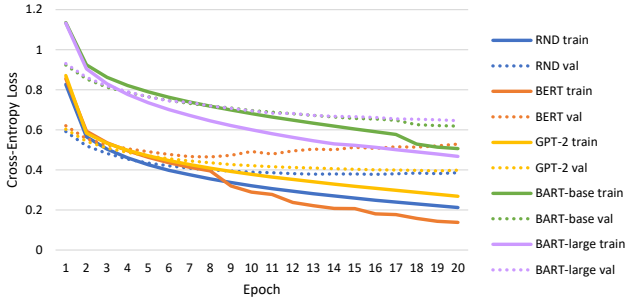

Figure 1: Training and validation curves of various models.

As shown in Table 2, RND generated more diverse tunes (higher DIST-N), but the tunes generated by GPT-2 and BART-base are closer to the ground truth (higher BLEU-N and EDS). We performed independent samples t-tests, and found statistically significant differences in the BART-base results of BLEU-N and EDS compared to that of RND, i.e., p-value < 0.05. These results show that the use of pre-trained checkpoints can improve the performance of the model on language-music tasks significantly.

For two randomly selected tunes from the Textune dataset, the average EDS is around 30%, while the results for RND, GPT-2 and BART-base on this metric are around 40%. Given the diversity of music, it indicates that these models can extract meaningful information from descriptions to generate music. However, the tunes generated by BERT and BART-large are not only low in diversity (lower DIST-N) but also far from the ground truth (lower BLEU-N and EDS). EDS suggests that the tunes generated by them are even more dissimilar to the ground truth compared to tunes randomly selected from Textune. We manually examined the tunes generated by these two models and found that there were a large number of instances of degeneration. We observed that they tend to repeatedly generate notes like `z8 | z8 | z8`. As shown in Fig. 1, we suggest that the cause of this problem is the severe overfitting of BERT and BART-large.

Discussions

To demonstrate the capabilities and limitations of language-music models, several representative examples of generated tunes are given in Fig. 2. All the text descriptions in Fig. 2 were hand-crafted by us and are not from Textune. Due to

Input Text A:
The tune "Twinkle, Twinkle, Little Star" is a popular English lullaby.

Output Tune A:
L:1/4
M:4/4
K:C
C C G G | A A G2 | F F E E | D D C2 | G G F F | E E D2 | G G F F | E E D2 | C C G G | A A G2 |
F F E E | D D C2 |]

Input Text B:
Lead sheet of the popular English lullaby "Twinkle, Twinkle, Little Star".

Output Tune B:
L:1/4
M:4/4
K:C
"C" C C G G |"F" A A"C" G2 |"G" F F"C" E E |"G" D D"C" C2 |"C" G G F F |"Em" E E"G" D2 |
"C" G G"F" F F |"Em" E E"G" D2 |"C" C C G G |"F" A A"C" G2 |"G" F F"C" E E |"G" D D"C" C2 |]

Input Text C:
Jazz arrangement of the popular English lullaby "Twinkle, Twinkle, Little Star".

Output Tune C:
L:1/8
M:4/4
K:F
"F" A3 G"Dm7" A c2 G |"Gm7" A B2"C7" G- G4 |"Gm7" d3 c"C7" d f2 c |"Gm7" f4-"C7" f2 z c |
"F" d3 c"F#dim7" d f2 c |"Gm7" f f2 d"A7b9" f2 z c |"Dm7" d3 c"G7" d f2 c |"C7" e4-"F#dim7" e2 z2 |
"Gm7" g3 f"C7" g b2 g |"F6" c' a2 f"Bb7" c'2 z c' |"Gm7" d'3 c'"C7" d' f2 c |
"F""Bbm6" f4-"F" f2 z2 ||"Dm7" f3 d"G7" f a2 f |"Gm7" b g2 f"C7" b2 z g |
"F" a3 g"Dm7" a c'2 g |"G7" a f2 d"A7b9" f2 z c |"Dm7" d3 c"G7" d f2 c |
"C7" e4-"F#dim7" e2 z2 |"Gm7" g3 f"C7" g b2 g |"F6""Bbm6" c' a2 f"A7b9" c'2 z c' |
"Dm7" d'3 c'"G7" d' f2 c |"C7" e4-"F#dim7" e2 z2 |"Gm7" g3 f"C7" g b2 g |
"F6" c' a2 f"Bbm6" c'2 z c' |"Gm7" d'3 c'"C7" d' f2 c |"F""Bbm6" f4-"F" f2 z2 |]

Input Text D:
Note Length-1/8
Key-D
Source-Dance Music of Ireland
Meter-6/8

Output Tune D:
L:1/8
M:6/8
K:D
f/e/ | dAA AFA | BGB A2 (f/e/) | dAA AFA | g3 f2 e | dAA AFA | BGB A2 f | aba gfe | fdd d2 ::
(a/g/) | (f/g/a)a b2 a | (f/g/a)a b2 a | (f/g/a)a faa | e2 f gfe | (f/g/a)a b2 a | (f/g/a)a b2 a | gfg eag | fdd d2 :|

Input Text E:
This is a traditional Irish dance music in the key of D and its meter is 6/8.

Output Tune E:
L:1/8
Q:1/4=100
M:4/4
K:C
CDE GAB | cde d2 B | cdc edc | B3 d3 | CDE GAB | cde d2 B | cde GAB | c3 c3 :: g2 g gfe |
f2 f fed | e2 e edc | d2 d dcB | CDE GAB | cde d2 B | cdc GAB | c3 c3 :|

Figure 2: Music generation examples of RND

space constraints, we only show RND-generated examples. We first tested whether the model could reproduce the tunes already present in Textune. The tune chosen here is "Twinkle, Twinkle, Little Star", which was present in Textune a total of 11 times. We found that when top-p was set to a low value (e.g., 0.5), the model almost always reproduced the tune perfectly, as shown in Fig. 2A. This means that the model does understand the relationship between the title and the tune. However, this also indicates that it is possible for the model to directly copy the music that exists in Textune. We recommend using a higher top-p when generating tunes using this model to avoid that problem.

We then tested the creativity of the model: generating the lead sheet and the jazz arrangement of "Twinkle, Twinkle, Little Star". According to Fig. 2B, the model does understand what a lead sheet is and succeeds in placing appropriate chord symbols for this tune. It should be noted that Textune does not contain any lead sheets for this tune. This demonstrates the potential of language-music models to be applied to the melody harmonization task. However, for the more creative task, melody style transfer, the model did not perform well. The tune in Fig. 2C, although it does have a very distinctive jazz style (e.g., rhythm, harmony), has a completely different melody from "Twinkle, Twinkle, Little Star". Given that the model can perform well on the melody harmonization task, we believe that the reason for its failure on the melody style transfer comes mainly from the small amount of text-music data. If the size of text-music datasets can reach the level of text-image datasets (**?**), achieving most music generation tasks, including those requiring a high degree of creativity, should not be a challenge anymore. We finally tested whether the model can follow the objective meta-information (e.g., key, meter) given in the text to generate tunes. We specified the key (D major), the meter (6/8), and the style of the music (Irish dance music). As shown in Fig. 2D and Fig. 2E, whether or not the model can generate music that matches the meta-information given in the text description depends on its format. When describing meta-information in a list format (Fig. 2D), the model can always follow the text accurately to generate tunes. The generated tune also exhibits distinctive characteristics of Irish dance music. For example, traditional Irish music is usually in a binary form (AABB), and the music generated here is exactly composed in that way. However, when the same information is given in a more loose way (Fig. 2E), the model does not follow the description well enough, even with a low top-$p = 0.5$. Although the actual meter of this generated tune is still 6/8 and is in keeping with the characteristics of Irish music, the generated music is in the key of C major and the meter in the header is 4/4. We tested this text format on the dual task (i.e., music-to-text generation) and found that when given the prompt "... in the key of", the model can always retrieve the meta-information correctly. Theoretically, the two tasks should be of equal difficulty, i.e., correctly translating the text to the header of ABC tunes or vice versa. More investigation is needed to determine the causes of this problem.

Conclusions

In this paper, we carry out the study of language-music models trained on large-scale text-music data. According to the experimental results, the use of pre-trained checkpoints leads to generated tunes that are much more similar to ground truth, but not improved in terms of diversity. Although the model can generate tunes that matched the semantic information of the text and exhibited a certain degree of creativity on some tasks, its creativity is limited, and it is input-sensitive. With a larger dataset, it is likely to develop a language-music model that performs well in music generation tasks that require a high degree of creativity. Voluptates eligendi quas labore ex, alias eveniet itaque ducimus odio quis aperiam dolor sunt, impedit nobis libero maiores odit magnam mollitia, ad eveniet adipisci pariatur.Amet quaerat eius iure in similique esse debitis sed reprehenderit alias, dicta ratione suscipit voluptas?Delectus amet quod corporis nam molestiae aperiam voluptatibus maiores, deserunt ipsum molestias cupiditate eum dicta officiis iure, debitis atque magnam nulla.Veniam repudiandae quia vel enim, dicta similique beatae quidem reprehenderit possimus et sed